Project Report : CS 7643
GitHub Repo: https://github.com/unsupervised-pumpkins/tone-toxicity

Tyler Dickson
tdickson30@gatech.edu

Byungwoo Sohn
bsohn8@gatech.edu

Rakin Hasan
rhasan34@gatech.edu

Eunah Kim
ekim669@gatech.edu

Georgia Institute of Technology
225 North Ave, Atlanta, Georgia 30332

Abstract

As online media continues to grow, platforms increasingly rely on automated tools to detect toxic and derogatory speech. Current tools for detecting toxicity rely primarily on textual representations to do this, but overlook the role of vocal delivery. In this project, we explore whether the addition of audio improves toxicity detection. We compile a multimodal dataset of clips from three well known social/political commentators and assess the performance of three models: a text-only classifier, an audio-only classifier, and a multimodal classifier that combines both representations. Our findings show that audio alone poorly predicts toxicity, but does provide information regarding the structure of speech such as tone and pace. However, once the audio and textual representations are combined, the structure of speech becomes a useful supplement to the text. The multimodal model generalizes better to unseen examples than either unimodal baselines, suggesting that models that combine both audio characteristics offer a more complete understanding of toxic speech.

1. Introduction/Background/Motivation

In this project, we set out to understand whether the way someone speaks. Specifically, their tone, energy, and other audio cues changes how toxic their words seem compared to reading the same words as plain text. To study this, we built three prediction systems. The first system looks only at the written transcript of a short spoken segment. The second looks only at the audio. The third looks at both together.

All three systems examine short clips from three well-known commentary figures: Ben Shapiro, Joe Rogan, and Jon Stewart. For each clip, the goal is to produce a single score between 0 and 1 that reflects how toxic or harsh the clip sounds.

Our objective was not only to build these systems, but to compare them. We wanted to see whether adding audio information meaningfully changes the toxicity score compared to judging the text alone.

1.1. Limitations of Text-Only Toxicity Models: How it's Done Today

Today, toxicity detection is dominated by text-only supervised learning models. Systems like Google's Perspective API [5] introduced the idea of assigning a continuous toxicity score to online comments to assist with content moderation. Although widely adopted, Perspective has been criticized for decision boundaries with weak justification, inconsistent handling of linguistic variation, and an inability to adapt to different communities or conversational norms. This led to open-source alternatives, and numerous Kaggle competitions aimed at addressing this issue. The results have produced open-source projects like Detoxify [1], which uses modern transformer encoders to improve text-based toxicity classification.

Despite these advancements, current practice shares a fundamental limitation: nearly all widely used toxicity models rely solely on textual input. They ignore cues such as pitch, emphasis, pacing, energy, or sarcasm features that often determine whether speech feels hostile or benign. For example, two statements that appear identical in writing may diverge drastically in perceived toxicity once spoken.

This gap is increasingly important as public discourse shifts toward podcasts, video commentary, and short-form spoken media. In these settings, text transcripts capture what was said but fail to capture how it was said. Existing models therefore operate on an incomplete representation of the underlying signal and may systematically mises-

timate toxicity when audio characteristics carry meaningful information.

1.2. Motivation and Real World Application

Toxic and hostile speech affects social platforms, news outlets, and moderation teams that increasingly rely on automated tools to assess harmful content. Yet most systems analyze only text. This may cause many of them to miss vocal cues such as tone, emphasis, sarcasm, or intensity that can change how toxic a message sounds once spoken aloud.

A model that incorporates both text and audio would give these groups a more accurate view of harmful speech in today's media landscape, where podcasts, video commentary, and short clips dominate. It can help identify content that appears neutral in transcript form but communicates hostility through delivery, which could improve detection of extremist rhetoric, targeted harassment, and manipulative commentary.

If successful, our multimodal approach brings toxicity detection closer to real human communication, which in turn can enable safer platforms.

1.3. Dataset Collection and Description

Our dataset consists of paired text and audio segments extracted from publicly available YouTube videos featuring three high profile commentary figures: Ben Shapiro, Joe Rogan, and Jon Stewart. These speakers were selected because they produce both polarizing and non-polarizing content and collectively represent a broad range of political viewpoints.

Composition: Using automated transcript extraction and audio slicing tools, we segmented each video into short contiguous clips (10–45 seconds). Each segment contains (1) the transcript text, (2) the corresponding audio waveform, and (3) metadata such as speaker identity, video ID, and timestamps. This produced 3,863 text segments and 3,863 aligned audio segments (7726 total instances), distributed across the three speakers (1,104 from Shapiro, 1,995 from Rogan, and 764 from Stewart). The list of source videos appears in the supplementary materials. The dataset contains a random sample of instances. No tests were run to determine representativeness.

