

Speech-to-Text and Tonal Analysis for Voice-based Sentiment Analysis using Small Language Modeling

Submitted in the Fulfillment of the Requirements in
CPE313 Data Science Track Elective 3:
Advanced Machine Learning and Deep Learning

Submitted by
Dejoras, Dylan James N.
Villamor, Kurt Russel A.

May 19 2025

Speech-to-Text and Tonal Analysis for Voice-based Sentiment Analysis using Small Language Modeling

Dejoras, Dylan James N.

Computer Engineering
Technological Institute of the Philippines
Quezon City
qdjndejoras@tip.edu.ph

Villamor, Kurt Russel A.

Computer Engineering
Technological Institute of the Philippines
Quezon City
qkravillamor@tip.edu.ph

ABSTRACT - *Natural language processing (NLP) and speech-to-text (STT) systems often face challenges in accurately detecting sentiment due to biases, noise, and limitations in handling spontaneous speech and low-resource languages. To address these issues, we compare the effectiveness of sentiment classification using transcribed text and raw audio. For text, we leverage pre-trained models such as RoBERTa, TinyBERT, and ELECTRA-Small, balancing performance with computational efficiency. On the audio side, we evaluate Wav2Vec 2.0, DaBloatCNN, and ResNet50. Results show that Wav2Vec 2.0 achieved the highest accuracy and F1 score (97.8%), outperforming all other models, while RoBERTa led among text-based approaches. These findings highlight the strength of raw audio models in capturing emotional tone and suggest that combining both modalities could improve the robustness and fairness of sentiment analysis systems.*

Keywords: *Sentiment Analysis, Wav2Vec 2.0, RoBERTa, TinyBERT, ELECTRA, speech emotion recognition, audio classification, natural language processing, multimodal learning, deep learning.*

I. Introduction

In several industries, sentiment analysis is used to understand satisfaction levels or customer feedback. This utilizes either text data or audio files for systems to make data-driven responsive decisions that improve user experience and efficiency.

Artificial Intelligence (AI) has become a transformative force across numerous domains in the modern world. One of its notable fields is 'Natural Language Processing (NLP)', a domain where computers are programmed to comprehend, and imitate human language. It is one of the fields where machine learning and deep learning techniques are applied to further

enhance the utilized language models to its fullest. NLP has multiple purposes by facilitating applications such as chatbots, speech recognition systems, search engines, etc. This paper focuses on sentiment analysis, a branch of NLP where a sentiment in a text is automatically determined. The challenge lies in performing accurate and efficient sentiment analysis using small language models that can run in resource-constrained environments without sacrificing performance. Sentiment analysis plays a critical role in understanding public opinion and evaluating customer feedback, both of which are key considerations in business decision-making. Using 'Large Language Models' (LLMs) are the norm for conducting sentiment analysis. However, as there are limitations for the machinery, an attempt in settling for 'Small Language Models' (SLMs) is made. This paper would tackle the application of variations of BERT and ELECTRA models such as 'RoBERTa', 'Tiny BERT', and 'ELECTRA-Small (Discriminator)' in an effort that would prioritize efficiency without sacrificing quality.

Text-based sentiment analysis has become popular in many applications like analyzing social media posts and customer reviews. However, relying only on text does not always give accurate results, especially in real-life situations where emotions are also expressed through voice. For example, a sentence can sound happy or angry depending on how it is spoken. This type of emotional information cannot be captured through text alone. Because of this, voice-based sentiment analysis has become an important area of research. It uses features like tone, pitch, and speed of speech to better understand how someone feels [16]. With recent progress in deep learning, especially with models that can learn directly from raw audio, speech emotion recognition systems are becoming more accurate and reliable [17], [18]. As a result, using voice is now seen as a strong alternative to text in understanding human emotions in many systems, such as virtual assistants or call centers [19].

In this work, we aim to compare the effectiveness of sentiment classification using text alone versus audio alone. We want to better understand which type of input carries more emotional information and which one performs better in practical classification tasks. To do this, we explore several deep learning models for audio-based sentiment recognition. We built a Convolutional Neural Network (CNN) from scratch using raw log-Mel spectrograms and fine-tuned a ResNet-50 model using log-Mel spectrogram images. We also fine-tuned Wav2Vec 2.0, a powerful pre-trained model that learns directly from raw waveforms. By comparing the results of these audio-based models with a text-only sentiment classifier, we aim to highlight the strengths and limitations of each approach.

II. Related Work

A. Introduction of BERT and ELECTRA models

BERT is a multi-layer transformer-based language model that is pre-trained on large-scale datasets using tasks such as masked word prediction and next sentence prediction [2]. It is known for its strong empirical performance and straightforward architecture. The BERT framework involves two main stages: pre-training and fine-tuning. During pre-training, the model learns general language representations from unlabeled data through various tasks. Once pre-training is complete, BERT can be adapted to specific downstream tasks in one of two ways: the feature-based approach or full fine-tuning. In the feature-based method, fixed representations are extracted from BERT and used as input features for task-specific models, while in the fine-tuning approach, a task-specific layer is added on top of BERT, and all model parameters are jointly trained using labeled data from the target task [3]. This allows BERT to effectively transfer its learned knowledge to a wide range of natural language processing applications.

