DeToxy: A Large-Scale Multimodal Dataset for Toxicity Classification in Spoken Utterances

Sreyan Ghosh¹, Samden Lepcha³, S Sakshi⁴, Rajiv Ratn Shah², S. Umesh¹

¹Speech Lab, Department of Electrical Engineering, IIT Madras, Chennai, India
²MIDAS Labs, IIIT-Delhi, India
³TEG Analytics, Bangalore, India, ⁴Cisco Systems, Bangalore, India

gsreyan@gmail.com, sam.lepcha@outlook.com, ssakshi1397@gmail.com, rajivratin@iiitd.ac.in, umeshs@ee.iitm.ac.in

Abstract

Toxic speech is regarded as one of the crucial issues plaguing online social media today. Most recent work on toxic speech detection is constrained to the modality of text and written conversations with very limited work on toxicity detection from spoken utterances or using the modality of speech. In this paper, we introduce a new dataset DeToxy, the first publicly available toxicity annotated dataset for the English language. DeToxy is sourced from various openly available speech databases and consists of over 2 million utterances. We believe that our dataset would act as a benchmark for the relatively new and unexplored Spoken Language Processing (SLP) task of detecting toxicity from spoken utterances and boost further research in this space. Finally, we also provide strong unimodal baselines for our dataset and compare traditional two-step cascade and End-to-End (E2E) approaches. Our experiments show that in the case of spoken utterances, text-based approaches are largely dependent on gold human-annotated transcripts for their performance and also suffer from the problem of keyword bias. However, the presence of speech files in DeToxy helps facilitate the development of E2E speech models which alleviate both the above-stated problems by better capturing speech clues.

Index Terms: Speech Toxicity Analysis, End-to-End, 2-step, Multimodal

1. Introduction

Social network platforms are generally meant to share positive, constructive, and insightful content. However, in recent times, people often get exposed to objectionable content like threats, identity attacks, hate speech, insults, obscene texts, offensive remarks, or bullying. Though previous research on these topics has often focused on them individually, in this work we denote such content as “toxic speech”, and consider it as a super-set of all the above sub-types [1].

With the rise of different forms of content available online beyond just written text, i.e., audio and video, it is crucial that we devise efficient content moderation systems for these forms of shared media. However, most prior work in literature and available datasets focus primarily on the modality of conversational text, with other modalities of conversation ignored at large. Thus, to alleviate this problem, in this paper, we propose a new dataset DeToxy, for the relatively new and unexplored SLP task of toxicity classification in spoken utterances. This remains a crucial problem to solve for interactive intelligent systems, with broad applications in the field of content moderation in online audio/video content, gaming, customer service, etc. DeToxy is a large-scale multimodal dataset with both speech and text cues, manually annotated for toxicity detection in spoken utterances, with potential applications in building unimodal or multimodal models for conversational utterance-level Speech Toxicity Classification (STC) and E2E speech systems which better capture semantics of the utterance.

The key challenge in toxicity classification in spoken speech is to learn good representations that capture toxicity signals from raw speech signals which are inherent in a rich set of acoustic and linguistic content, including semantic meaning, speaker characteristics, emotion, tone, and possibly even sentiment information while remaining invariant under different speakers, acoustic conditions, and other natural speech variations. However, different from Speech Emotion Recognition (SER) and similar to Speech Sentiment Recognition (SSER), toxicity classification from spoken utterances involves the system capturing semantic and syntactic properties of the utterance beyond just acoustic and prosodic cues which makes this task all the more difficult.

