Performance Analysis between different Machine Learning Models for Emotion Analysis

Soham Bhokare, Omkar Masur, Ozan Erdal, Garima Dave and **Yutao Huang**
University of Southern California, Los Angeles, CA 90007

Abstract

In this paper, we perform a comprehensive study of text-only, audio-only, and multimodal models using text and audio for Human Speech Emotion Recognition (SER). We also propose a novel method for emotion detection using text and audio. Our approach involves extracting each word from the input audio and concatenating it with each text word, creating a joint audio-text representation for emotion detection.

1 Introduction

Emotion recognition models help doctors to diagnose diseases such as depression and dementia by identifying emotion patterns using voice analysis. It can help analyze stress and anxiety levels, which can help in early diagnosis. It can also help increase the usefulness of virtual assistants such as Google Assistant by giving them more data to respond to user input accurately. Similar benefits would be seen with Chatbot accuracy and usefulness.

The project's primary goal is to analyze different machine learning/deep-learning models that can accurately predict emotions based on text and corresponding audio inputs. We start by analyzing the models that only accept the text as input, which include traditional SVM, Perceptron, and Naive Bayes models, then move to Deep learning based LSTM models, then finally use the newly invented Transformer models. We then analyze the possibility of accuracy gain when using a multimodal model that accepts text and audio as input, similar to the model proposed in (Akbari et al., 2021). Features such as tone, pitch, and loudness in speech data are analyzed alongside the text to predict emotions.

Natural Language Processing and encoding methods like Word2Vec are used to extract the semantics of the sentence, which can be correlated to different emotions. Different encoding and feature extraction methods are explored to get the best results. We use spectral features and pass them into the various models for audio.

2 Related Work

Emotion detection using Natural Language Processing has been a popular area of research. It is imperative to social media, where tweets can be analyzed to identify how a person is feeling. (Desmet and Hoste, 2013) presents an application where texts in suicide notes are analyzed to detect their emotions. (Gaind et al., 2019) classifies tweets into six categories: Happiness, Sadness, Fear, Anger, Surprise, and Disgust. (Acheampong et al., 2021; Kane et al., 2022; Prasad et al., 2022) present a transformer-based approach to detecting emotions in pure text. (Talegaonkar et al., 2019; Minaee et al., 2021) gives insight into detecting emotions based on facial expressions captured from images using CNNs. Transformer-based emotion detection using voice is explained in (Gaind et al., 2019)

Multimodal learning enables the model to learn from multiple forms of data rather than relying on one. For emotion recognition, analyzing text alone may not be enough to recognize the underlying emotion, which is more evident in speech analysis. However, this form of learning poses multiple challenges. (Akbari et al., 2021) proposes a novel architecture called Video, Audio, Text Transformer (VATT). As the name suggests, it takes a video file, audio, and text as input, and performs classification tasks. The model uses a transformer model and encodes the input modality using an encoding layer before passing it to the transformer layer.

We analyze the advantages of multimodal learning over unimodal learning by using the models described in the individual papers mentioned above. We propose a model similar to the one proposed in (Akbari et al., 2021) for multimodal learning.

3 Dataset

Our dataset is from the USC website: the Interactive Emotional Dyadic Motion Capture (IEMO-

CAP) Database[1]. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions from 10 male and female actors during their effective dyadic interaction. Multiple annotators annotate the dataset into categorical and dimensional emotions, which include Neutral, Happy, Sad, Angry, Surprised, Fear, Disgusted, Frustrated and Excited. The data is divided into a total of 5 sessions.

4 Data Preparation

Research generally focuses on four emotions — Neutral, Angry, Sad, and Happy, so we first extracted these. We split the dataset into a ratio of 80:20 for training and testing, so across the five sessions, we got a total of 3592 training examples and a total of 898 testing examples. Input data with no consensus emotion label or with a consensus label outside of the 4 labels $y \in \{neu, ang, sad, hap\}$ were dropped.

Text-only data: The IEMOCAP data is prepared such that each instance of dialogue is a single data point. The inputs are sequences of English words, and the labels are the corresponding emotion tags for the sequence. Stop word removal and/or lemmatization is performed on the input text for cases where doing so is necessary. The data is also converted to various embeddings such as TF-IDF, GloVe, RoBERTa, etc.