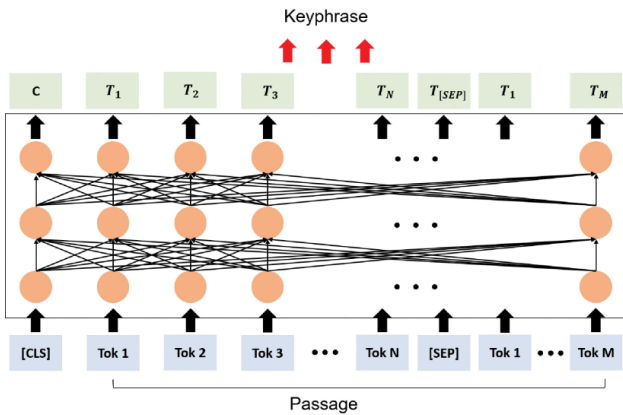


Fig 2.1. Transformer Architectures [15]

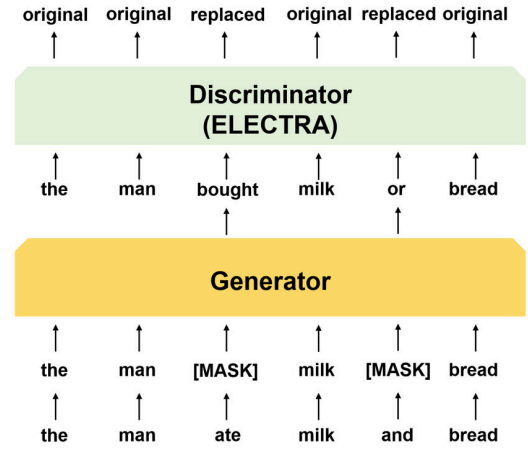


Fig 2.2. ELECTRA Architecture [15]

ELECTRA is a pre-training method designed to identify altered tokens within a sentence by training a discriminator model to distinguish between original and replaced tokens [4]. Unlike traditional masked language modeling approaches, ELECTRA employs a generator network to produce high-quality negative samples, which are then used to train the main encoder to effectively detect token substitutions. This approach enables more computationally efficient pre-training while achieving superior performance with greater parameter efficiency compared to other language modeling techniques [4]. Electra primarily utilizes two neural networks: the Generator and the Discriminator.

B. Analysis of Model Variants

For this study, various adaptations of BERT and ELECTRA models, including RoBERTa, TinyBERT, and ELECTRA-Small (Discriminator), are utilized to optimize performance while maintaining computational efficiency without compromising model quality. The approach ensures a balance between resource utilization and predictive accuracy, making it a viable solution for real-world applications.

RoBERTa shares the same underlying architecture as BERT [5] but employs a byte-level BPE tokenizer, similar to GPT-2, while adopting an alternative pre-training approach. TinyBERT's learning process consists of both general distillation and task-specific distillation. In the general distillation phase, the maximum sequence length is set to 128 [6]. The used model for this study is a compact BERT model with just 2 layers ($L=2$) and 128-dimensional embeddings ($H=128$), making it much smaller and faster than standard BERT. It's ideal for low-resource settings or quick experiments while still offering useful language understanding capabilities. The compact ELECTRA-Small model follows a downscaled version of the BERT-Base architecture, featuring reduced sequence length (from 512 to 128), smaller batch size

(from 256 to 128), and decreased hidden dimension size (from 768 to 256), along with smaller token embeddings (from 768 to 128) [7]. These modifications allow ELECTRA-Small to achieve strong performance on NLP tasks while significantly lowering computational requirements, making it highly suitable for fast training on limited hardware.

C. Prior Applications of Transformer-Based Language Models in Related Domains

The key challenge in maximizing the accuracy of language models lies in maintaining a balance between performance and computational efficiency. Although large pre-trained models like BERT-base and BERT-large demonstrate strong accuracy across various NLP tasks, their implementation incurs significant costs. These include prolonged training and inference times, high memory usage, substantial energy consumption, and demanding hardware requirements such as multiple GPUs or TPUs. Consequently, deploying such models in practical applications, particularly in low-resource environments or edge-device settings becomes increasingly difficult.

In the real world, the mentioned transformer models are widely used in several fields. As for RoBERTa, a COVID-19 fake news Detection [8] and online review sentiment analysis [9] were conducted using the mentioned model. An AI-Driven smishing detection application for Android devices [10] and rumor detection [11] were created with the help of the TinyBERT transformer. Lastly, ELECTRA-Small could be used for classification of depression from texts and also sentiment analysis [12].

D. The Challenges of Having Limited Computational Resources and Machinery

Despite significant advancements in large language models, several challenges remain. While smaller models are more cost-effective and energy-efficient, larger models deliver better performance and broader functionality but come with higher environmental costs and resource consumption [14]. Integrating large models into real-world or edge-based systems, such as mobile apps, embedded devices, or IoT environments, is often impractical due to substantial computational requirements, including memory usage, inference speed, and power demands. Enhancing the efficiency of large language models to improve sustainability and deployment feasibility remains a key focus in AI research.

E. Model Variants and Representations

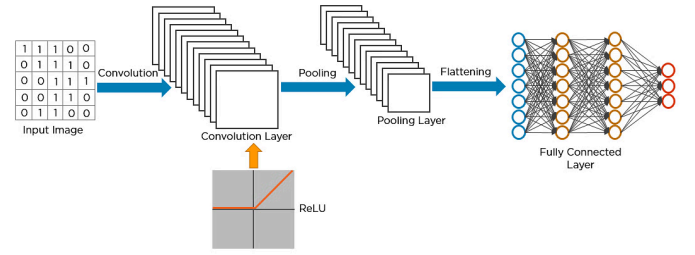


Fig 2.3. CNN Model Architecture

Several architectures have been explored in recent studies to enhance speech emotion recognition (SER), especially when using spectrogram representations of audio data. CNNs trained on log-Mel spectrograms have demonstrated effectiveness in capturing time-frequency emotion features. In particular, treating spectrograms as images enables CNNs to extract spatial and temporal cues relevant to emotional expression [20].