Traditional approaches for various speech downstream tasks involving spoken utterance classification employ low-level acoustic features, such as band-energies, filter banks, and MFCC features [2], or raw waveform [3]. We acknowledge the fact that models trained on these low-level features can easily overfit to noise or signals irrelevant to the task. One way to remove variations in speech is to transcribe the audio into text and use text features to predict toxicity as done with SSER by [4]. Nonetheless, toxicity signals in the speech, like tone and emotion, can be lost in the transcription. Thus, it is essential for any system to learn speech representations that make high-level information from speech signals available to solve downstream SLP tasks. Inspired by this, we provide both 2-step cascade and E2E baselines for STC on DeToxy and analyze the advantages and disadvantages in both, concluding why the modality of speech is crucial for the task of STC and makes DeToxy an important contribution to the speech community. Additionally, we show how prosodies and linguistic cues in natural speech aid our E2E model, which outperforms the 2-step methodology using less than 10% of the total amount of labeled data available to our 2-step system.

2. Related Work

Toxic speech detection is a challenging task extensively studied in the literature, with techniques including dictionary-based search [5], distributional semantics [6], and recent literature exploring the power of neural network architectures for the task [7]. However, most of the work done on toxic speech detection is constrained to just text, in English and some other foreign languages [8, 9, 10] with the only existing work done by [11] for...
audio and videos. Another notable mention is the work done by authors of [13], where they employ speech representations learned on a downstream SER task, as a second modality of input to the final model they learn for hate speech detection.

On the other hand, the most commonly explored SLP tasks include Automatic Speech Recognition (ASR), SER, speaker verification, speaker identification, speech separation, speech enhancement, Named Entity Recognition (NER), phoneme recognition, etc. and the most recently explored speech sentiment classification which is very related to our task. All of these tasks are well studied with a lot of open-source datasets available online. With recent advancements in Natural Language Processing (NLP) and SLP, a lot of the systems achieving state-of-the-art in these tasks leverage E2E multi-layer neural networks like CNNs or Transformers, including self-supervised learning (SSL) and semi-supervised techniques which have shown to help models in the low-resource data paradigm and proven to learn powerful speech representations [14, 15, 16].

With downstream SLP tasks like speech sentiment classification and NER requiring an understanding of the contents and semantics of the spoken utterance, people have employed both 2-step [17] and E2E methodologies [18, 19]. Some of these self-supervised methodologies, originally invented for pushing the performance of state-of-the-art ASR systems, have also shown success in other downstream tasks [16].

3. DeToxy and DeToxy-B

In this paper, we present DeToxy, the first annotated dataset for toxic speech detection in the English language. Our dataset is a subset of various open-source datasets, detailed statistics about which can be found in Table 1. We also present DeToxy-B, a balanced version of the dataset, curated from the original larger version taking into consideration auxiliary factors like trigger terms and utterance sentiment labels.

3.1. Dataset Collection and Annotation

For annotation, we define toxicity as rude, disrespectful, or otherwise likely to make someone leave a discussion. This definition is reminiscent of the wealth of literature on toxic speech classification [8, 9, 20]. For the initial version of our dataset, we primarily focus on openly available speech databases with or without text transcripts available. For obtaining transcripts of datasets for which transcripts were not available, we use pre-trained wav2vec-2.0 [14] to obtain the transcriptions from a similar domain. In the process of consolidation of our dataset, primarily, after an empirical analysis of the transcripts of a majority of the open-source datasets available online, we found that most of them did not contain toxic utterances. Thus, we follow a 2-step approach to annotate our dataset. First, we train a textual toxicity classifier using BERTbase [21] and use that to select datasets that had at least 10% of their total number of utterances flagged as toxic by the model. Some datasets which did not fit this criterion were LibriSpeech [22], TIMIT [23], TEDLIUM [24] and MASS [25]. Next, all these utterances from all datasets filtered through step 1 were manually annotated by 3 professional annotators using audio [26] taking both the textual transcript and speech files into account. Finally, we do a simple majority voting among the 3 annotations to determine the final toxicity class of each utterance. The Cohen’s Kappa Score for inter-annotator agreement is 0.76. Table 1 and Table 2 show detailed descriptions of DeToxy and DeToxy-B datasets respectively.