The dataset is a sample, not an exhaustive collection of videos, and is not intended to be representative of YouTube as a whole. All content was collected from public sources, and no personally identifiable information beyond speaker identity is included.

Labeling process: Each segment received a weak toxicity label generated using the *meta-llama/Llama-3.1-8B-Instruct*[4] model, producing a continuous toxicity score in [0,1]. To produce higher quality evaluation labels, team members manually reviewed and corrected these weak labels to ensure consistent interpretation of toxicity across speakers and segments.

Each instance is self-contained, consisting only of text, audio, and metadata needed for supervised learning.

2. Approach

To solve this problem, we frame toxicity prediction as a supervised classification problem over paired text and audio. Our objective is to compare three systems: text only, audio only, and a combined multimodal model, and evaluate whether combining these modalities improves toxicity classification in comparison to our unimodal baselines.

After generating continuous toxicity scores with *meta-llama/Llama-3.1-8B-Instruct*[4], continuous toxicity scores are discretize into several equally sized toxicity bins using fixed threshold ranges. Each text and audio segment is mapped into one toxicity bins making up ordered categories. The model then predicts a probability distribution over labels (toxicity classes) y given the input x is modeled by the function f with parameters θ, taking both text and audio features: $p(y|x) = f_\theta(x_{\text{text}}, x_{\text{audio}})$.

Data Pipeline: Both text and audio models were processed using aligned pipelines. We tokenized text inputs using the pre-trained *RoBERTa-base* model from Hugging Face [6] padded to a maximum sequence length, and converted to integer token IDs. Audio was preprocessed using a pretrained *facebook/wav2vec2-base model*[2]. The model was resampled to 16 kHz because Wav2Vec2 is specifically trained on 16,000 samples per second, and the model expects 16kHz output. Audio segments are cropped or padded to a fixed duration corresponding to the chosen segment window. Training, testing, and validation splits were made at the video level to avoid audio leakage between segments originating from the same video.

2.1. Model Creation

We trained three models of increasing complexity: a text classifier, an audio classifier, and a multimodal classifier combining both representations. All models optimize the cross-entropy loss over discretized toxicity bins. Group members explored different hyperparameter settings during tuning based on validation accuracy or MAE over toxicity bins.

Text Classifier: The text classifier fine tunes a pretrained RoBERTa-base transformer encoder model imported from Hugging Face. Input tokens are passed through the transformer, and the pooled sequence representation (CLS tokens) are extracted and fed into a feedforward classification head that outputs class logits (toxicity score).

Audio Classifier: The audio classifier fine tunes a pretrained facebook/wav2vec2-base model. Audio waveforms are transformed into latent speech representations, which are aggregated using mean pooling to obtain a fixed-length

embedding. A linear classification head maps this embedding to toxicity logits. Because the dataset contains clips of varying intensity and speaking style, the audio model captures diction, tone, and emphasis that are not present in the transcript.

Multimodal Classifier: The multimodal classifier combines both text and audio embeddings into a fused architecture. After both text and audio models produce their respective hidden vectors, they are concatenated, and passed into a network consisting of fully connected layers. The fused representation captures semantic cues, which is passed into the final classification layer.

This fusion strategy is motivated by the fact that toxicity often depends on *how* something is said in addition to *what* is said. Sarcasm, tone, or emotional delivery may not be identifiable from transcripts alone, and audio-only systems lack the detailed information contained in text.

2.2. Challenges and Limitations

Weak Labels: The dataset relies on weak labels generated from an LLM and later corrected manually. To frame this as a classification problem, continuous values were discretized into five bins:

$$\text{Bin}(s) = \begin{cases} 0 & \text{if } s < 0.2 \\ 1 & \text{if } 0.2 \leq s < 0.4 \\ 2 & \text{if } 0.4 \leq s < 0.6 \\ 3 & \text{if } 0.6 \leq s < 0.8 \\ 4 & \text{if } s \geq 0.8 \end{cases}$$

Although manual revisions improved quality, some uncertainty remains with our designated bin boundaries. Some segments contain questionable levels of toxicity that is difficult to identify and discretize accurately. During the tuning and evaluation process, this resulted in noisy gradients, and early overfitting, especially with the audio model where cues such as tone, emphasis, and pace are subtle. More importantly, because the initial toxicity scores were produced by an LLM reading the transcript, the final labels likely remain more sensitive to what was said than how it was said. This creates a structural disadvantage for the audio-only model, which is trained to predict targets that may not strongly encode acoustic delivery.

