MuTox: Universal MUltilingual Audio-based TOXicity Dataset and Zero-shot Detector

Marta R. Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, Carleigh Wood
FAIR, Meta
{costajussa,mfcoria,mortimer,daviddale,prangthiphansanti, ekalbassi,alexmourachko,chrisropers,carleighwood}@meta.com

Abstract

Research in toxicity detection in natural language processing for the speech modality (audio-based) is quite limited, particularly for languages other than English. To address these limitations and lay the groundwork for truly multilingual audio-based toxicity detection, we introduce MuTox, the first highly multilingual audio-based dataset with toxicity labels which covers 14 different linguistic families. The dataset comprises 20,000 audio utterances for English and Spanish, and 4,000 for the other 28 languages. To demonstrate the quality of this dataset, we trained the MuTox audio-based toxicity classifier, which enables zero-shot toxicity detection across a wide range of languages. This classifier performs on par with existing text-based trainable classifiers, while expanding the language coverage more than tenfold. When compared to a wordlist-based classifier that covers a similar number of languages, MuTox improves F1-Score by an average of 100%. This significant improvement underscores the potential of MuTox in advancing the field of audio-based toxicity detection.

Warning: This article includes examples of language that can be considered offensive or upsetting.

1 Introduction

Text toxicity detection has been largely explored for different tasks in Natural Language Processing (NLP) (Kaggle, 2018). Wordlist-based toxicity classifiers—e.g., ETOX (Costa-jussà et al., 2023)—scale well to a large number of languages (NLLB Team et al., 2022) and context-based classifiers are able to detect beyond lexical toxicity with tools such as DETOXIFY[1].

When exploring audio-based toxicity detection, there are either cascaded systems which extend text toxicity detection with speech recognition (Seamless Communication et al., 2023a); or end-to-end audio-based toxicity classification (Ghosh et al., 2021) which provides an English dataset together with end-to-end toxicity detection results. This work shows that gains of English text-less audio-based classifiers over text-based classifiers are specially relevant when applied to out-of-domain, coherently with the previous study on a non-disclosed dataset (Yousefi and Emmanouilidou, 2021).

In this paper, we go far beyond existing research in audio-based toxicity detection by providing the first highly multilingual audio-based toxicity annotated dataset (MuTox dataset, 30 languages, see Table 6 in appendix A) together with the first text-less massive multilingual metric (MuTox classifier, 100+ languages). Note that multilinguality for audio-based toxicity detection becomes even more crucial for the task of added toxicity in the context of multimodal and multilingual translation, where the case of adding or deleting toxicity may be considered as a critical error (Seamless Communication et al., 2023a).

In particular, the main contributions of this paper are: providing guidelines for audio-based toxicity annotation (section 3); releasing the first highly multilingual audio-based toxicity dataset and benchmark with human annotations for 30 languages (section 4); analyzing the performance of text-based classifiers when applied to audio-based toxicity detection (section 6); proposing MuTox, a massively multilingual audio-based toxicity classifier (section 5). Our results can be summarized:

- When compared to the strongest performing systems, which are composed of speech recognition plus trainable text-based toxicity detectors, MuTox performs on par, while offering more than 10 times the language coverage [2].

[1] https://github.com/unitaryai/detoxify

[2] We want to clarify that while MuTox dataset evaluates 30 languages, the MuTox classifier relies on SONAR (Duquenne et al., 2023) embeddings. As of March 2024, SONAR encoders are available for 57 languages in speech and 200 in text. Thanks to this architecture, by design, the MuTox classifier

Findings of the Association for Computational Linguistics ACL 2024, pages 5725–5734
August 11-16, 2024 ©2024 Association for Computational Linguistics

- When compared to systems with the highest coverage, which are composed of speech recognition plus wordlists toxicity detectors, MuTox improves F1-Score by an average of 100%.

These results highlight the effectiveness of MuTox in multilingual audio-based toxicity detection, demonstrating its potential to significantly advance the field.

2 Background

Text-based toxicity classifiers. In this work, we use ETOX (Costa-jussà et al., 2023) and DETOXIFY as primary text-based toxicity detectors. ETOX is chosen for offering the widest language coverage in toxicity detection. DETOXIFY is chosen for being one of the available tools with the highest performance on several JigSaw benchmarks with a single model (Thewlis et al., 2021).

ETOX is an open-source[3] wordlist-based classifier covering 200 languages. While this metric has several advantages - namely, it covers a massive number of languages and is highly transparent - it has other limitations such as only detecting lexical toxicity and not distinguishing polysemous words which may be toxic in some contexts and not others.

DETOXIFY [1] is a text-based toxicity classifier trained mainly in JigSaw (Kaggle, 2018) which is freely available in 7 languages (English, French, Italian, Portuguese, Russian, Spanish, Turkish).

Datasets A well-known resource for textual toxicity detection is the JigSaw dataset (Kaggle, 2018), which consists of a large number of Wikipedia comments which have been labeled by human raters for toxic meaning. This dataset has been used for several Kaggle competitions covering a broad range of tasks, from detecting types of toxicity to analyzing bias in toxicity detection systems. There are also other related datasets e.g. (Röttger et al., 2022).