3.2. Sampling DeToxy-B and Test-Trigger Datasets

For DeToxy-B, we keep all the toxic utterances from DeToxy and sample thrice the number of toxic utterances as our non-toxic utterances. We acknowledge the fact that sentiment is a strong controlling factor for various speech cues, and oversampling any one of the sentiments might lead our E2E model to overfit the same. Thus, we don’t sample at random but use sentiment labels to sample non-toxic utterances so that the final DeToxy-B dataset has an equal distribution of sentiments for its non-toxic utterances. These sentiment labels were annotated by a single annotator only for the purpose of constructing DeToxy-B and thus we don’t include these labels in our final release of the dataset. In this paper, we evaluate all our proposed baselines on only DeToxy-B, with a Training, Development and Test stratified split of 70:15:15.

Finally, we also collate an explicit test set with just non-toxic utterances where each utterance consisted of at least one trigger term which we call the Test-Trigger (Test-T)\(^1\). We do this to evaluate how our baselines perform against not getting biased to trigger terms, a long-standing problem in toxicity classification [27, 28]. Our Test-T dataset is sampled as a subset from the DeToxy-B Validation and Test sets.

4. Baselines

4.1. Problem Formulation

The task of toxicity detection from spoken utterances involves assigning a probability score, denoting the toxicity, to each utterance fed to the model. Formally, if \( X = \{x_1, \cdots, x_i, \cdots, x_N \} \) are the set of input utterances from the dataset \( D \) with \( N \) utterances, then \( Y = \{y_1, \cdots, y_i, \cdots, y_N \} \) are their corresponding labels where \( y_i \in \{0, 1\} \).

In the next 2 sections, we describe in detail our proposed 2-step and E2E baselines for DeToxy, the different components involved, and the training procedure involved in each step.

4.2. 2-step Baselines

Our 2-step approach consists of two primary components, an ASR component that produces transcriptions from spoken utterances, and a Sequence Classification Model for classifying the toxicity of the output transcript. For our ASR model, we resort to using wav2vec-2.0. Wav2vec-2.0 is built on the transformer architecture and has a Convolutional Neural Network (CNN) feature encoder layer, which encodes raw audio into speech frames. Wav2vec-2.0 follows the pre-training and fine-tuning paradigm wherein we first pre-train the model using a large unlabeled audio corpus in a self-supervised fashion solving a contrastive task as follows:

\[
\mathcal{L}_m = -\log \frac{\exp \left( \frac{\text{sim} (c_i, q_m) \cdot \kappa}{\kappa} \right)}{\sum_{q \neq q_m} \exp \left( \frac{\text{sim} (c_i, q) \cdot \kappa}{\kappa} \right)}
\]

where \( \mathcal{L}_m \) is the InfoNCE loss that wav2vec-2.0 tries to minimize, \( c_i \) is the context representation of a masked frame obtained from the last transformer layer, \( q_m \) is the quantized representation of obtained through a product quantization based quantization module and \( q \). Precisely, for every masked input frame which is input from the feature encoder module to the transformer-based context encoder, wav2vec-2.0 tries to identify the true quantized representation among a set of distractors. For more implementation-specific details about wav2vec-2.0, we refer our readers to [14]. For our 2-step approach, we

\(^1\)The list of trigger terms were obtained from https://hatebase.org/.
use the wav2vec-2.0_{large} architecture to push performance on ASR.

Post SSL pre-training, the model is then fine-tuned by mini-
mizing the Connectionist Temporal Classification (CTC) objective
function on labeled speech data. We do not use a language
model in our final setup to decode our utterances during infer-
ence, as we did not find significant differences in Word Error
Rate (WER) with it in our experiments, and also to keep the
setup simple.

