NED University of Engineering & Technology

Department of Computer & Information Systems Engineering

NLP Project Report

Voice-to-Text with Sentiment & Intent Analysis

[AI-Powered Voice Transcription and Sentiment Analysis System]

Student Name	Khadija, Sundus Baloch	
Roll No/Batch	CS-103/2023-24, CS-110/2023-24	
Program / Specialization	☐	M.Engg (CS) – "Computer Network and System Security"
	☐	MS – Data Engineering & Information Management
	☒	MS – Artificial Intelligence
Subject Instructor	Dr. Shahab Tehzeeb	
Subject	Natural Language Processing	

Table of Contents

1. **Introduction**

With the increasing demand for automated speech processing in industries such as customer service, media analytics, and healthcare, voice transcription systems combined with natural language understanding capabilities have become a key focus in NLP. This project presents a pipeline that transcribes audio using OpenAI's Whisper-small model and performs sentiment analysis using CardiffNLP’s RoBERTa model. The system aims to serve as a foundation for real-world applications requiring voice-based emotion and intent recognition.

2. Objective

- To develop a robust speech-to-text system using OpenAI Whisper.

- To classify sentiments (Positive, Negative, Neutral) using a fine-tuned RoBERTa model.

- To lay the groundwork for intent and urgency detection in conversational audio.

3. Scope

The project is divided into the following key modules:

- Voice Transcription using Whisper.

- Sentiment Analysis using RoBERTa.

- Intent Analysis (Future Scope).

4. Methodology

4.1 Research and Data Collection

Hugging Face was explored for pre-trained models suitable for transcription and sentiment analysis.

Audio data was collected and organized in .wav format, with transcripts saved for supervised fine-tuning.

4.2 Development

OpenAI Whisper is used for transcribing speech. A RoBERTa model is used for sentiment classification. A preprocessing step ensures all audio is standardized to 16 kHz mono audio.

4.3 Testing and Evaluation

The dataset is split into training and testing sets (80:20). Transcriptions and sentiment results are manually verified. Evaluation metrics include transcription accuracy and sentiment classification confidence.

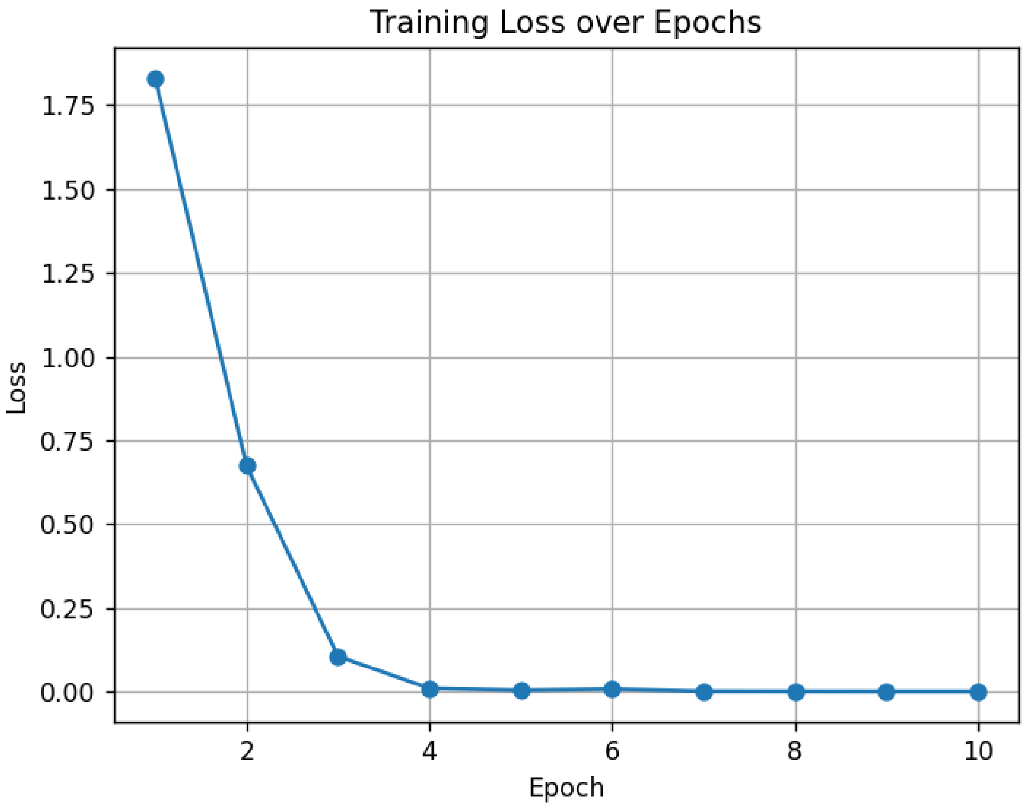


Figure 1 Training Loss Curve

The graph of training loss over epochs illustrates the learning dynamics of the Whisper model during fine-tuning for audio transcription. The initial phase of training is characterized by a steep decline in the loss, indicating a rapid improvement in the model's ability to transcribe the audio data. This suggests that the model quickly identifies and learns fundamental patterns within the dataset relevant to the task. As training progresses, the rate at which the loss decreases gradually diminishes, and the curve begins to plateau. This leveling off signifies that the model is approaching convergence, where subsequent training epochs contribute incrementally less to reducing the transcription error. Such a trend suggests that the model has largely captured the underlying relationships within the training data and that further training on this specific dataset may yield limited additional gains in performance.