There are extremely few speech datasets with toxicity labels. Recently, ADIMA covers multilingual profanity detection audio dataset in 10 Indic languages (Gupta et al., 2022). Relatedly, Detoxy (Ghosh et al., 2021) estimated the amount of toxicity for several English spoken datasets using text-based classifiers. In this work, we follow a similar approach for pre-selecting data for annotation to maximize chances of the annotators confirming toxicity. We used two spoken datasets which cover diverse domains and are highly multilingual: COMMONVOICE (Ardila et al., 2020) and SEAMLESSALIGN (Seamless Communication et al., 2023b).

COMMONVOICE is a massively-multilingual collection of transcribed speech intended for speech technology research and development. SEAMLESSALIGN is an automatically collected pairs of natural speech from raw web corpora through parallel data mining following methodology described in (Seamless Communication et al., 2023a). Additionally, for English and Spanish, we use SEAMLESSALIGNEXPRESSIVE which extends the process used for SEAMLESSALIGN to create a large collection of multilingual speech/speech and speech/text pairs, aligned not only in meaning but also expressivity. Text and audio sources are identical to SEAMLESSALIGN. The modified mining algorithm is described in section 4 from (Seamless Communication et al., 2023b).

Experimental Task. Toxicity detection in natural language processing is the task of assigning a toxicity label to a speech or text utterance.

3 Annotation Guidelines

This section reports the detailed guidelines that we provide to annotators to detect toxic content in audio speech at the level of single utterances. This toxicity could be due to aspects of lexical semantics or of perlocutionary effects. For this we provide annotators with the following definitions:

- **Utterance** refers to a unit of audio speech that is comparable to what a sentence is for writing.

- **Lexical semantics** refers to meaning clearly attached to a particular word or phrase, as opposed to meaning that can be conveyed by other aspects of audio speech than words, such as by tone of voice.

- **Perlocutionary** refers to the effect that an utterance has on the interlocutor or listener (as opposed to the locutionary or illocutionary aspects of the same utterance). For example, if Interlocutor A says "it would be a shame if

provides zero-shot toxicity detection for all of them. This is more than 8 times language coverage in speech and more than 25 times in text.

[3] https://github.com/facebookresearch/seamless_communication/tree/main/src/seamless_communication/toxicity

something happened to your family" to Interlocutor B, the utterance has a locutionary aspect (its literal meaning; i.e. something could happen to your family, and that would be bad), an illocutionary aspect (the thinly veiled threat against Interlocutor B's family; i.e. if you don't do as I say, something bad will happen to your family), and a perlocutionary effect (the fear it causes to Interlocutor B).

What qualifies as toxicity for this task? The notion of toxicity can be elusive because of its inherent subjectivity; i.e. it greatly depends on the lived experience of the person who is on the receiving end of an utterance. For the task at hand, we arbitrarily considered the following cases as toxic:

- **Profanities** include slurs and language that is regarded as obscene, repulsive, or excessively vulgar, as well as scatological. Examples of profanities in English include words such as *shit*, *asshole*, *fucking*, etc.

- **Hate speech** is language that is used to demean, disparage, belittle, or insult groups of people. Hate speech in English includes words and expressions such as *women are sluts*, *men are trash*, *wetbacks*, *towelheads*, etc.

- **Pornographic language** is language that refers to sexual acts or refers in a vulgar way to body parts typically associated with sexuality. Examples of pornographic language include words or expressions such as *suck my dick*, *cumshot*, etc.

- **Physical violence or bullying language** is language that is used to bully, threaten, silence individuals. Examples of such language include words or expression such as *son of a bitch*, *shut the fuck up*, *retard*, etc.

What does not qualify as toxicity for this task? Annotators were instructed to refrain from labeling as toxic the following types of language:

- **Common innocuous slang** is language that is particularly colloquial but isn't offensive; e.g., using the word *cops* to refer to police officers.

- **Appearance of hate** shows up as language that conveys hate but isn't directed to other human beings; e.g., *I hate this movie!*

- **Appearance of obscenity** shows up as language that sounds similar to profanities or pornographic language but isn't directed towards people; e.g., *school sucks!*

To complete the task, annotators were asked to answer the below questions:

1. Does the utterance contain toxicity?

2. If you answered positively to Question 1, is the toxicity related to:

   (a) Specific words or phrases? If so, please specify the word or phrase.

   (b) Specific perlocutionary effects? If so, please specify which effect among the following options:
      - aggressive raising of the voice
      - aggressive tone
      - (veiled) threat

To help annotators reach a faster determination as to which words or phrases could qualify as toxic in this exercise, we also pointed them to the publicly available Toxicity-200[4] word list repository.

4 MuTox DataSet Description

MuTox is composed of 30 languages. Two languages, English and Spanish, contain a larger amount of annotated data (20k utterances/sentences each language), while the rest of languages (see section 4.2), contain a smaller amount of annotated data (4k utterances/sentences each language).

4.1 Annotation for English and Spanish

Preliminary selection. We used speech transcriptions of datasets and a text toxicity classifier. We use mainly highly multilingual datasets: COMMONVOICE[5] and SEAMLESSALIGN (Seamless Communication et al., 2023b), in addition to the English/Spanish SEAMLESSALIGNEXPRESSIVE dataset, presented in Section 2. In these datasets we use the text-based toxicity classifier DETOXIFY, presented in section 2, to detect toxicity in the transcribed text.