Both self-supervised and supervised training paradigms for
ASR tend to get biased to the domain from where the data
originates [29] and therefore we make careful data considera-
tions for pre-training and fine-tuning our model. Most works
in the literature report result on a single benchmark dataset and
so we resort to multiple dataset pre-training but single dataset
fine-tuning. We choose to pre-train our model using a combi-
nation of Libri-Light, Common Voice (CV) [30], SwitchBoard
(SWBD) [31], and Fisher [32]. These 4 datasets cover a di-
verse set of domains and are well-suited to our downstream
ASR fine-tuning needs without creating a reasonable domain
shift [29]. For downstream ASR fine-tuning, we resort to 3 dif-
f erent dataset setups as follows: 1) To fit our model to the con-
versational domain, we fine-tune it on 300 hours of SWBD, 2)
for read speech (e.g. Audio Books) we resort to 960 hours of
LibriSpeech fine-tuning and finally 3) We also fine-tune on CV
which comprises of a majority of DeToxy.

For our text-based sequence-classification model, which is
the second step of our 2-step approach, we employ the
BERT_{base} architecture from the transformer family. Since its
inception, transformers have achieved SOTA on sequence clas-
sification tasks including toxicity or hate-speech classification
[33, 34]. Formally put, we tokenize each word in the sen-
tence and feed it as input through the transformer architec-
ture. Next, we utilize the hidden state embedding \( e \) corre-
sponding to the \([CLS]\) token, where \( e \) is obtained through a softmax activation function to get the prob-
ability distribution \( p_y \) of toxicity for each sentence. For train-
ing our model we minimize the Cross Entropy (C.E.) loss as
\[
\text{C.E.} = - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log (p_{ij}),
\]
where \( y_{ij} \) corresponds to the \( j^{th} \) gold annotated class from the \( i^{th} \) training instance in our dataset and \( p_{ij} \) is the probability score given by the model that the \( j^{th} \) instance belongs to the \( j^{th} \) class. Dur-
ing inference, we simply perform the \( \text{argmax} \) on \( p_i \) to find the
final class for each utterance. We fine-tune BERT_{base} on a pub-
licly available large-scale hate-speech classification dataset [35]
or on the relatively smaller gold toxicity annotations from the
DeToxy-B dataset.

### 4.3. End-to-End Baselines

For our E2E Baselines, we choose to experiment with different
feature extractors, where we build baselines employing either
low-level feature extractors (eg. Filter-Banks) or high-level
features from contextual feature encoder models learned through
SSL on unlabelled speech data. In the next 2 sub-sections, we
briefly describe both our approaches.

| Dataset       | Utterances | Toxic | Non-Toxic | Sp | Toxic | Emo | Sent | TL (hh:mm:ss) | Source          |
|---------------|------------|-------|-----------|----|-------|-----|------|----------------|-----------------|
| CMU-MOSEI     | 860        | 217   | 643       | -  | 233   | -   | 151  | 476            | 1:44:25         |
| CMU-MOSI      | 260        | 67    | 193       | -  | 77    | 49  | 134  | 0:18:03        |
| Common Voice  | 11,551     | 2,888 | 8,663     | 6,350 | 3,241 | 2,199 | 6,111 | 12:38:17       |
| IEMOCAP       | 1,090      | 274   | 816       | -  | 313   | 195 | 582  | 1:19:26        |
| LJ Speech     | 148        | 40    | 108       | 1  | 30    | 27  | 91   | 0:14:57        |
| MELD          | 565        | 142   | 423       | 67* | 145   | 136 | 284  | 0:31:05        |
| MSP-Improv    | 523        | 129   | 394       | 12 | 115   | 91  | 317  | 0:36:32        |
| MSP-Podcast   | 2,772      | 692   | 2,080     | 415* | 842   | 516 | 1,414 | 4:01:57        |
| Social-IQ     | 479        | 122   | 357       | -  | 147   | 93  | 239  | 0:36:40        |
| SwitchBoard   | 1,824      | 456   | 1,368     | -  | 500   | 345 | 979  | 2:28:57        |
| VCTK          | 44,583     | 50    | 44,533    | 110 | ✓     | ✓   | 70:22:28 | Newspaper Articles |