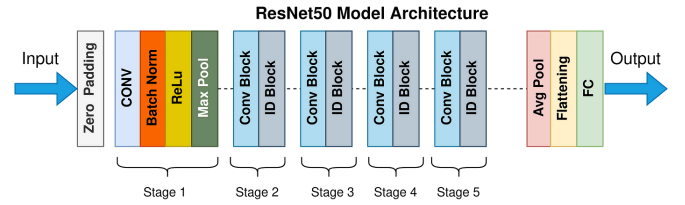


Fig 2.4. ResNet50 Model Architecture

Extending this idea, ResNet-50—originally designed for large-scale image classification—has been applied to spectrogram inputs in SER. Its residual connections support deeper networks without vanishing gradients, aiding the detection of subtle emotion cues in speech. Additionally, techniques such as saturation-based data augmentation have proven effective in improving performance [20].

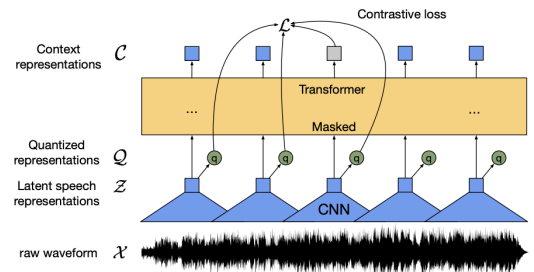


Fig 2.5. Wav2Vec 2.0 Model Architecture

Wav2Vec 2.0, a self-supervised model developed by Facebook AI, offers an alternative approach by learning contextual speech features directly from raw audio. It has demonstrated strong results when fine-tuned on emotion recognition tasks, especially in low-resource scenarios [21]. Recent work

has also explored combining Wav2Vec 2.0 and transformer models like BERT for multimodal emotion recognition, showing that integrating acoustic and textual features improves overall performance [22].

F. Computational Constraints and Lightweight Solutions

While deep models like Wav2Vec 2.0 and ResNet-50 are powerful, their high memory and computation demands pose challenges for deployment on edge devices or low-resource environments. This has led to interest in lightweight alternatives like TinyBERT and simplified CNN variants. These models maintain reasonable performance with fewer parameters and lower inference cost [23].

To further enhance small model performance, multi-feature fusion and attention mechanisms have been proposed. These help the models focus on critical regions within spectrograms, improving emotion recognition even in noisy environments. Such strategies are especially useful when full-scale feature extraction is computationally expensive [24].

III. Methods

III.A. Text-Based Sentiment Analysis

III.A.1. Motivation for Using Small Language Models

The key challenge in maximizing the accuracy of language models lies in maintaining a balance between performance and computational efficiency. Although large pre-trained models like BERT-base and BERT-large demonstrate strong accuracy across various NLP tasks, their implementation incurs significant costs. These include prolonged training and inference times, high memory usage, substantial energy consumption, and demanding hardware requirements such as multiple GPUs or TPUs. Consequently, deploying such models in practical applications, particularly in low-resource environments or edge-device settings becomes increasingly difficult

III.A.2. Data Preprocessing (Sentiment Analysis)

After concatenating four sentiment analysis datasets with the purpose of having one generalized dataset. The sentiment analysis dataset used in this study consists of a total of 118,334 labeled text samples, partitioned into 95,039 training instances and 23,295 test instances. This corresponds to an approximate 80–20 train-test split, ensuring sufficient data for model learning while preserving a representative evaluation set. Each sample is annotated

with one of three sentiment labels: negative, neutral, or positive.

III.A.3. Model Adaptation and Fine-Tuning

The ELECTRA-Small Discriminator model was fine-tuned for multi-class sentiment classification using a custom dataset composed of 118,334 labeled text samples. The pre-trained model was loaded via Hugging Face Transformers and adapted for the target task by adding a classification head supporting three output classes: negative, neutral, and positive. Text sequences were tokenized using the ELECTRA tokenizer with a maximum sequence length of 128 tokens, applying uniform padding and truncation.

Training was conducted using PyTorch, with data loading handled through a custom SentimentDataset class that enabled efficient batching and GPU acceleration. A batch size of 16 was used, along with the AdamW optimizer at a learning rate of 2×10^{-5} . A linear learning rate scheduler was applied over 12 training epochs to improve convergence.

During training, the model was evaluated on both training and validation sets using metrics such as accuracy, macro and weighted F1-score, precision, recall, Cohen's Kappa, and Matthews Correlation Coefficient (MCC). These metrics provided a comprehensive view of model performance across imbalanced classes and ensured robustness in real-world deployment scenarios.

III.B. Voice-Based Sentiment Analysis

III.B.1 Dataset Preparation

For this project, we used a combination of four datasets: TESS, SAVEE, RAVDESS, and CREMA-D. These datasets consist of raw audio files that capture spoken expressions of emotion, performed by various actors of different ages. Each dataset includes 6 to 7 predefined emotions such as happy, sad, angry, fearful, disgusted, surprised, and neutral.

To enable a direct comparison between voice-based and text-based sentiment analysis models, we grouped these emotions into three broader sentiment categories. Negative sentiment includes audio samples labeled as sad, angry, fearful, and disgusted. Positive sentiment comprises happy and surprised emotions, while neutral remains as its own category.

Each raw audio file undergoes a series of preprocessing steps before being fed into the model. Initially, the raw audio is converted into a Mel spectrogram. This spectrogram is then transformed

using a logarithmic scale to produce a log-Mel spectrogram. To enhance model robustness and generalization, data augmentation techniques are applied exclusively to the training set. These include pitch shifting by 70%, time stretching by 80%, adding random noise, and further spectrogram stretching.