We screen audio files by length between 2 and 8 seconds for reasons that relate to both semantic and cognitive loads. On the one hand, annotators reported that it was particularly difficult to

[4] https://tinyurl.com/NLLB200TWL
[5] https://commonvoice.mozilla.org/

assert the meaning of very short utterances (under 2 seconds). On the other hand, longer utterances (over 8 seconds) can have too much cognitive load or too much information to annotate. For pre-selecting toxicity samples, we perform sampling across DETOXIFY toxicity categories.For pre-selecting clean samples, we use cases where all toxicity scores fall below 0.5.

Data statistics. Annotation results are reported in Table 1. The percentages of confirmed toxic samples are 16 % and 19% for English and Spanish, respectively. Figure 1 reports the types of toxicity for each language using annotated categories.The proportion of toxicity increases with the toxicity quantile (see Figure 5 in appendix B) . The correlation coefficient between the toxicity in the annotation with DETOXIFY threshold is 0.6.

	English		Spanish	
	utterances	hours	utterances	hours
Total	20K	21	20K	22.2
Cannot say	547	0.52	391	0.41
No-Toxicity	16210	17.1	15709	17.5
Toxicity	3243	3.42	3900	4.24

Table 1: Results on toxicity annotated results for English and Spanish.

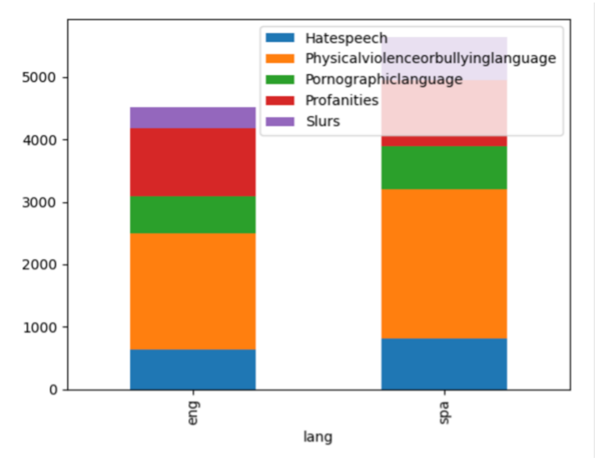

Figure 1: Amount of toxicity per toxic category proposed in this paper.

Dataset splits. The final MuTox dataset includes the 20k annotated English and Spanish corpora from Table 1. We supplement this with the English COMMONVOICE annotated data released in Detoxy (Ghosh et al., 2021). We split the data in subsets of training, dev, devtest and 3k test with stratified samples of types of toxicity and corpora. Detailed statistics are given in Table 2.

Subset	Language	Modality	Dataset	Size	Toxicity
Train	Eng	Speech	MuTox	13617	2270
			Detoxy	9818	2455
		Text	Jigsaw	21924	2928
	Spa	Speech	MuTox	13726	2730
	HP			1212-1498	39-363
Devtest	Eng	Speech	MuTox	973	162
	Spa			981	195
	HP			203-250	7-60
Devtest	Eng	Speech	MuTox	1945	324
	Spa			1960	390
	HP			606-749	20-182
Test	Eng	Speech	MuTox	2918	486
	Spa			2918	486
	HP			1213-1499	38-362

Table 2: Audio speech utterances specified by dataset subset. HP languages footnote. MuTox is our new labelled data that has been annotated in this work. Additionally, we used data from Detoxy (Ghosh et al., 2021) and JigSaw (Kaggle, 2018).

4.2 Annotation for 28 additional languages

Preliminary selection. Based on annotated data in English and Spanish (section 4.1), we devise new criteria to select 4k sentences across 28 extra languages which we consider high-priority (HP) for the context of this work and related previous projects (Seamless Communication et al., 2023a). The purpose of this extension is to form a highly multilingual benchmark for audio-based toxicity classification.

Languages. Modern Standard Arabic, Bengali, Bulgarian, Catalan, Czech, Danish, Dutch, Estonian, Finish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Mandarin Chinese, Persian, Polish, Portuguese, Russian, Slovak, Swahili, Tagalog, Turkish, Urdu, Vietnamese (see table 6).

Classifiers. We use a text toxicity classifier that covers all these languages (ETOX) in combination with the English toxicity classifier used in section 4.1 (DETOXIFY).

Datasets. We use SEAMLESSALIGN which has parallel data eng–X for all languages of interest.

Methodology. We aim at finding 2,500 sentences with the n-largest toxicity scores, and randomly sample the dataset to complete the 4k set. Motivated by results from section 4.1, sentences are selected as follows:

We include all the samples positive per ETOX and DETOXIFY scores higher than 0.8 (i.e. intersection). This results in a small number of samples, ranging from more than 2k (pol) to 0 (cmn,jpn). Then, we include ETOX detections (ignoring the DETOXIFY threshold) but allow only for a maximum of 200 sentences for each ETOX token and add a maximum of 1,000 ETOX sentences. We set these maxima so as to ensure toxicity diversity in the selection. Finally, we include samples detected with DETOXIFY: we use DETOXIFY in Italian, Portuguese, Turkish, Russian and French (which are supported by the tool). For the rest of non-supported languages, we resort to English DETOXIFY in the parallel text. The effectiveness of this method is supported by the following evidence: when using DETOXIFY in English with a threshold of 0.80 on the corpus from section 4.1, we observe a total toxicity of 0.25 in Spanish.