Total 2,336,177 5,077 2,331,100 - - - - 2770:04:15*

Table 1: Statistics for the DeToxy dataset. The * symbol signifies that the numbers reported are approximate.

| Dataset       | # Utterances | # Toxic | # Non-Toxic | # Sp | # Positive | # Neutral | # Negative | TL (hh:mm:ss) |
|---------------|--------------|---------|-------------|------|------------|-----------|------------|----------------|
| CMU-MOSEI     | 860          | 217     | 643         | -    | 233        | 151       | 476        | 1:44:25       |
| CMU-MOSI      | 260          | 67      | 193         | -    | 77         | 49        | 134        | 0:18:03       |
| Common Voice  | 11,551       | 2,888   | 8,663       | 6,350 | 3,241      | 2,199     | 6,111      | 12:38:17      |
| IEMOCAP       | 1,090        | 274     | 816         | -    | 313        | 195       | 582        | 1:19:26       |
| LJ Speech     | 148          | 40      | 108         | 1    | 30         | 27        | 91         | 0:14:57       |
| MELD          | 565          | 142     | 423         | 67*  | 145        | 136       | 284        | 0:31:05       |
| MSP-Improv    | 523          | 129     | 394         | 12   | 115        | 91        | 317        | 0:36:32       |
| MSP-Podcast   | 2,772        | 692     | 2,080       | 415* | 842        | 516       | 1,414      | 4:01:57       |
| Social-IQ     | 479          | 122     | 357         | -    | 147        | 93        | 239        | 0:36:40       |
| SwitchBoard   | 1,824        | 456     | 1,368       | -    | 500        | 345       | 979        | 2:28:57       |
| VCTK          | 44,583       | 50      | 44,533      | 110  | ✓          | ✓         | 70:22:28    | -              |

Total 20,271 5,077 15,194 - 5,701 3,835 10,735 24:39:10

Table 2: Statistics for the DeToxy-B dataset. In this table, we include the sentiment label statistics annotated by us for sampling DeToxy-B. Additionally, “Sp” and “TL” refers to the number of speakers and total length respectively.
In this paper, we introduce the first publicly available human-annotated dataset DeToxy for the task of toxic speech classification and propose two strong baselines for this task. Future work includes expanding the dataset using more naturally spoken utterances, exploring different modalities, and developing better neural architectures to solve this problem. Additionally, we would like to explore the factors and features learned by our E2E system that explain its superiority over the 2-step system.
8. References

[1] P. Fortuna, J. Soler, and L. Wanner, “Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets,” in *Proceedings of the 12th International Conference on Language Resources and Evaluation* (LREC), 2020, pp. 6786–6794.

[2] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, “Speech emotion classification using attention-based lstm,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[3] P. Tzirakis, J. Zhang, and B. W. Schuller, “End-to-end speech emotion recognition using deep neural networks,” in *IEEE ICASSP 2019*, pp. 5089–5093.

[4] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, “Incorporating end-to-end speech recognition models for sentiment analysis,” in *IEEE ICRRA 2019*, pp. 7976–7982.

[5] R. Guermazi, M. Hammani, and A. B. Hamadou, “Using a semi-automated keyword dictionary for improving violent web site filtering,” in *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*. IEEE, 2007, pp. 337–344.

[6] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.

[7] P. Badatjaya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in *Proceedings of the 26th international conference on World Wide Web Companion*, 2017, pp. 759–760.

[8] T. Davidson, D. Warmelsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in *ICWSM 2017*, vol. 11, 2017.

[9] A.-M. Founta, C. Djojavas, D. Chatzakou, I. Leonitiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Strivianos, and N. Kouttellis, “Large scale crowdsourcing and characterization of twitter abusive behavior,” 2018.

[10] J. A. Leite, D. Silva, K. Rontcheva, and C. Scarton, “Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis,” in *ACL-IJCNLP 2020*, Suzhou, China, pp. 914–924.