5. Tools and Technologies

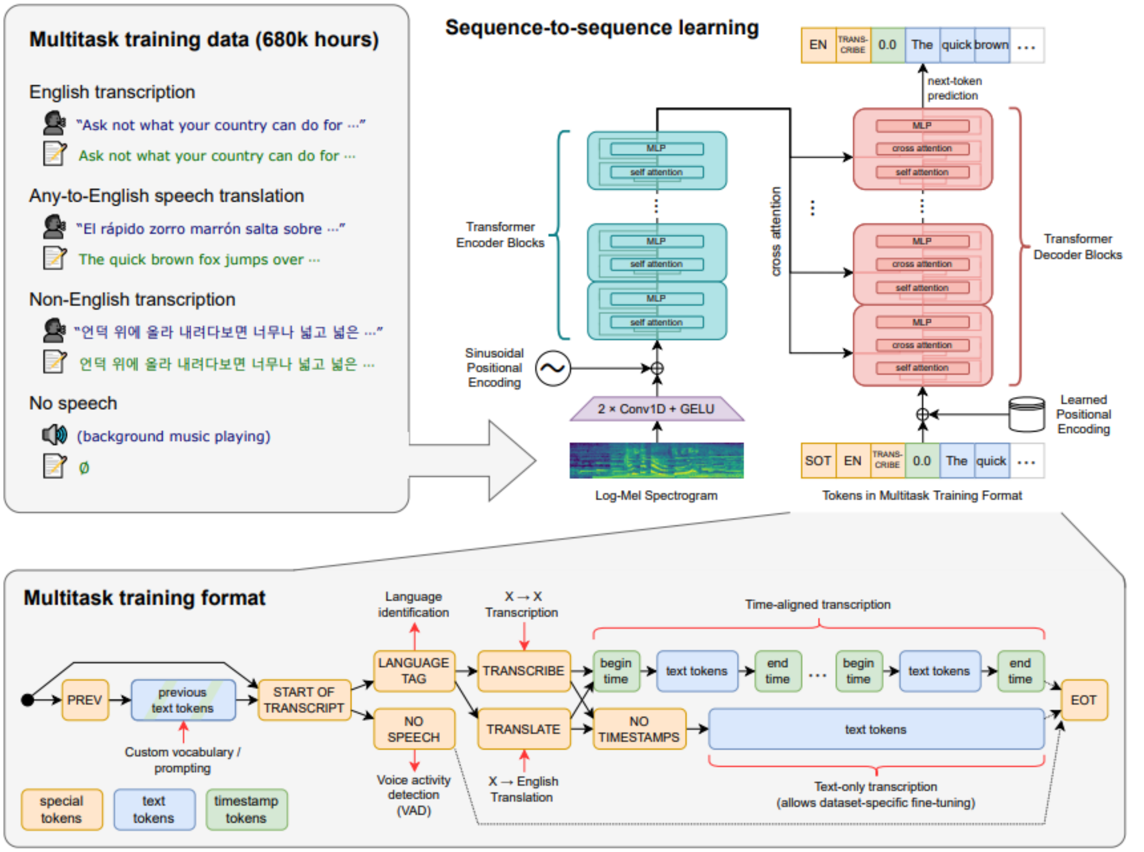
Component	Technology
Speech-to-Text	OpenAI Whisper-small
Sentiment Analysis	CardiffNLP Twitter-RoBERTa
Programming Language	Python
Libraries	Transformers, torchaudio, torch, datasets, matplotlib

6. Model Definition and Inference Pipeline

To develop a comprehensive voice-to-text system integrated with sentiment analysis, two advanced pre-trained models from Hugging Face were utilized: Whisper-small and Twitter RoBERTa-base.

6.1 Model Overview

Whisper Model Architecture



Multitask training data (680k hours)
English transcription
"Ask not what your country can do for …"
Ask not what your country can do for …
Any-to-English speech translation
"El rápido zorro marrón salta sobre …"
The quick brown fox jumps over …
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 …"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 …
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
…
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
…
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
…
begin time
text tokens
end time
NO TIMESTAMPS
text tokens
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure 2 Whisper Model Architecure

Whisper-small is an automatic speech recognition (ASR) model created by OpenAI. It is designed to transcribe spoken language into written text across multiple languages with high accuracy. This model is well-suited for real-world audio scenarios due to its ability to handle background noise and varied speech patterns.