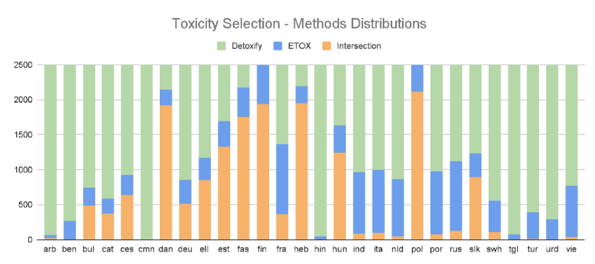

Figure 2: Toxicity selection distribution per language (x-axis) and for each method: DETOXIFY, ETOX and their intersection.

Dataset statistics and splits. Figure 2 shows the results of this selection, grouped by the methodology used (1 - intersection, 2 - ETOX, 3 - DETOXIFY). Table 2 shows the dataset splits. The final dataset of 28 languages includes 83 hours.

5 MuTox audio-based toxicity Classifier

Methodology: MuTox Classifier. We feed our toxicity classifier, MuTox, with both audio-based and text-based toxicity-labeled data. The audio-based toxicity classifier depicted on Figure 3 follows a simple architecture consisting of an encoder, turning input text or audio speech into a fixed-size representation vector, and a binary classifier composed of three simple feed-forward layers.

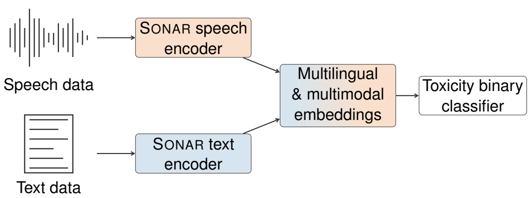

Figure 3: MuTox toxicity classifier diagram

MuTox Implementation. We use multimodal and multilingual SONAR encoders from (Duquenne et al., 2023). This choice is motivated by the broad language coverage at the time of this study (English, Spanish, and HP languages) and the zero-shot capabilities of SONAR: once trained on a set of languages, the classifier head can be used on top of any compatible SONAR encoder handling another language. For the classifier, we use variable input sizes for the 3 feedforward layers (1024, 512, and 128). We train with a Binary Cross Entropy loss with logits and Adam optimizer with an initial learning rate of 0.001. In order to compare zero-shot (ZS) vs supervised performance, we train a first classifier with English and Spanish training data only and test on HP languages. We then train a second classifier on all training data available and report results on the same HP languages. The number of parameters of our models is around 600k. Results are listed in Table 4. Throughout our experiments, we use WHISPER-LARGE-V2 to generate transcriptions.

6 Experiments and Results

Experimental Framework. We evaluate MuTox (text and speech) on Devtest and Test from Table 2. We compare the performance against ASR-DETOXIFY (hereinafter, DETOXIFY for simplicity) using the available tool[1] and ASR-ETOX (hereinafter, ETOX) using the available tool[3]. Speech recognition is done with WHISPER-LARGE-V2. We report Area Under the Curve (AUC), Precision, Recall and F1-Score across toxicity detection.

Correlation across classifiers We computed Pearson correlation across classifiers (see results in Table 3). We observe that correlation is higher between MuTox and DETOXIFY than MuTox and ETOX or even DETOXIFY and ETOX. In future work, we would like to do a manual analysis of the differences across classifiers.

	ETOX	DETOXIFY	MuTox	ASR-MuTox
ETOX	1	0.26	0.06	0.12
DETOXIFY	0.26	1	0.46	0.49
MuTox	0.06	0.46	1	0.79
ASR-MuTox	0.12	0.49	0.79	1

Table 3: Pearson correlation across toxicity classifiers.

Improvements of MuTox compared to one of the Strongest Quality Text-based Toxicity Classifier Table 4 compares MuTox and Detoxify (for available languages) in terms of AUC. When comparing the 7 languages covered by DETOXIFY, MuTox trained on all languages shows on par results with DETOXIFY. MuTox however, scales to 10 times more languages than DETOXIFY.

Improvements of MuTox compared to the Text-based Toxicity Classifier with the Largest Coverage Table 5 compares MuTox and ETOX in terms of recall at fixed precision. ASR-MuTox with a fixed precision of $max(\text{ASR-ETOX}, 0.3)$ (meaning 0.3 average precision)[6] improves F1-Score over ETOX both in devtest and test from 0.19 to 0.38. We observe variations in recall rates for several HP languages (e.g., cat, heb, ind, rus, swh) between development and test datasets. This is due to the fact that we have very low representation of toxicity in devtest and test for those languages. That is why, it can be expected to have different results.

MuTox Configurations Supervised MuTox slightly improves zero-shot MuTox by 2% on average. When comparing MuTox and ASR-MuTox averaging over 30 languages, results are almost comparable for zero-shot, but ASR-Mutox is better than MuTox in supervised setting. While it is unclear why MuTox vs. ASR-MuTox show different results depending on the language, we could hypothesize that the imbalances in the complexities of pronunciation/writing in various languages lead to variations in transcription quality.

Performance across toxicity categories Figure 4 reports the model performances by toxicity categories. We note that DETOXIFY and MuTox have less variance (<0.001) across categories than ETOX (>0.01). ETOX performs well on Profanities (recall>0.8), since those can easily be identified with explicit wordlists. It struggles however with more complex/implicit types of toxicity such as Hate Speech or Physical Violence.