Audio-only data: For feeding data into the audio-only and multimodal model, we extract features from the audio data using the Python Librosa Library. As proposed in (Zhang et al., 2018), we extract the following five spectral representations of each audio sample: Mel-frequency Cepstral Coefficients, Mel-scaled spectrogram, Chronogram, Spectral contrast feature, and Tonnetz representation. We then concatenate each extracted spectral representation into a single vector and pass an input of size 281×126.

Multimodal data: We use previously described techniques and extract features for the text and audio individually before passing them into the model.

Multimodal using segmented audio and text data: The emotion of a given sentence can be predicted not just by how the whole sentence is interpreted, but also by how microscopically each word is said out loud. For example, the emotion could be interpreted as either neutral or angry based on how loud "No" is vocalized. Based on this intuition, we propose a novel method for emotion detection where we feed individual text-based words *and* individually segment words from the audio input to the model. We segment the audio samples into their constituent word segments, which include individual word audio and some overlap between the next and previous word audios. Features similar to the audio-only model were extracted for training for each of these segmented word-level audio samples. We then concatenate these extracted features with the Word2vec features of text and pass it to a model. We give a detailed overview of the segmentation module in Section 5.

5 Audio Segmentation Module

Python audio manipulation modules like Pydub have audio segmentation functions that split audio based on silences between the different words in the file. This works well for speech audio with consistent pauses between words. However, upon further analysis, we observed many dialog audios without sufficient pauses between words to accurately segment the audio.

Figure 1 shows the audio for an example where there are sufficient pauses between each word (as observed in the waveform). Using Pydub and manually fine-tuning the silence length and threshold parameters, we could segment the audio into seven segments. However, the audio contains 23 words, and thus we need 23 segments. The results are even worse when we apply the segmentation function to an audio example where the pauses between words are minimal, as shown in Figure 2. As seen in Fig 3, we only get 2 segments compared to the expected 16 segments.

The outcomes from using our developed audio segmentation module are illustrated in Figure 4. The number of segments produced corresponds to the number of words in the audio. Nevertheless, some words, like "see", represented by segment 5, can be effortlessly recognized by the Wav2Vec model even with a frame size smaller than the sampling rate, resulting in the generation of a frame that provides a sparse waveform. This is one of the aspects of the module that can be improved. The segments also require a certain overlap between the previous and next segments for the module to accurately recognize the word produced in the center of the frame. Thus, the resulting audio segment includes the overlap. This overlap can be removed in post-processing, and the result of doing so is yet

[1] https://sail.usc.edu/iemocap/index.html

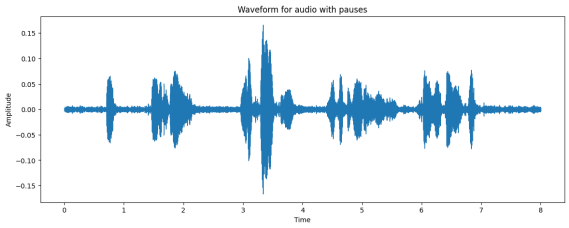

Figure 1: Waveform for audio without pauses

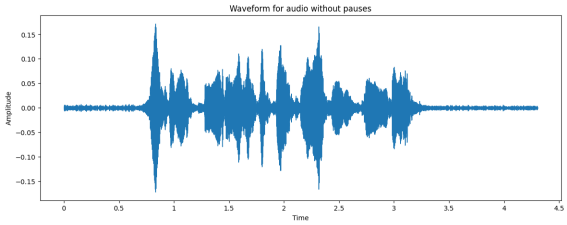

Figure 2: Waveform for audio without pauses

to be analyzed.

6 Models

Text-only models

For the text portion of the (Speech Emotion Recognition) SER analysis, we explore several text classification models in a unimodal manner using a variety of word embeddings. The comprehensive list of experimental accuracies is shown in Table 1. The most effective model was then used for the multimodal SER task. From our exploration, we found that text-based transformers performed the best. Specifically, we used RoBERTa base, a large transformer model from HuggingFace, for tokenizing our text. RoBERTa was created using self-supervised training on the Masked Language Modeling (MLM) task, and recent research has shown that this lends itself to better generalization on an unrelated downstream task such as SER. We perform batch-less optimization using the Adam optimizer and Cross-Entropy Loss; using a learning rate of 10^{-6}, our model converges in around 25 epochs.