Metric Choices: Because our training objective was five-class cross-entropy, we prioritized validation loss and accuracy as the most direct indicators of whether the models were learning the intended task. We also reported MAE by converting predicted class probabilities into continuous scores using fixed bin midpoints, which provided an interpretable check while remaining consistent with the classification setup. We considered additional cross-model analyses such as mean absolute score shift (MASS) and correlation across modalities during the initial proposal. However, these measures assume stable, well-calibrated continuous predictions and benefit from stronger continuous ground truth. Given that our labels originated from a text-driven weak-label pipeline with limited manual correction, and that our validation split was relatively small, we judged MASS and correlation to be less reliable for this iteration and more likely to reflect label construction artifacts rather than true modality-driven shifts. We therefore treated these analyses as future work better suited to improved labeling and repeated-run evaluation.

Model Size Constraints: The decision to use RoBERTa-base for text and Wav2Vec2-base for audio were motivated by their historically strong performance with language and speech tasks. However, both models are computationally heavy. Even fine-tuning the models by a single epoch required substantial GPU memory. This proved to limit batch size and ultimately constrain our hyperparameter search by a meaningful amount. In practice, these models can demand close to 32GB of VRAM under certain configurations, and having two 16GB GPUs did not meaningfully help because a single training process cannot easily combine memory across two devices. This prevented us from increasing batch sizes to the levels we initially planned, which likely limited training stability and reduced the potential gains from broader tuning.

Training Instability: Although segments were aligned by time, the amount of information within a given snippet between audio and text differed significantly. For example, some audio clips contained long pauses, while the corresponding text transcript only contained a few tokens. It was more challenging than anticipated to ensure that both text and audio models contributed meaningful information. It also reinforces a broader limitation of audio-only modeling in this setting: a speaker can deliver a clearly negative comment in a calm, flat, or monotone voice, in which case the acoustic signal may not reliably convey toxicity even when the textual meaning is overtly harsh. This suggests that audio cues are more likely to improve performance as a complementary signal rather than as a standalone basis for toxicity prediction in our dataset.

Data Collection: We adapted YouTube's scraping and audio downloading functionality from Youtube's transcript API [3]. We adapted the api and audio download utilities to generate JSONL metadata.

The following pre-trained models were used:

- *meta-llama/Llama-3.1-8B-Instruct* (Noisy Labels) [4]
- *RoBERTa-base* (Text Processing) [6]
- *facebook/wav2vec2-base* (Audio Processing) [2]

3. Experiments and Results

Table 1 and Table 2 below summarizes the final results of our experiments.

Hyperparameter	Value	Scope
# classes	5	All
Batch size	32	All
Epochs	7	All
Learning rate	2×10^{-5}	All
Text backbone	`roberta-base`	T, TA
Max text length	256	T, TA
Audio backbone	`facebook/wav2vec2-base`	A, TA
Audio sample rate	16,000	A, TA
Max audio duration (s)	15.0	A, TA

Table 1. Hyperparameters used across all three models.

Metric	Model T (Text)	Model A (Audio)	Model TA (Text+Audio)
Training Loss	0.5489	1.3734	0.5489
Training Accuracy	0.7944	0.3686	0.7786
Validation Loss	1.1652	1.3965	0.9408
Validation Accuracy	0.5688	0.3523	0.6632
Validation MAE	0.1306	0.1984	0.1108

Table 2. Metric comparison across text-only (T), audio-only (A), and multimodal (TA) models at epoch=4.

3.1. Model T (Text Only)

We first evaluate Model T, our text-only baseline, to understand how far we can go using transcripts alone. Because the original label is a continuous toxicity score in [0,1], we trained the model as a 5 class classification problem using fixed bins. To measure success, we report both validation accuracy and validation MAE. Accuracy helps us see whether the model predicts the correct bin, while MAE gives a more direct sense of how close the predicted toxicity level is to the original continuous score.

Quantitatively, Model T shows strong baseline performance. Using the epoch 4 setting reported in Table 2, the model achieves train loss = 0.5489, train accuracy = 0.7944, validation loss = 1.1652, validation accuracy = 0.5688, and validation MAE = 0.1306. These results suggest that the text signal alone already carries meaningful information for toxicity prediction. In many segments, the semantic content and explicit wording likely provide clear cues about harshness or negativity.