7 Conclusions

In this paper, we introduced MuTox, a highly multilingual dataset and a massive multilingual toxicity detector that lays the groundwork for the largely unexplored task of multilingual audio-based toxicity detection. Our MuTox dataset enables benchmarking of multilingual audio-based toxicity detection across 30 languages. Our MuTox classifier, compared to cascade tools (speech recognition followed by text-based toxicity classifiers) that have similar coverage (more than hundreds of languages), shows superior performance for all evaluated tasks.

Data[7] and code[8] with their corresponding data and model cards are freely available for further research and development. In future work, we primarily aim to explore more complex architectures

[6] Except for swh and cat where precision is of $max(\text{ASR-ETOX}, 0.1)$

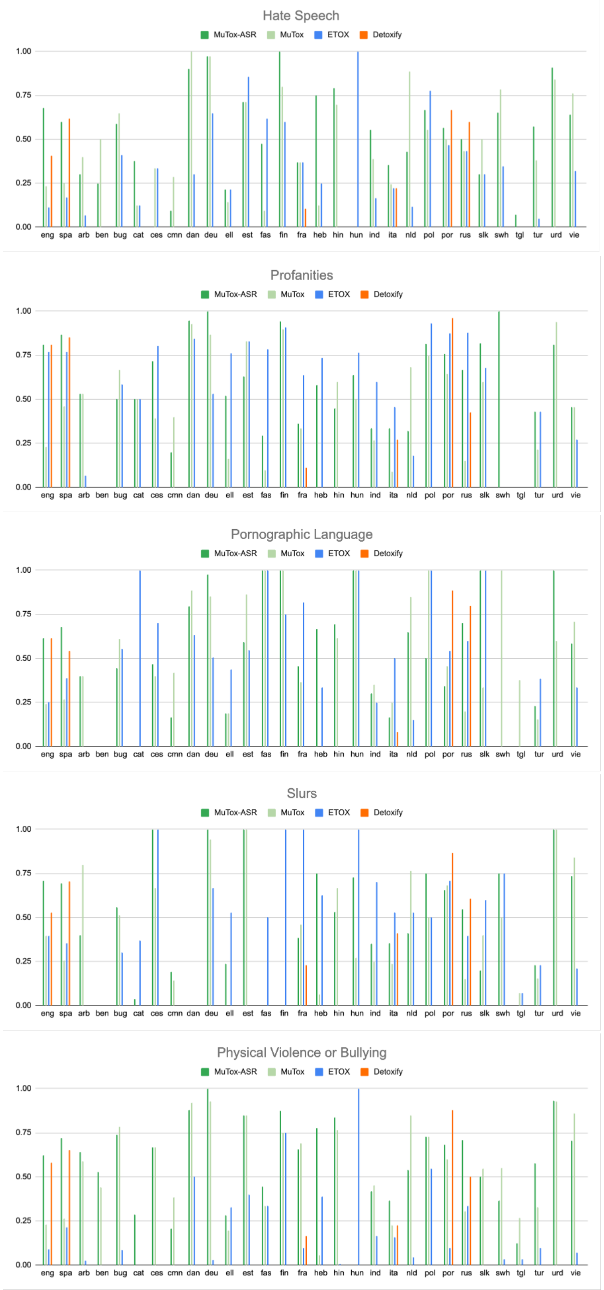

Figure 4: Recall per toxicity category at fixed precision of $max(\text{ETOX}, 0.3)$.

[7] Will be revealed after the anonymized review.
[8] Will be revealed upon anonymized review.