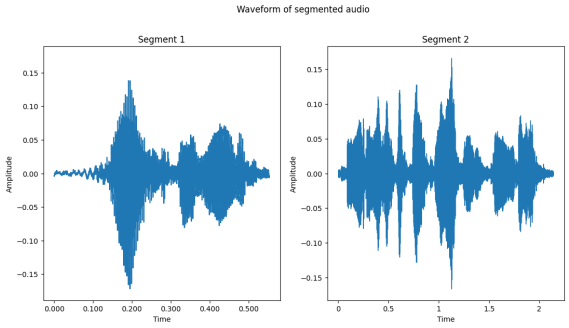

Figure 3: Audio Segmentation using Pydub for Figure 2

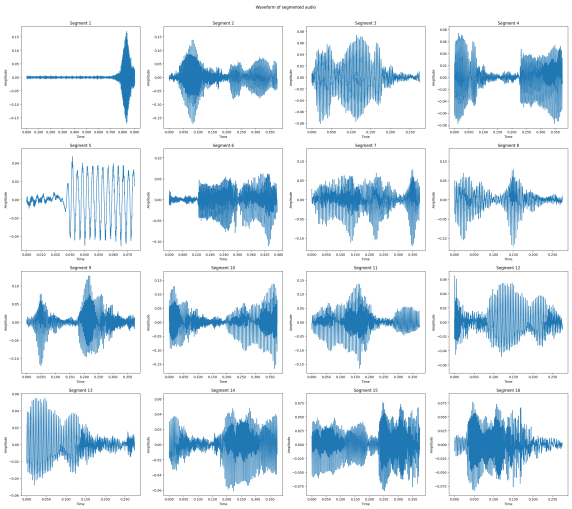

Figure 4: Audio Segmentation using our own module for Figure 2

Audio-only model

We propose a CNN-based classifier that takes input audio features extracted from Section 4. The architecture for the same is presented in Figure 5

Multimodal model using shallow fusion

Upon training the text-only model and Audio-only models, we use the best-performing text-only model and the audio-only model in building the multimodal model. Our proposed model architecture removes the final Linear layers from both models and concatenates the outputs after passing the outputs from the resultant model. We then pass these concatenated outputs to a classifier. The proposed model architecture is shown in Figure 6

7 Results

We evaluate the proposed models using three metrics; Accuracy, Macro-F1, and Weighted F1. These results comply with numerous state-of-the-art models proposed in (Akbari et al., 2021) and (Zhang et al., 2018). The results for the text-only model is shown in Table 1, that for audio-only model is shown in Table 2 and that for Multimodal models is shown in Table 3.

8 Conclusions

Our multimodal noticeably outperforms the unimodal text-only and audio-only models. However, our suggested segmented model results in a lower accuracy when compared to a traditional multimodal approach where the entire text and audio