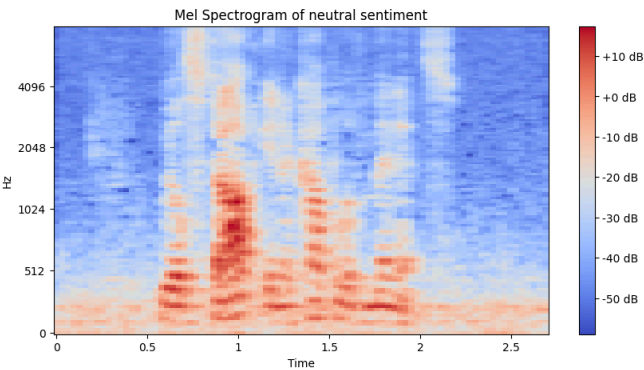


Fig 3.1 Sample Mel Spectrogram of an audio file with Neutral sentiment

In the second stage of preprocessing, each log-Mel spectrogram is converted into a 224×224 RGB image to make it compatible with vision-based neural network architectures. To further enhance the variability and robustness of the training data, a series of image augmentation techniques are applied. These include Time Masking and Frequency Masking to simulate missing information along the temporal and spectral dimensions, respectively. Additionally, visual augmentations such as color jittering, random cropping, and horizontal flipping are employed to introduce subtle distortions that prevent overfitting and improve the model’s generalization on unseen data.

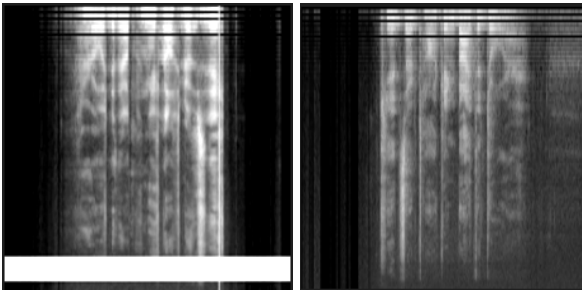


Fig 3.2 Sample 224x224 images of a Negative sentiment audio (train, test)

The final preprocessing method involves directly extracting the raw waveform array from each audio file, sampled at a consistent rate of 16,000 Hz. This raw audio representation preserves the original time-domain signal and is typically used as input for models like Wav2Vec 2.0 that operate directly on waveform data without the need for spectrogram transformation.

III.B.2 Model Preparation

For this study, a custom convolutional neural network named DaBloatCNN was developed using PyTorch and made Hugging Face-compatible by subclassing the PreTrainedModel and PretrainedConfig classes. This allowed the model to be trained and evaluated using Hugging Face's Trainer API, aligning it with modern NLP and speech models.

DaBloatCNN is a 1D CNN designed to classify speech sentiment using raw audio array inputs. The architecture features five convolutional blocks, each followed by batch normalization, max pooling, and dropout layers where applicable. These layers progressively extract increasingly abstract features from the 1D input. After convolutional processing, the output is flattened and passed through two fully connected layers to predict one of the three sentiment labels: positive, negative, or neutral.

The model dynamically computes the flattened size of the convolutional output to accommodate different input lengths. This allows flexibility for varying audio lengths while maintaining architectural integrity.

Tab 3.1 Sample Mel Spectrogram of an audio file with Neutral sentiment

Layer Block	Description	Output Shape (Example)
Conv1 + BN + Pool (x2)	Conv1d layers with BatchNorm and MaxPool reduce time dimension by 4x	[B, 512, T/4]
Dropout	Regularization (p=0.2)	[B, 512, T/4]
Conv1 + BN + Pool (x3)	Conv1d layers with BatchNorm and MaxPool further reduce time dimension by 8x	[B, 128, T/32]
Dropout	Regularization (p=0.2)	[B, 128, T/32]
Flatten + FC + BN	Flatten features and fully connected layer to 512 units	[B, 512]
FC (Output)	Final fully connected layer to 3 output logits	[B, 3]

ResNet50 is a popular deep learning model designed for image recognition. It uses a special technique called “residual learning” to help train very deep networks by allowing layers to skip connections, which prevents the problem of vanishing gradients. In this project, ResNet50 is used as a feature extractor to analyze spectrogram images of speech, helping the model learn important audio patterns for classification.

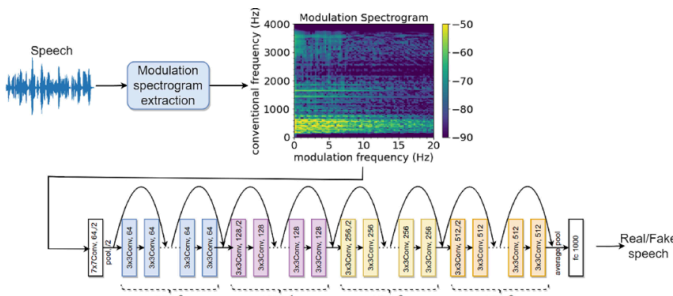


Fig 3.3 Sample Spectrogram in Fine Tuning ResNet50

Wav2Vec 2.0 is a state-of-the-art self-supervised model designed to process raw speech audio without relying on handcrafted features. It learns directly from the raw waveform by first pre-training on large amounts of unlabeled audio data. During this pre-training phase, the model masks parts of the audio signal and learns to predict these masked sections using the surrounding context. This helps the model develop rich and general speech representations without any labeled data.