lang	Devtest (D)		Test (T)		MuTox$_{ZS}$		ASR-MuTox$_{ZS}$		MuTox		ASR-MuTox		DETOXIFY	
	Size	Tox	Size	Tox	D	T	D	T	D	T	D	T	D	T
eng	1945	325	2918	486	-	-	-	-	0.61	0.63	0.72	0.75	0.68	0.71
spa	1960	390	2941	585	-	-	-	-	0.63	0.65	0.72	0.73	0.69	0.71
arb	720	63	1440	125	0.73	0.77	0.76	0.79	0.84	0.82	0.81	0.83	-	-
ben	738	20	1474	38	0.88	0.87	0.88	0.81	0.84	0.88	0.90	0.89	-	-
bul	675	95	1350	191	0.73	0.71	0.75	0.76	0.80	0.79	0.81	0.78	-	-
cat	606	22	1213	43	0.80	0.71	0.80	0.77	0.77	0.69	0.79	0.75	-	-
ces	744	43	1488	85	0.72	0.85	0.75	0.85	0.73	0.85	0.75	0.86	-	-
cmn	744	70	1487	140	0.78	0.79	0.76	0.77	0.82	0.78	0.80	0.74	-	-
dan	737	133	1481	266	0.82	0.82	0.86	0.83	0.83	0.82	0.90	0.85	-	-
deu	744	182	1486	362	0.77	0.80	0.78	0.79	0.83	0.85	0.85	0.86	-	-
ell	737	59	1474	118	0.71	0.70	0.79	0.78	0.71	0.68	0.80	0.79	-	-
est	684	86	1369	173	0.79	0.77	0.81	0.78	0.87	0.85	0.84	0.83	-	-
fas	743	40	1487	81	0.85	0.83	0.79	0.81	0.84	0.86	0.84	0.84	-	-
fin	738	87	1476	173	0.89	0.88	0.92	0.90	0.90	0.91	0.93	0.93	-	-
fra	738	62	1477	124	0.80	0.80	0.76	0.78	0.81	0.80	0.77	0.80	0.79	0.83
heb	725	30	1450	60	0.73	0.71	0.72	0.76	0.70	0.74	0.71	0.75	-	-
hin	699	82	1399	166	0.77	0.77	0.77	0.79	0.80	0.83	0.81	0.84	-	-
hun	749	88	1499	177	0.72	0.77	0.77	0.80	0.77	0.76	0.76	0.81	-	-
ind	745	72	1490	143	0.70	0.73	0.67	0.69	0.74	0.76	0.73	0.75	-	-
ita	693	98	1385	197	0.61	0.60	0.63	0.60	0.64	0.66	0.62	0.69	0.79	0.62
nld	729	87	1458	174	0.82	0.82	0.70	0.72	0.86	0.87	0.77	0.76	-	-
pol	741	64	1484	129	0.87	0.84	0.88	0.86	0.84	0.86	0.92	0.88	-	-
por	724	109	1449	218	0.78	0.76	0.80	0.78	0.79	0.75	0.78	0.76	0.81	0.83
rus	741	81	1481	161	0.79	0.76	0.80	0.82	0.81	0.75	0.83	0.83	0.84	0.81
slk	741	45	1482	90	0.82	0.84	0.84	0.84	0.81	0.87	0.83	0.87	-	-
swh	729	45	1458	89	0.69	0.69	0.66	0.66	0.70	0.68	0.69	0.67	-	-
tgl	736	43	1473	88	0.74	0.73	0.69	0.70	0.77	0.77	0.76	0.74	-	-
tur	740	53	1480	107	0.80	0.74	0.78	0.73	0.82	0.78	0.79	0.81	0.80	0.82
urd	741	139	1486	278	0.78	0.79	0.76	0.78	0.81	0.83	0.79	0.84	-	-
vie	738	93	1477	185	0.78	0.81	0.75	0.76	0.81	0.84	0.81	0.80	-	-
avg7	-	-	-	-	0.75	0.73	0.75	0.74	0.73	0.72	0.75	**0.77**	**0.77**	0.76
avg	-	-	-	-	0.77	0.77	0.77	0.78	0.78	0.79	**0.79**	**0.80**	-	-

Table 4: Toxicity detection AUC results of MuTox vs DETOXIFY. We show different MuTox configurations: ZS, trained only with English and Spanish; supervised, trained on English, Spanish, and HP languages; in audio speech (MuTox) or text (ASR-MuTox). The best results are bolded.

for training MuTox detector. Additionally, we want to evaluate MuTox's performance on the task of added toxicity. We also intend to use it to analyze and improve ETOX wordlists, as detailed in the appendix C. This ongoing work will continue to advance the field of multilingual audio-based toxicity detection.

Ethical Considerations and Limitations

Annotators. Annotations were provided by professional annotators, who were informed of the nature of the content to be annotated and were given the opportunity to opt out. We understand that the perception of toxicity is subjective, and therefore varies greatly from individual to individual, and from group of individuals to group of individuals. Even within groups of any size, the probability is high that toxicity will be perceived differently from one individual to the next. However, we involved a large pool of annotators with native level in each of the MuTox languages. While our guidelines specify the types of toxicity annotators should look out for, it is ultimately down to annotator judgment to interpret whether a term or phrase is toxic, given the context, and given their personal interpretation.

Bias in pre-selected sentences. Our dataset may be biased towards text-based toxicity because we pre-selected annotations using this criterion. The use of different text classifiers, as well as a variety of thresholds, mitigates this potential bias. Moreover, the data is sourced from (e.g. COMMONVOICE), and it represents a broad and geographically diverse sample.

Unintended bias. Our evaluation does not cover unintended biases[9], which we intend to cover in

