**Model Name**:  
**Speech Emotion Recognition with OpenAI Whisper Large V3**

**Hugging Face Link**:  
[firdhokk/speech-emotion-recognition-with-openai-whisper-large-v3](https://huggingface.co/firdhokk/speech-emotion-recognition-with-openai-whisper-large-v3)

**Architecture/Backbone**:  
Fine-tuned version of OpenAI's Whisper Large V3, an encoder-decoder transformer model originally designed for speech recognition.

**Task**:  
Audio-based Emotion Classification (Speech Emotion Recognition)

**Dataset Used for Pretraining**:  
Combined datasets:

* RAVDESS
* SAVEE
* TESS
* URDU

Emotion distribution:

* Sad: 752 samples
* Happy: 752 samples
* Angry: 752 samples
* Neutral: 716 samples
* Disgust: 652 samples
* Fearful: 652 samples
* Surprised: 652 samples
* Calm: 192 samples (excluded from training due to underrepresentation)

**Input Format**:

* Audio files processed using Librosa to load and convert to NumPy arrays.
* Feature extraction via Whisper Feature Extractor, standardizing and normalizing audio features.
* Sampling rate: 16,000 Hz (as per Whisper's requirements)

**Output Format**:  
Emotion labels:

* Angry
* Disgust
* Fearful
* Happy
* Neutral
* Sad
* Surprised

**Languages Supported**:  
Primarily English, with some inclusion of Urdu (from the URDU dataset). However, the model's performance is optimized for English speech emotion recognition.

**Model Size**:  
Based on Whisper Large V3:

* Parameters: Approximately 1.55 billion
* Memory footprint: Requires significant GPU memory (recommendation: ≥16GB VRAM)

**Evaluation Metrics**:

* Loss: 0.5008
* Accuracy: 91.99%
* Precision: 92.30%
* Recall: 91.99%
* F1 Score: 91.98%