[11] M. Yousefi and D. Emmanouilidou, “Audio-based toxic language classification using self-attentive convolutional neural network,” in *EUSIPCO 2021*, pp. 11–15.

[12] C. S. Wu and U. Bhandary, “Detection of hate speech in videos using machine learning,” in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2020, pp. 585–590.

[13] A. Rana and S. Jha, “Emotion based hate speech detection using multimodal learning,” *arXiv preprint arXiv:2202.06218*, 2022.

[14] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” 2020.

[15] S. Ling and Y. Liu, “Decoor 2.0: Deep contextualized acoustic representations with vector quantization,” *arXiv preprint arXiv:2012.06659*, 2020.

[16] A. T. Liu, S.-w. Yang, P.-h. Chi, P.-c. Hsu, and H.-y. Lee, “Mock-ingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders,” in *IEEE ICASSP 2020*.

[17] I. Cohn, I. Laish, G. Beryozkin, G. Li, I. Shafран, I. Szpektor, T. Hartman, A. Hassidim, and Y. Matias, “Audio de-identification - a new entity recognition task,” in *NAACL 2019*, Minneapolis, Minnesota, pp. 197–204.

[18] H. Yadav, S. Ghosh, Y. Yu, and R. R. Shah, “End-to-end named entity recognition from english speech,” in *Interspeech 2021*, 2021, pp. 4268–4272.

[19] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, “Speech sentiment analysis via pre-trained features from end-to-end asr models,” in *IEEE ICASSP 2020*, pp. 7149–7153.

[20] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *IEEE ICASSP 2015*, pp. 5206–5210.

[23] J. S. Garofolo, “Timit acoustic phonetic continuous speech corpus,” *Linguistic Data Consortium*, 1993, 1993.

[24] F. Hernandez, V. Nguyen, S. Ghanay, N. Tomashenko, and Y. Es-teve, “Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation,” in *International conference on speech and computer*. Springer, 2018, pp. 198–208.

[25] M. Z. Boito, W. N. Havard, M. Garnerin, É. L. Ferrand, and L. Besacier, “Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible,” *arXiv preprint arXiv:1907.12895*, 2019.

[26] M. S. Grover, P. Bamdev, Y. Kumar, M. Hama, and R. R. Shah, “audiine: A modern annotation tool for audio and speech,” *CoRR*, vol. abs/2006.05236, 2020.

[27] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The risk of racial bias in hate speech detection,” in *ACL 2019*, pp. 1668–1678.

[28] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, “Handling bias in toxic speech detection: A survey,” *arXiv preprint arXiv:2202.00126*, 2022.

[29] W.-N. Hsu, A. Srimar, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, “Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training,” in *Interspeech 2021*, 2021, pp. 721–725.

[30] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.

[31] J. Godfrey and E. Holliman, “Switchboard-1 release 2 ldc97s62,” *LDC*, 1993.

[32] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, “Fisher english training speech part 1 transcripts,” *Philadelphia: Linguistic Data Consortium*, 2004.

[33] M. Malik, A. Aggrawal, and D. K. Vishwakarma, “Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks,” in *IEEE ICCMC 2021*, pp. 1254–1259.

[34] Z. Zhao, Z. Zhang, and F. Hopfgartner, “A comparative study of using pre-trained language models for toxic comment classification,” in *Companion Proceedings of the Web Conference*, 2021, pp. 500–507.

[35] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, “Nuanced metrics for measuring unintended bias with real data for text classification,” *CoRR*, vol. abs/1903.04561, 2019.

[36] P. Flandrin, G. Rilling, and P. Goncalves, “Empirical mode decomposition as a filter bank,” *IEEE signal processing letters*, vol. 11, no. 2, pp. 112–114, 2004.

[37] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, “What all do audio transformer models hear? probing acoustic representations for language delivery and its structure,” *arXiv preprint arXiv:2101.00387*, 2021.