[9] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/

lang	ETOX				MuTox$_{ZS}$		ASR-MuTox$_{ZS}$		MuTox		ASR-MuTox	
	Devtest		Test		Devtest	Test	Devtest	Test	Devtest	Test	Devtest	Test
	Prec	Rec	Prec	Rec	Recall (Rec)							
eng	0.39	0.30	0.40	0.31	-	-	-	-	0.18	0.23	0.58	0.67
spa	0.40	0.33	0.41	0.33	-	-	-	-	0.19	0.32	0.73	0.74
arb	0.09	0.02	0.21	0.04	0.16	0.26	0.17	0.15	0.60	0.54	0.44	0.53
ben	0.02	0.05	0.00	0.00	0.35	0.11	0.10	0.03	0.30	0.45	0.45	0.50
bul	0.19	0.29	0.17	0.21	0.22	0.32	0.53	0.65	0.71	0.69	0.74	0.65
cat	0.09	0.36	0.08	0.30	0.18	0.07	0.18	0.19	0.59	0.05	0.77	0.16
ces	0.12	0.44	0.18	0.73	0.19	0.45	0.30	0.59	0.21	0.41	0.30	0.61
cmn	0.00	0.00	0.00	0.00	0.14	0.37	0.27	0.17	0.43	0.35	0.41	0.16
dan	0.26	0.77	0.23	0.68	0.92	0.91	0.93	0.90	0.89	0.92	0.96	0.88
deu	0.42	0.36	0.43	0.36	0.81	0.88	0.95	0.99	0.86	0.90	0.98	0.99
ell	0.11	0.39	0.12	0.41	0.02	0.16	0.03	0.36	0.19	0.18	0.44	0.31
est	0.15	0.49	0.15	0.51	0.57	0.54	0.71	0.63	0.86	0.84	0.80	0.76
fas	0.07	0.75	0.07	0.68	0.15	0.23	0.18	0.33	0.33	0.14	0.50	0.36
fin	0.17	0.90	0.17	0.89	0.91	0.86	0.94	0.94	0.87	0.88	0.94	0.94
fra	0.11	0.42	0.10	0.39	0.26	0.31	0.08	0.14	0.40	0.53	0.48	0.50
heb	0.05	0.63	0.04	0.53	0.10	0.05	0.07	0.10	0.13	0.05	0.03	0.70
hin	0.18	0.02	0.06	0.01	0.56	0.36	0.54	0.33	0.63	0.72	0.66	0.79
hun	0.20	0.73	0.22	0.79	0.34	0.51	0.57	0.57	0.55	0.49	0.56	0.64
ind	0.10	0.22	0.11	0.28	0.18	0.19	0.07	0.17	0.15	0.41	0.32	0.40
ita	0.20	0.38	0.16	0.27	0.10	0.05	0.28	0.18	0.16	0.21	0.31	0.35
nld	0.08	0.15	0.06	0.11	0.77	0.73	0.21	0.21	0.78	0.84	0.61	0.52
pol	0.12	0.89	0.12	0.88	0.86	0.60	0.83	0.73	0.67	0.73	0.86	0.79
por	0.28	0.43	0.30	0.50	0.68	0.69	0.75	0.77	0.74	0.59	0.66	0.63
rus	0.17	0.44	0.19	0.47	0.54	0.42	0.47	0.52	0.65	0.27	0.69	0.63
slk	0.14	0.73	0.09	0.48	0.49	0.41	0.36	0.52	0.47	0.56	0.49	0.66
swh	0.08	0.18	0.07	0.15	0.04	0.02	0.64	0.06	0.02	0.61	0.60	0.46
tgl	0.27	0.09	0.12	0.03	0.19	0.05	0.05	0.06	0.35	0.23	0.07	0.09
tur	0.06	0.08	0.12	0.16	0.13	0.93	0.15	0.12	0.34	0.28	0.28	0.47
urd	0.00	0.00	0.00	0.00	0.87	0.87	0.81	0.85	0.88	0.88	0.83	0.92
vie	0.12	0.16	0.10	0.15	0.54	0.69	0.46	0.61	0.68	0.80	0.73	0.68
avg	0.17	0.36	0.17	0.35	0.40	0.43	0.42	0.42	0.47	0.48	**0.57**	**0.58**

Table 5: Toxicity detection precision and recall results. MuTox recall at the precision of $max(\text{ETOX}, 0.3)$ vs. ETOX. The best results are bolded.

future work.

Acknowledgements

The authors want to thank Samuel Bell for enriching discussions and his valuable feedback.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Marta R. Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale.

Paul-Ambroise Duquenne, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk. 2023. Expressivity-aware SONAR: speech decoder.

Sreyan Ghosh, Samden Lepcha, Sahni Sakshi, Rajiv Ratn Shah, and Sharma Umesh. 2021. Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances. In *Interspeech*.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. 2022. Adima: Abuse detection in multilingual audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6172–6176.

Kaggle. 2018. Toxic comment classification challenge. `https://www.kaggle.com/c/jigsaw-toxic-comment-\classification-challenge`. Accessed: 2022-05-03.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023a. Seamlessm4t-massively multilingual & multimodal machine translation.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023b. Seamless: Multilingual expressive and streaming speech translation.

James Thewlis, Sasha Haco, and Laura Hanu. 2021. How ai is learning to identify toxic online content.

Midia Yousefi and Dimitra Emmanouilidou. 2021. Audio-based toxic language classification using self-attentive convolutional neural network. In *29th European Signal Processing Conference, EUSIPCO 2021, Dublin, Ireland, August 23-27, 2021*, pages 11–15. IEEE.

A Languages

arb	Modern Standard Arabic	Mesopotamian Arabic
ben	Bengali	Indo-Aryan
bul	Bulgarian	Balto-Slavic
cat	Catalan	Romance
ces	Czech	Balto-Slavic
cmn	Mandarin Chinese	Sino-Tibetan
dan	Danish	Germanic
deu	German	Germanic
ell	Greek	Hellenic
eng	English	Germanic
est	Estonian	Uralic
fas	Western Persian	Iranian
fin	Finnish	Uralic
fra	French	Romance
heb	Hebrew	Afro-Asiatic
hin	Hindi	Indo-Aryan
hun	Hungarian	Uralic
ind	Indonesian	Austronesian
ita	Italian	Romance
nld	Dutch	Germanic
pol	Polish	Romance
por	Portuguese	Romance
rus	Russian	Balto-Slavic
spa	Spanish	Romance
slk	Slovak	Balto-Slavic
swh	Swahili	Atlantic-Congo
tgl	Tagalog	Austronesian
tur	Turkish	Turkic
urd	Urdu	Indo-Aryan
vie	Vietnamese	Austroasiatic

Table 6: The 30 languages covered in this work.