The learning curves also support this interpretation. As shown in the training and validation loss plot for Model T (Figure 1), training loss decreases steadily across epochs, indicating stable learning and good fit to the training distribution.

Meanwhile, validation loss improves early but does not keep decreasing, and even shows mild fluctuation afterward. This pattern suggests that the model may start to overfit after a small number of epochs, which is reasonable given the limited dataset size and the use of weak labels. The validation MAE curve (Figure 2) shows a similar trend, where the error appears lowest around the mid epochs rather than at the very end. This implies that an early stopping point around epoch 3,4 is a sensible choice for Model T.

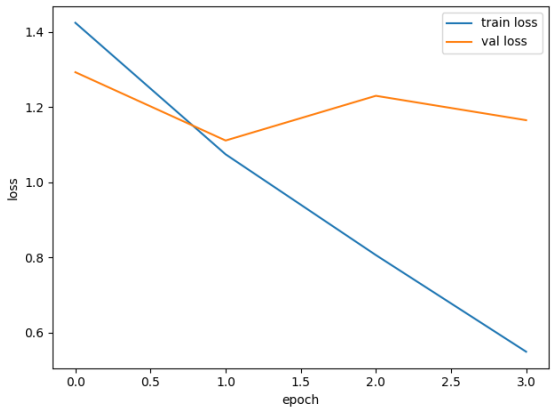

Figure 1. Training and validation loss across epochs for Model T (text-only).

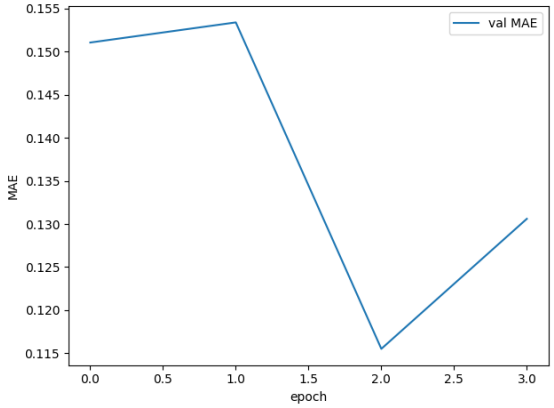

Figure 2. Validation MAE across epochs for Model T (text-only).

Overall, Model T provides a reliable and competitive baseline. At the same time, the gap between improving training loss and less consistent validation improvement hints at the limitations of using text alone. Spoken toxicity may depend not only on what was said but also how it was delivered. This motivates our audio-only and multimodal experiments, where tone, emphasis, and energy may provide additional informative signals beyond the transcript.

3.2. Model A (Audio Only)

We evaluated the audio-only model (Model A) using validation accuracy, cross-entropy loss, and MAE, where MAE was computed by converting predicted class probabilities into a continuous score using fixed bin midpoints. We report these metrics to directly compare Model A against the text-only and multimodal models under the same evaluation lens. Our main focus here is whether audio cues alone carry enough useful signal to track the toxicity labels in our dataset.

Across four epochs, Model A shows limited but mea-

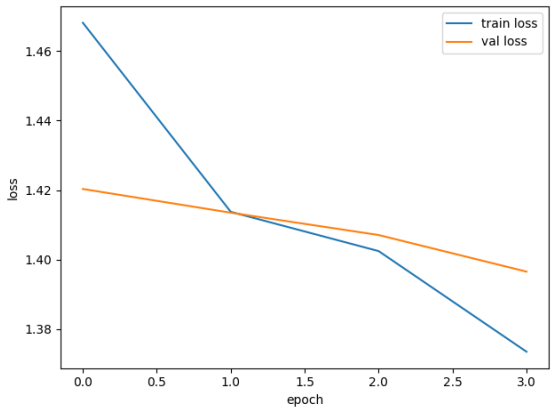

Figure 3. Training and validation loss across epochs for Model A (audio-only).