After pre-training, Wav2Vec 2.0 is fine-tuned on smaller, labeled datasets for specific tasks such as speech recognition or emotion classification. The model architecture combines a convolutional feature encoder, which extracts local acoustic features, with a Transformer network that captures long-range dependencies in the speech signal. This design allows Wav2Vec 2.0 to effectively understand complex speech patterns, making it highly accurate for various speech-related applications.

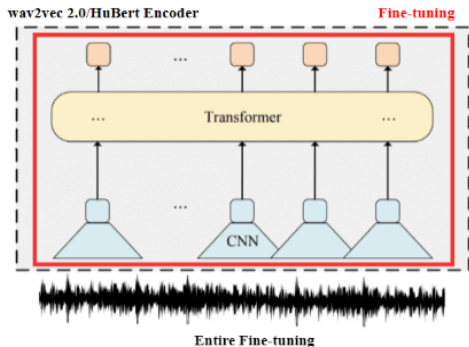


Fig 3.4 Sample Fine Tuning the Whole Wav2Vec2

B.3 Training or Finetuning

The training phase involved fine-tuning three different models—DaBloatCNN, ResNet50, and Wav2Vec 2.0—each tailored to a specific type of input: spectrograms (1D and 2D) and raw audio. All models were trained using Hugging Face’s Trainer API, which provided a consistent interface for managing training, evaluation, and checkpointing. Performance was monitored using accuracy and weighted F1-score, ensuring that the models not only predicted correctly but also handled class imbalances effectively.

For DaBloatCNN, a custom 1D CNN designed for log-Mel spectrograms, the model was trained for 25 epochs with a batch size of 64 and a learning rate of 5e-4. Mixed-precision training (fp16) was enabled to accelerate performance and reduce memory usage. Evaluation, saving, and logging were all performed per epoch, and the best model was selected based on validation accuracy. The training process used dropout, batch normalization, and cross-entropy loss to improve generalization and stability.

The ResNet50 model, adapted for 2D spectrogram image classification, was trained over 10 epochs with a batch size of 32 and a learning rate of 1e-5. Evaluation and saving were done at the end of each epoch, and the model checkpoint with the best F1-score was retained. A custom metric function computed accuracy and weighted F1, and a processor was used to ensure input consistency. This setup allowed for stable fine-tuning of ResNet’s deeper layers for emotion classification.

For Wav2Vec 2.0, which directly processes raw waveforms, training was performed over 5 epochs with a small batch size of 4 and the same learning rate of 1e-5. Regularization techniques such as weight decay and a warm up schedule were applied to stabilize the training process. Evaluation and checkpointing occurred every 500 steps, with logging every 100 steps. Mixed-precision (fp16) was used for efficiency, and the best model was chosen based on the highest F1-score. The built-in feature extractor handled raw waveform preprocessing, enabling end-to-end training without handcrafted features.

B.4 Evaluation

To assess the performance of DaBloatCNN, ResNet50, and Wav2Vec 2.0, each model was evaluated on a held-out validation dataset using a consistent set of metrics. These metrics provided a balanced view of classification quality, taking into account both general performance and class-specific behavior—especially

important for emotion and sentiment tasks where class imbalance can be an issue.

The primary evaluation metrics included accuracy, loss, and weighted F1-score, which measure overall prediction correctness and class-sensitive performance. Additionally, a classification report was generated for each model to show precision, recall, and F1-score per class. Confusion matrices were used to visualize misclassifications between emotion categories, helping to identify patterns such as overlapping or frequently confused classes. Lastly, Cohen’s kappa was computed to quantify the agreement between predicted and actual labels beyond chance level. This comprehensive evaluation allowed for a fair and insightful comparison across models.

III.C. Deployment

The final system was deployed as a dual-pathway sentiment analysis pipeline to align with our goal: to compare the effectiveness of sentiment classification using transcribed text alone versus raw audio alone. The system processes audio input by dividing it into two parallel streams—one focusing on linguistic content, and the other on vocal tone.

In the text-based branch, the input audio is transcribed using an automatic speech recognition (ASR) model. The resulting text is then analyzed by a suite of NLP sentiment classifiers, including RoBERTa, TinyBERT, and ELECTRA, which predict sentiment based purely on linguistic features.

In the audio-based branch, the same raw waveform is processed directly using speech emotion recognition models—DaBloatCNN, ResNet50, or Wav2Vec 2.0—which infer sentiment from acoustic features such as pitch, tone, and rhythm.

This dual-path design allows a direct comparison of text-only and audio-only sentiment classification. The entire system is deployed as an interactive web application using Streamlit, enabling easy access and real-time testing of both modalities in a user-friendly interface.