B Annotations Details

Figure 5 shows the percentage of toxicity obtained with the annotation (y-axis) by the quantile of toxicity (x-axis) in the text data.

C Wordlists analysis.

We report an analysis on the toxic words detected with ETOX in order to understand the limitations of the Toxicity-200 word-lists.

For English we have a total of 110 different toxic tokens detected and 59 tokens show reasonable precision (>0.4). On the one hand, the worst

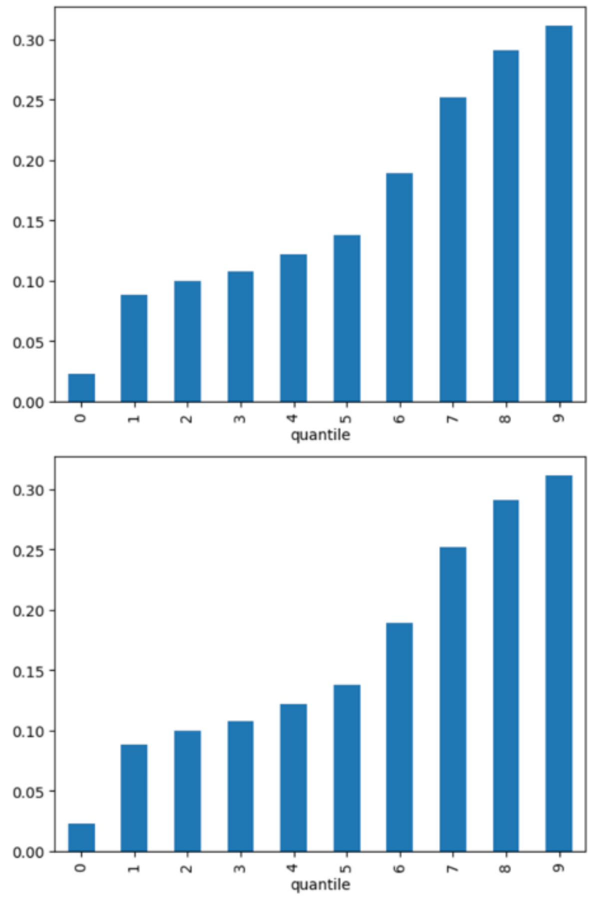

Figure 5: Percentage (y-axis) of toxicity in the audio speech dataset in English (top) and in Spanish (bottom) per toxicity quantile (x-axis) in the text toxicity classification.

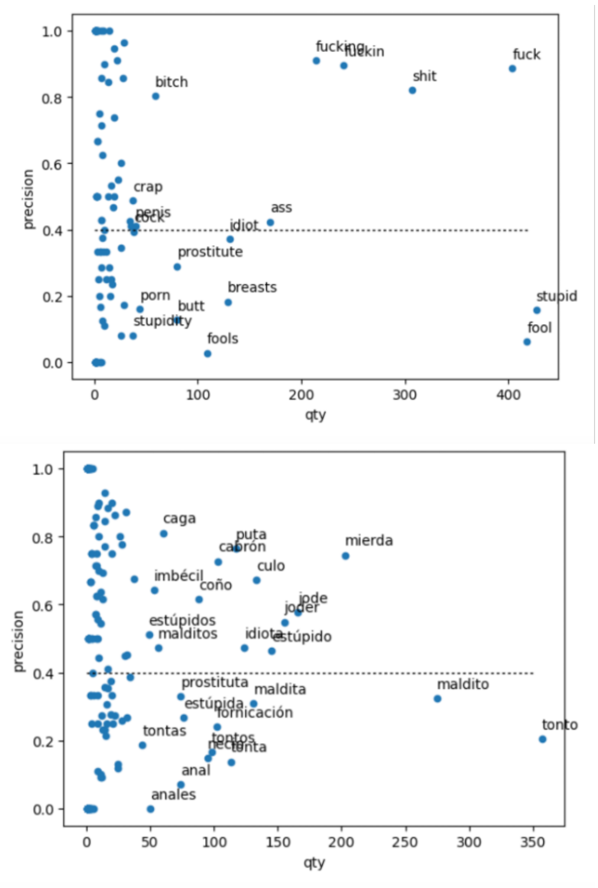

Figure 6: ETOX's tokens performance for our English Dataset (top) and Spanish Dataset (bottom). Vertical axis representing the precision and Horizontal axis the total output.

performing tokens are insults such as stupid* and fool*, these insults are not so harsh and probably some times are not considered non-toxic by native speakers. These show a high output, but very low precision and hence are negatively affecting ETOX overall's performance. On the other hand, some of the best performing tokens are variations of fuck* and shit*. These tokens show a high output and precision (>0.8), being responsable for a considerable share of recall (>0.15). Slurs tend to have a high precision but low recall.

For Spanish we have a total of 187 different toxic tokens detected and 110 tokens show reasonable precision (>0.4). The worst performing tokens are insults such as tonto* and maldito* and some of the best performing tokens are are m*erda and variations of j*der. The recall is much more distributed compared to English, with several terms having a precision above 0.4.

This analysis could be further extended to HP languages.