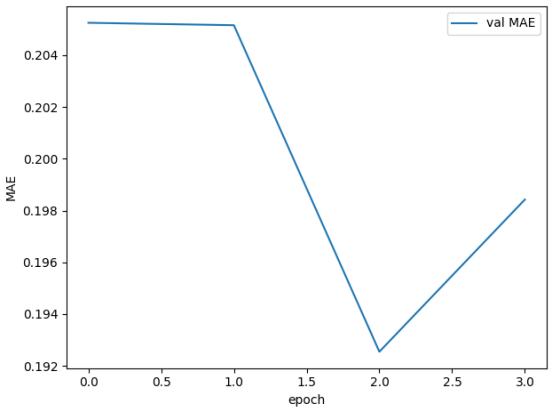

Figure 4. Validation MAE across epochs for Model A (audio-only).

surable learning. The training and validation loss curves in Figure 3 both trend downward slightly across epochs, which suggests the model is not stuck and is picking up some consistent patterns from the audio. This aligns with the numeric results at epoch 4, where the model reaches a training accuracy of 0.3686 and a validation accuracy of 0.3523, with a validation loss of 1.3965. However, the overall gains remain modest compared to text-based models, and the final validation accuracy is still close to a weak baseline for a five-class task. However, the final performance remains near a weak baseline for a five-class task and clearly trails the text-only model (validation accuracy 0.5688, MAE 0.1306). In other words, audio-only prediction does not appear to recover enough information to reliably approximate our toxicity targets.

The MAE trend in Figure 4 is also consistent with this interpretation. The curve improves most noticeably around epoch 3 and then slightly rebounds by epoch 4, indicating mild instability and limited generalization. The final validation MAE of 0.1984 remains substantially worse than the text-only result (0.1306). A likely explanation is that the labels themselves are more aligned with textual meaning than vocal delivery. Since our initial scores were produced from transcripts and only lightly corrected by humans, Model A may be trying to predict a target that is not strongly determined by speech characteristics. This also fits a simple real-world case: someone can say something strongly negative in a calm or controlled tone. In that situation, the audio may not sound especially hostile even when the content is clearly toxic, so an audio-only model is inherently limited if the ground truth is primarily content-driven.

Overall, Model A was partially successful as an analysis tool but not as a standalone predictor. It provides an important baseline showing that, under our current labeling scheme and dataset, audio alone is not enough to match text-based performance. This supports the interpretation that audio is most valuable as a complementary signal in multimodal settings rather than a replacement for text.

3.3. Model TA (Text + Audio)

Since the multimodal classifier contains the text embeddings of the Roberta-Base transformer encoder (in addition to the audio embeddings), it should be able to perform close to if not slightly better than Model T. Given the much less accurate results of model A compared to model T, model TA should not have too much of a difference compared to model T.

The respective performances of Model T and Model TA (Shown in Table 2) in the training data support this conclusion, as they have nearly the same loss and accuracy. However, Model TA shows a noticeable improvement when it comes to performing on the validation set. When going from model T to TA, the loss on the validation set goes down from 1.1652 to 0.9408, and the accuracy goes up from 0.5688 to 0.6632. At the fourth epoch, the MAE goes from 0.1306 to 0.1108. The trend lines of the validation loss and validation MAE (Shown in Figure 5 and Figure 6 respectively) further reflect this. There is a consistent downward trajectory across the epochs, and any rebounds (such as epoch 3 in the validation loss and epoch 2 in the MAE) are relatively small. This shows that the model gets consistently better at generalization over time, under a reasonable limit. When we tested different configurations of all the models, they would start to fluctuate around epoch 4 or higher. These results clearly show that Model TA is better at generalization than Model T and that the added audio embeddings are adding some useful information. If they were not, the model would have more useless parameters and that would make it more susceptible to overfitting to the training data (thus being worse at generalization).

The results of Model TA compared to Model A show that Model A is likely only considering the structure of the audio

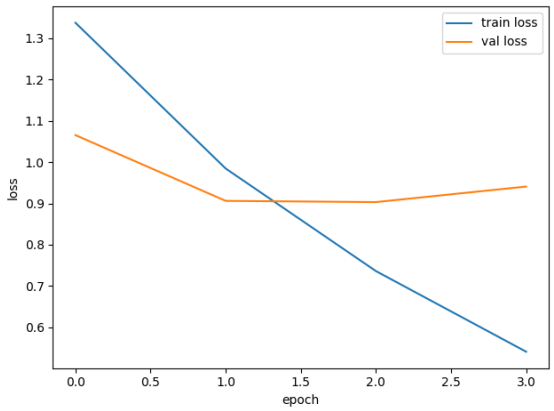

Figure 5. Training and validation loss across epochs for Model TA (text + audio).