IV. Results and Discussion

IV.A Text-Based Sentiment Analysis

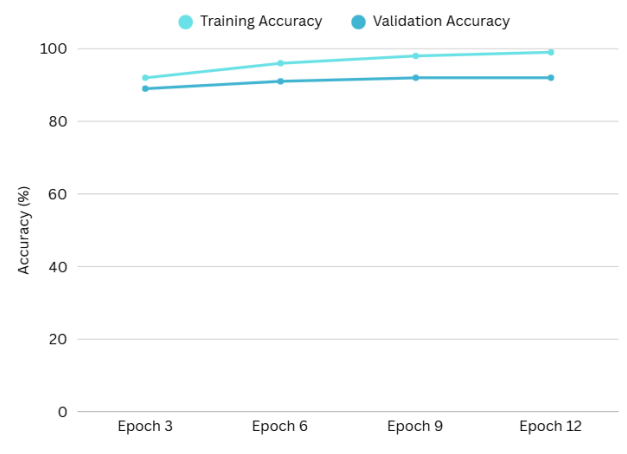


Fig 4.1. RoBERTa Accuracy

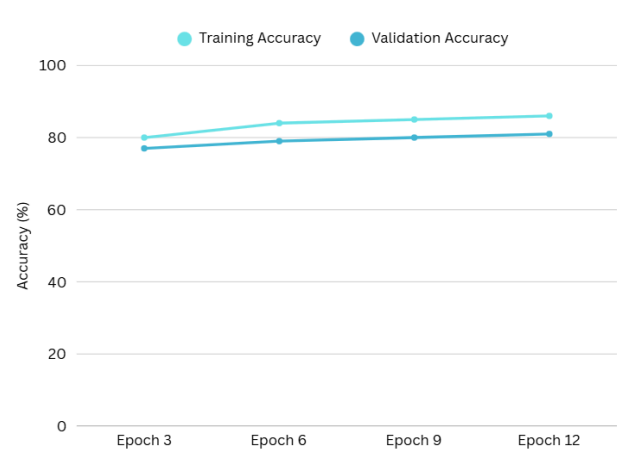


Fig 4.2. TinyBERT Accuracy

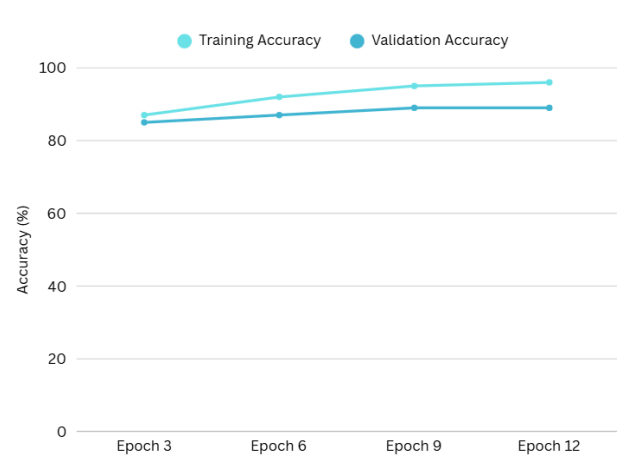


Fig 4.2. ELECTRA-Small Accuracy

Kappa (0.8761), and MCC (0.8761). This indicates that RoBERTa offers the best overall performance among the evaluated models. However, its computational demands may limit its applicability in resource-constrained environments.

In contrast, ELECTRA-Small demonstrated competitive performance with an accuracy of 89%, Macro F1 of 0.8884, and Weighted F1 of 0.8895. While slightly lower than RoBERTa, ELECTRA-Small strikes a favorable balance between accuracy and efficiency. TinyBERT, however, performed significantly worse across all metrics, achieving only 81% accuracy and subpar scores in Macro F1 (0.7979), Precision (0.8061), and Recall (0.8057). This suggests that TinyBERT may not be suitable for this particular sentiment analysis task.

Both Cohen’s Kappa and Matthews Correlation Coefficient (MCC) were used to assess model reliability across all classes. Interestingly, the values matched closely, indicating consistent performance across diverse sentiment categories.

IV.B Voice-Based Sentiment Analysis

Tab 4.2. Overall Evaluation Metrics				
Models	Epoch	Accuracy	F1	Cohen’s Kappa
DaBloatCNN	25	0.962	0.962	0.944
ResNet50	10	0.626	0.627	0.60
Wav2Vec2	5	0.978	0.978	0.967

The three models—DaBloatCNN, ResNet50, and Wav2Vec 2.0—were evaluated using accuracy, F1 score, and Cohen’s Kappa. The results highlight significant performance differences across architectures and input modalities.

Wav2Vec 2.0 achieved the highest performance with an accuracy and F1 score of 0.978, and a Cohen’s Kappa of 0.967. This demonstrates the model’s strong ability to capture emotional tone directly from raw audio, benefiting from its powerful pretraining on large-scale speech data. DaBloatCNN, trained for 25 epochs, followed closely with an accuracy and F1 of 0.962, and a Cohen’s Kappa of 0.944, showing that a custom CNN-based approach can also be highly effective when trained on sufficient data. In contrast, ResNet50, which used spectrograms as input, showed significantly lower performance (accuracy: 0.626, F1: 0.627, Kappa: 0.600), indicating that it struggled to generalize sentiment from visual features alone.