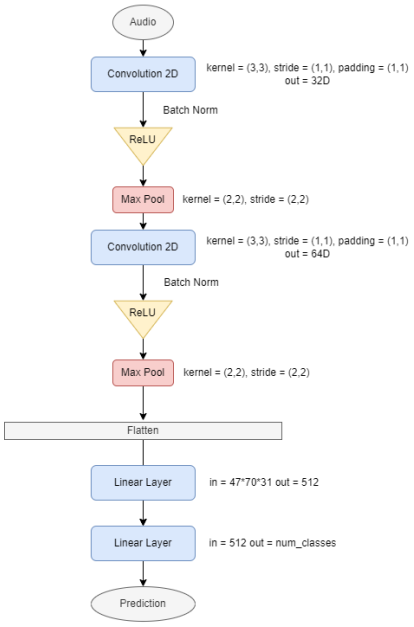

Figure 5: CNN Model for classifying Audio

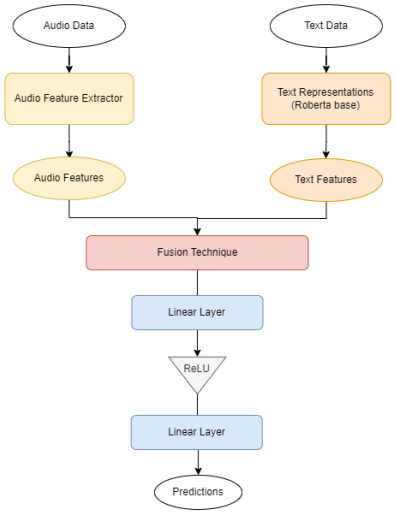

Figure 6: Multimodal Model using Shallow Fusion

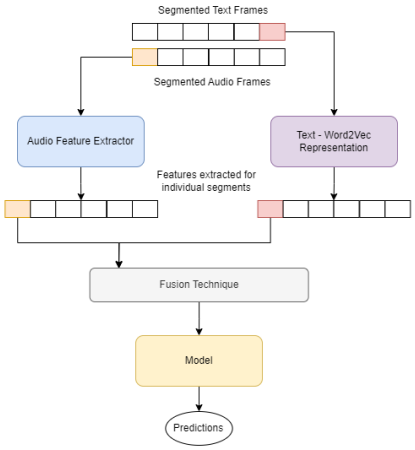

Figure 7: Multimodal Model using Segmented Audio and Text Inputs

Model	Accuracy	f1$_{macro}$	f1$_{weight}$
Perceptron	57.6%	55.7%	57.4%
SVM	61.0%	59.1%	60.1%
Logistic Regression	59.8%	55.8%	58.8%
Naive Bayes	58.7%	51.7%	56.3%
XGBoost	60.8%	58.3%	60.4%
Random Forest	59.9%	54.6%	58.4%
RoBERTa	**69.4%**	**68.7%**	**69.2%**

Table 1: Results for Text-only models

Model	Acc	f1$_{macro}$	f1$_{weight}$
(Zhang et al., 2018)	64.3%	–	–
Proposed model (5)	**64.35%**	**54.4%**	**60.4%**

Table 2: Results for Audio-only models

features are combined together. The biggest factor is the lack of a dataset where speech audio is pre-segmented into its respective word audios. Thus, the final accuracy largely depends on our audio segmentation module's accuracy in recognizing individual words and segmenting them. Furthermore, when words are not recognized in the audio, the module defaults to segmenting the audio into interval widths based on the character length of individual words. This can generate noise in the training data, as the aforementioned process doesn't always yield matching word and audio segments.

9 Future Scope

Our novel architecture, which takes the concatenated features of text and audio representations, currently uses a normal LSTM without any attention mechanism. Yet, it performs considerably well. Hence, work can be done where we can incorporate attention mechanisms into it to further exploit the latent relationships between text and audio. Not just that, but we could also use RoBERTa or other pre-trained models. There is also scope for improvements in the post-processing done by the audio segmentation module and better default segmentation algorithms in cases where a word in the sentence audio is not recognized. Nevertheless, we strongly feel there are gains to be made by interlinking text and audio features individually.

Model	Acc	f1$_{macro}$	f1$_{weight}$
Model (Fig. 6)	72.04%	70.1%	71.8%
Model (Fig. 7)	65.29%	60.1%	62.05%

Table 3: Results for Multimodal models

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems* **2021**, *34*, 24206–24221.

Desmet, B.; Hoste, V. Emotion detection in suicide notes. *Expert Systems with Applications* **2013**, *40*, 6351–6358.

Gaind, B.; Syal, V.; Padgalwar, S. Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458* **2019**,

Acheampong, F. A.; Nunoo-Mensah, H.; Chen, W. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review* **2021**, 1–41.

Kane, A.; Patankar, S.; Khose, S.; Kirtane, N. Transformer based ensemble for emotion detection. *arXiv preprint arXiv:2203.11899* **2022**,

Prasad, J.; Prasad, G.; Gunavathi, C. GJG TamilNLP-ACL2022: emotion analysis and classification in Tamil using Transformers. Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages. 2022; pp 86–92.

Talegaonkar, I.; Joshi, K.; Valunj, S.; Kohok, R.; Kulkarni, A. Real time facial expression recognition using deep learning. Proceedings of international conference on communication and information processing (ICCIP). 2019.

Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046.

Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia* **2018**, *20*, 1576–1590.