Roberta-GRU Model Architecture

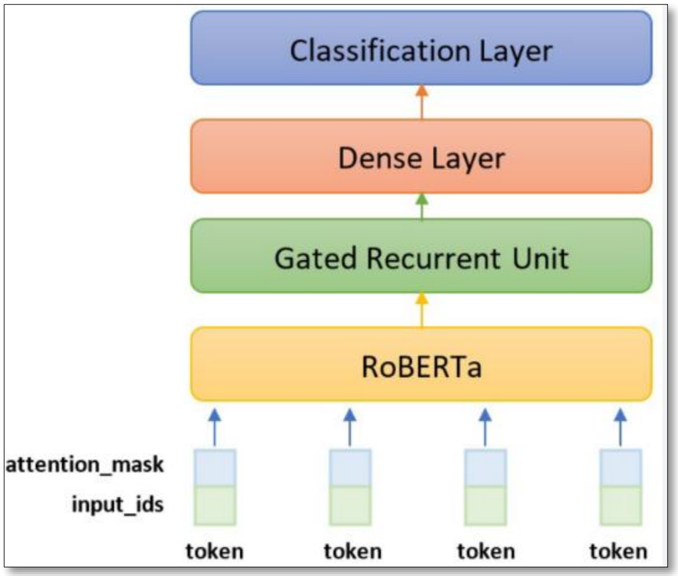


Figure 3 Roberta Model Architecture

Twitter RoBERTa-base is a sentiment analysis model fine-tuned specifically on social media data. It is capable of classifying input text into sentiment categories: Positive, Neutral, or Negative, along with a confidence score that reflects the model's certainty in its prediction.

Tweets Sentiment Analysis

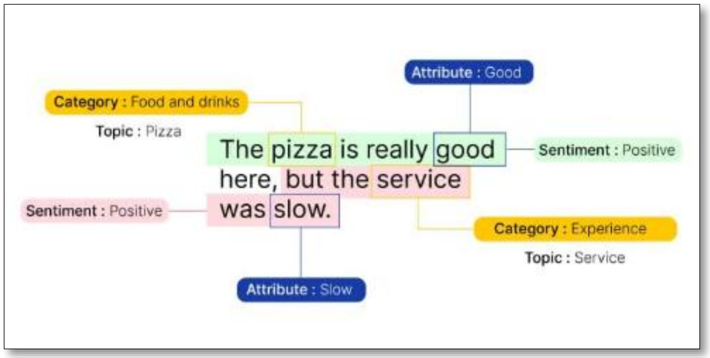


Figure 4 Tweets Sentiment Analysis

6.2 Audio Preprocessing and Transcription Workflow

Before audio is analyzed, it undergoes a series of preprocessing steps. The audio files are first standardized by converting them to mono (single-channel) format and resampling them to a consistent frequency of 16 kHz, which is required for optimal performance of the Whisper model.

This ensures compatibility and reliable transcription results across different recording devices and formats.

Once the audio is in the correct format, it is passed through the Whisper model, which processes the waveform and generates the corresponding text transcription.

6.3 Sentiment Analysis Integration

The transcribed text is then analyzed for sentiment using the RoBERTa sentiment model. This step involves evaluating the emotional tone of the text and assigning it a label: Negative, Neutral, or Positive. Each label is accompanied by a numerical confidence score, indicating how strongly the model believes in its prediction.

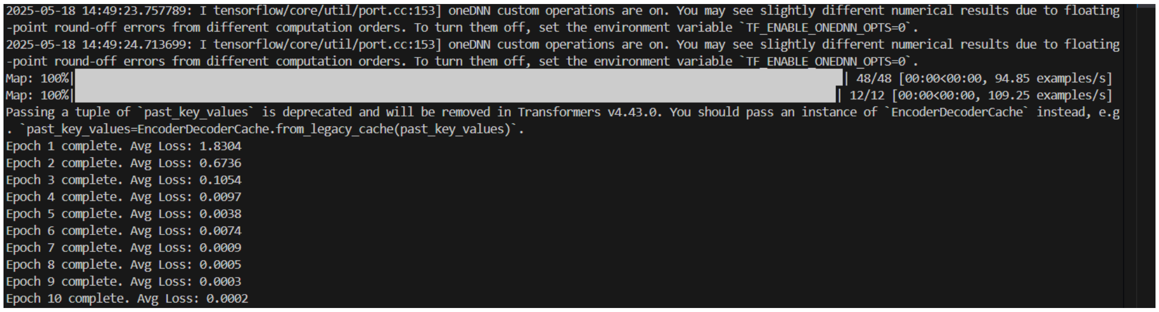
6.4 Sample Result

For instance, in one test case involving an emotionally charged audio clip, the system transcribed the speech as a clear expression of frustration and correctly classified the sentiment as Negative with high confidence. This highlights the system's capability to not only understand spoken language but also assess the underlying emotional context, making it suitable for applications in customer service monitoring, behavioral analysis, and support systems.