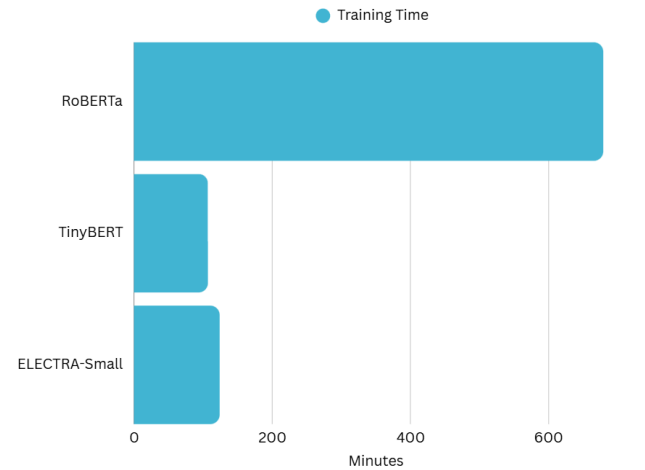


Fig 4.3. Training Time

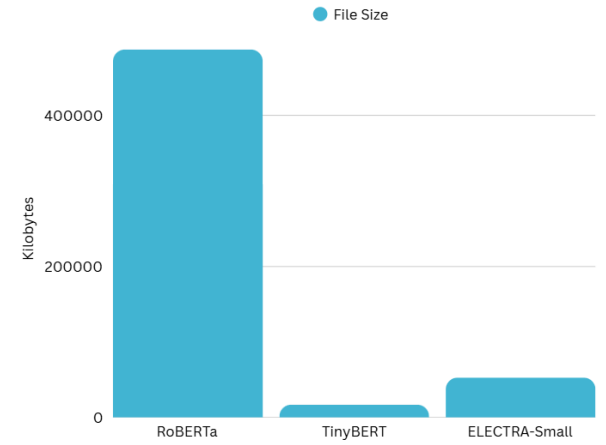


Fig 4.4. File Size

Tab 4.1. Overall Evaluation Metrics

Metric	RoBERTa	TinyBERT	ELECTRA-Small
Accuracy	92%	81%	89%
Macro F1	0.9186	0.7979	0.8884
Weighted F1	0.9196	0.8058	0.8895
Precision	0.9196	0.8061	0.8896
Recall	0.9196	0.8057	0.8896
Cohen’s Kappa	0.8761	0.7007	0.8295
MCC	0.8761	0.7007	0.8295

The metrics used are necessary for tracking the performance of classifying sentiments in the text dataset. Table 1 presents the comparative performance of RoBERTa, TinyBERT, and ELECTRA-Small across various evaluation metrics for sentiment analysis. RoBERTa achieved the highest accuracy (92%) and superior scores in Macro F1 (0.9186), Weighted F1 (0.9196), Precision (0.9196), Recall (0.9196), Cohen’s

These results suggest that direct raw audio models (like Wav2Vec 2.0) are more effective for emotion recognition than spectrogram-based visual models. Moreover, the custom CNN (DaBloatCNN) demonstrates that with well-designed architecture and adequate training, competitive performance can be achieved even without pre-training. This supports the conclusion that audio-based sentiment classification, especially with models like Wav2Vec 2.0, may outperform traditional visual CNNs and even rival text-based approaches.

V. Conclusion and Recommendations

This project set out to explore and compare two different ways of understanding human sentiment from speech. The first approach used the spoken words (transcribed into text), while the second focused on the sound of the voice itself, including tone and emotion. By testing both methods separately, we aimed to find out which one was more accurate and effective for detecting sentiment such as positive, negative, or neutral feelings.

For the text-based models, RoBERTa performed the best overall, reaching 92% accuracy and a strong F1 score of 0.91. It was closely followed by ELECTRA-Small, which also showed excellent results, while TinyBERT was slightly less accurate but still useful, especially for faster or lightweight applications. These results show that advanced language models can accurately understand and classify emotional tone based on the content of what is being said.

On the audio side, Wav2Vec 2.0 delivered the highest performance among all models tested. It achieved 97.8% accuracy and a Cohen's Kappa of 0.967, which means it was not only highly accurate but also consistent in its predictions. The custom-built DaBloatCNN model also performed very well, showing that carefully designed convolutional networks can learn emotional patterns from voice signals. On the other hand, ResNet50, which used visual representations of sound (spectrograms), did not perform as well. This suggests that raw audio models may capture emotional cues more effectively than image-based approaches.

In conclusion, both text and audio models have their strengths. Text-based models like RoBERTa are excellent when clean, accurate transcriptions are available. However, audio-based models, especially Wav2Vec 2.0, can recognize emotions directly from voice tone, making them more flexible in situations where the exact words are unclear or unimportant.

For future work, we recommend combining both approaches into a multimodal system that takes advantage of both the words and the sound of speech. This could lead to even better sentiment detection in real-world settings. We also suggest testing these models on more natural conversations and noisy environments to improve their robustness. Finally, lightweight models like TinyBERT and DaBloatCNN are promising for use in real-time applications, such as a Streamlit-based web app, making this technology more accessible to users and developers alike.

VI. Acknowledgements





The researchers would like to express their gratitude to Engr. Roman M. Richard for his guidance and support throughout the duration of this project. We also acknowledge the resources provided by the Computer Engineering Department at the Technological Institute of the Philippines.

VII. References

- [1] M. Naseer, M. Asvial and R. F. Sari, "An Empirical Comparison of BERT, RoBERTa, and Electra for Fact Verification," *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, Jeju Island, Korea (South), 2021, pp. 241-246
- [2] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding", 2018.
- [3] H. Sohn and H. Lee, "MC-BERT4HATE: Hate speech detection using multi-channel bert for different languages and translations", *IEEE Int. Conf. Data Min. Work. ICDMW*, pp. 551-559, 2019.
- [4] K. Clark, M. T. Luong, Q. V. Le and C. D. Manning, "Electra: Pretraining text encoders as discriminators rather than generators", *arXiv*, pp. 1-18, 2020.
- [5] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint, arXiv:1907.11692*, 2019. [Online]. Available: <https://arxiv.org/abs/1907.11692>.
- [6] X. Jiao et al., "TinyBERT: Distilling BERT for natural language understanding," in *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2352-2362, 2020. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.372.pdf>.