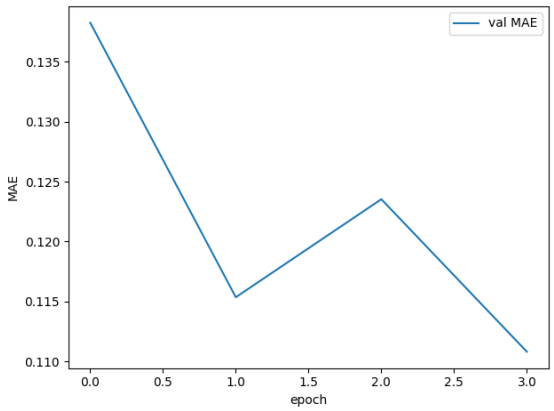

Figure 6. Validation MAE across epochs for Model TA (text + audio).

(which contains aspects such as pitch,speed, and coherence of the speech) without considering the semantic meaning of the words being said. The audio embedding vector space is configured in such a way that two separate clips of people speaking in a similar manner will be very close to each other, even if the words and subject matter are very different. This is useful information, but given all the various nuances in speech, such as potential sarcasm and the natural differences in people's unique voices, this alone is not a good predictor of toxicity. Once the semantic meaning of the speech is added through the text embeddings, then the audio becomes a useful addition that gives the model more context to make its prediction. This effect could potentially have been even greater if the ground truth labels were less biased towards the textual meaning.

4. Conclusion

We observed cases where the text-only model predicted high toxicity even when the audio delivery sounded neutral, as well as cases where the multimodal model lowered the predicted toxicity compared to the text-only baseline. This supports the intuition that tone can soften or intensify how a statement is perceived. At the same time, because our labels were initially generated from transcripts and only lightly corrected, the training signal likely remained more sensitive to what was said than how it was said. This helps explain why the added value of audio appeared meaningful but not large. In this sense, our results suggest that audio is most useful as a complementary cue under our current dataset and labeling scheme rather than a standalone driver of toxicity prediction.

A practical implication of this finding is the tradeoff between performance and cost. The text-only model was substantially more efficient to train, taking roughly 3 minutes per epoch, while the audio-only model required about 15 minutes per epoch and the multimodal model about 20 minutes per epoch. Given that the observed improvement from adding audio was modest relative to the additional compute time, the multimodal approach may be most appropriate for settings where capturing delivery-based nuance is a priority and resources permit longer training cycles. Conversely, for many academic or production contexts that prioritize speed and simplicity, a strong text-only baseline remains a competitive and more efficient choice.

5. Future Studies

Future work for this project is fairly clear. The biggest extension would be to move beyond text and audio and add visual signals from video, such as facial expressions, head movement, posture, and hand or shoulder gestures. These cues often shape how people interpret intent, sarcasm, intensity, or hostility. A richer multimodal model that combines transcript meaning with vocal delivery and visible affect may better match real human judgments of toxicity in spoken commentary, especially in cases where the words alone are ambiguous.

A second improvement is to strengthen the labeling strategy to better capture delivery. Because our weak labels were generated from transcripts, the targets likely reflect what was said more than how it was said. Future work could use a smaller but higher quality set of multimodal human ratings to better measure the added value of audio and video. Finally, we should treat compute as a practical constraint by adopting multi GPU and memory efficient training to enable broader tuning without relying on a single high memory device.

Student Name	Contributed Aspects	Details
Tyler Dickson	Data Creation, Implementation and Analysis	Scraped the dataset for this project and created training model templates to be tuned by teammates. Wrote Introduction / Background / Motivation / Approach / Challenges and Limitations sections. Wrote Project Proposal.
Byungwoo Sohn	Implementation and Analysis	Github Repo setup, set fixed random seeds for reproducibility, fixed bugs, implemented MAE calculation, tuned Model A and Model TA. Generated final charts and values for the report. Wrote Challenges and Limitations, Model A analysis and Future Studies sections.
Eunah Kim	Implementation and Analysis	Implemented visualizations: loss curves and validation MAE curves for all three models. Tuned Model T. Wrote Model T analysis and Conclusion sections.
Rakin Hasan	Analysis	Tuned Model TA. Wrote Abstract / Model TA analysis sections.

Table 3. Contributions of team members.

References

[1] Unitary AI. Detoxify: Toxic comment classification models. `https://github.com/unitaryai/detoxify`, 2025. Python library for toxicity classification.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*.

[3] Jonas Depoix. Youtube transcript api. `https://pypi.org/project/youtube-transcript-api/`, 2025. Python library for retrieving YouTube video transcripts.

[4] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[5] Jigsaw and Google. Perspective api. `https://developers.perspectiveapi.com/`, 2025. Online tool for detecting toxic and harmful language.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.