6.4 Fine-Tuning and Model Adaptation

To further improve the model's performance on our specific audio dataset, we performed a fine-tuning process on the Whisper-small model. This was done to adapt the model more effectively to the unique characteristics of our recordings, which may include varying speech patterns, accents, or emotional tones not fully represented in the original pretraining dataset.

The fine-tuning strategy involved freezing the encoder part of the model, which is responsible for extracting audio features, while allowing the decoder—the part that generates text—to learn from our data. This approach reduces computational requirements while still providing adaptation benefits.

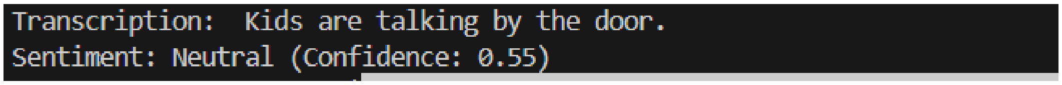


```
2025-05-18 14:49:23.757789: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different numerical results due to floating
-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-05-18 14:49:24.713699: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different numerical results due to floating
-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
Map: 100%|██████████████████████████████████████████████████████████████████| 48/48 [00:00<00:00, 94.85 examples/s]
Map: 100%|█████████████████████████████████████████████████████████████████| 12/12 [00:00<00:00, 109.25 examples/s]
Passing a tuple of `past_key_values` is deprecated and will be removed in Transformers v4.43.0. You should pass an instance of `EncoderDecoderCache` instead, e.g
. `past_key_values=EncoderDecoderCache.from_legacy_cache(past_key_values)`.
Epoch 1 complete. Avg Loss: 1.8304
Epoch 2 complete. Avg Loss: 0.6736
Epoch 3 complete. Avg Loss: 0.1054
Epoch 4 complete. Avg Loss: 0.0097
Epoch 5 complete. Avg Loss: 0.0038
Epoch 6 complete. Avg Loss: 0.0074
Epoch 7 complete. Avg Loss: 0.0009
Epoch 8 complete. Avg Loss: 0.0005
Epoch 9 complete. Avg Loss: 0.0003
Epoch 10 complete. Avg Loss: 0.0002
```

Figure 5 Whisper Model Finetune Result

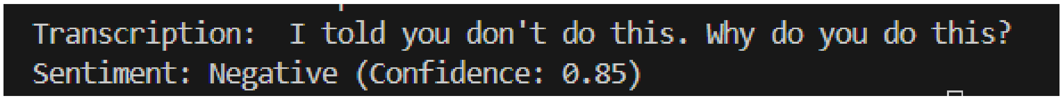
7. Results

The following outputs show transcriptions and sentiment predictions for different audio inputs:



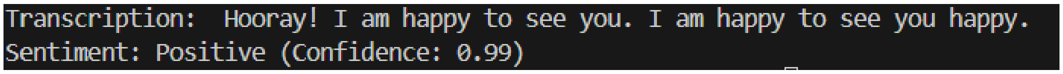
```
Transcription:  Kids are talking by the door.
Sentiment: Neutral (Confidence: 0.55)
```

Figure 6 Model Test: Neutral result



```
Transcription:  I told you don't do this. Why do you do this?
Sentiment: Negative (Confidence: 0.85)
```

Figure 7 Model Test: Negative result



```
Transcription:  Hooray! I am happy to see you. I am happy to see you happy.
Sentiment: Positive (Confidence: 0.99)
```

Figure 8 Model Test: Positive result

8. Expected Outcomes

- A working prototype that transcribes audio.

- Sentiment analysis with reliable predictions.

- Scalable design for future intent and urgency detection.

9. Challenges and Risks

- Audio noise and accents can reduce accuracy.

- Sentiment models may struggle with sarcasm or ambiguous expressions.

- Real-world datasets with diverse samples are needed for generalization.

10. Conclusion

This project successfully demonstrates the integration of advanced models for speech-to-text and sentiment analysis. The system offers valuable insights for industries handling spoken communication and sets a foundation for future enhancements in intent detection and real-time analysis.

11. References

1. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust Speech Recognition via Large-Scale Weak Supervision*. OpenAI. https://cdn.openai.com/papers/whisper.pdf
2. Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification*. In Proceedings of the Findings of EMNLP 2020. https://arxiv.org/pdf/2010.12421