- [7] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," *arXiv preprint, arXiv:2003.10555*, 2020. [Online]. Available: <https://arxiv.org/abs/2003.10555>.
- [8] T. Pavlov and G. Mirceva, "COVID-19 Fake News Detection by Using BERT and RoBERTa models," 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2022, pp. 312-316
- [9] C. Chen and X. Hu, "The Research on an Online Review Sentiment Analysis Model Based on Improved RoBERTa," 2024 3rd International Conference on Electronics and Information Technology (EIT), Chengdu, China, 2024, pp. 624-627
- [10] A. Gaurav, B. B. Gupta and K. T. Chui, "AI-Driven Smishing Detection in Android Devices Using TinyBERT and Aquila Optimization," 2025 27th International Conference on Advanced Communications Technology (ICTACT), PyeongChang, Korea, Republic of, 2025, pp. 99-105, doi: 10.23919/ICTACT63878.2025.10936701.
- [11] A. Qazi, R. H. Goudar, R. Patil, G. S. Hukkeri and D. Kulkarni, "Leveraging BERT, DistilBERT, and TinyBERT for Rumor Detection," in *IEEE Access*, vol. 13, pp. 72918-72929, 2025, doi: 10.1109/ACCESS.2025.3563301.
- [12] M. Rizwan, M. F. Mushtaq, U. Akram, A. Mehmood, I. Ashraf and B. Sahelices, "Depression Classification From Tweets Using Small Deep Transfer Learning Language Models," in *IEEE Access*, vol. 10, pp. 129176-129189, 2022, doi: 10.1109/ACCESS.2022.3223049.
- [13] C. Chen, X. Feng, Y. Li, L. Lyu, J. Zhou, X. Zheng, and J. Yin, "Integration of large language models and federated learning," *Patterns*, vol. 5, no. 12, p. 101098, Dec. 2024. doi: 10.1016/j.patter.2024.101098.
- [14] Z. Örpek, B. Tural and Z. Destan, "The Language Model Revolution: LLM and SLM Analysis," 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Türkiye, 2024, pp. 1-4, doi: 10.1109/IDAP64064.2024.10710677.
- [17] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network Based Speech Emotion Recognition," in *Proc. Interspeech*, 2017, pp. 1263-1267.
- [18] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162-1181, 2006.
- [18] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162-1181, 2006.
- [19] Y. Zhang et al., "Learning Speech Emotion Representations with Cross-Modal Supervision," in *Proc. IEEE ICASSP*, 2021, pp. 6324-6328.
- [20] J. Lee and H. Kim, "Performance Improvement of Speech Emotion Recognition Using ResNet Model with Data Augmentation-Saturation," *Appl. Sci.*, vol. 15, no. 4, p. 2088, 2023.
- [21] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using Wav2Vec 2.0 Embeddings," *arXiv preprint arXiv:2104.03502*, 2021. [Online]. Available: <https://arxiv.org/abs/2104.03502>
- [22] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level Fusion of Wav2Vec 2.0 and BERT for Multimodal Emotion Recognition," *arXiv preprint arXiv:2207.04697*, 2022. [Online]. Available: <https://arxiv.org/abs/2207.04697>
- [23] X. Chen and Y. Zhang, "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neural Network," *Appl. Sci.*, vol. 12, no. 19, p. 9518, 2022.
- [24] X. Zhang and Y. Li, "ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments," *Sustainability*, vol. 15, no. 14, p. 10762, 2023.

Evaluation Metrics

Student Outcome 7							
Criteria	Ratings						Pts
 T.I.P. SO 7.1 Acquire and apply new knowledge from outside sources threshold: 4.2 pts	6 pts [Excellent] Educational interests and pursuits exist and flourish outside classroom requirements, knowledge and/or experiences are pursued independently and applies knowledge learned into practice	5 pts [Good] Educational interests and pursuits exist and flourish outside classroom requirements, knowledge and/or experiences are pursued independently	4 pts [Satisfactory] Look beyond classroom requirements, showing interest in pursuing knowledge independently	3 pts [Unsatisfactory] Begins to look beyond classroom requirements, showing interest in pursuing knowledge independently	2 pts [Poor] Relies on classroom instruction only	1 pts [Very Poor] No initiative or interest in acquiring new knowledge	6 pts
 T.I.P. SO 7.2 Learn independently threshold: 4.2 pts	6 pts [Excellent] Completes an assigned task independently and practices continuous improvement	5 pts [Good] Completes an assigned task without supervision or guidance	4 pts [Satisfactory] Requires minimal guidance to complete an assigned task	3 pts [Unsatisfactory] Requires detailed or step-by-step instructions to complete a task	2 pts [Poor] Shows little interest to complete a task independently	1 pts [Very Poor] No interest to complete a task independently	6 pts
 T.I.P. SO 7.3 Critical thinking in the broadest context of technological change threshold: 4.2 pts	6 pts [Excellent] Synthesizes and integrates information from a variety of sources; formulates a clear and precise perspective; draws appropriate conclusions	5 pts [Good] Evaluate information from a variety of sources; formulates a clear and precise perspective.	4 pts [Satisfactory] Analyze information from a variety of sources; formulates a clear and precise perspective.	3 pts [Unsatisfactory] Apply the gathered information to formulate the problem	2 pts [Poor] Gather and summarized the information from a variety of sources but failed to formulate the problem	1 pts [Very Poor] Gather information from a variety of sources	6 pts
 T.I.P. SO 7.4 Creativity and adaptability to new and emerging technologies threshold: 4.2 pts	6 pts [Excellent] Ideas are combined in original and creative ways in line with the new and emerging technology trends to solve a problem or address an issue.	5 pts [Good] Ideas are creative and adapt the new knowledge to solve a problem or address an issue	4 pts [Satisfactory] Ideas are creative in solving a problem, or address an issue	3 pts [Unsatisfactory] Shows some creative ways to solve the problem	2 pts [Poor] Shows initiative and attempt to develop creative ideas to solve the problem	1 pts [Very Poor] Ideas are copied or restated from the sources consulted	6 pts
Total Points: 24							

Evaluated by:

Engr. Roman M. Richard
 Course Instructor