FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

STUDENT
SCIENTIFIC CONFERENCE 2024

Multi-modal Emotion Analysis in Textual and Audio data

Matěj Zeman[1]

1 Introduction

Emotions are a complex behavioral phenomenon. Humans express emotions in various ways, but the most common ones involve speech characteristics and facial expressions. The way a sentence is said can change its meaning drastically and the incorporated emotion takes a big part in the way a sentence is said. Understanding the text alone is insufficient if we want to understand the whole semantics of a sentence. It is thus beneficial to incorporate nonlinguistic information such as emotion with the content of the sentence together.

2 Proposed Solution

The goal of this thesis was to implement a system for multimodal emotion recognition. To accomplish that, 3 audio emotion recognition models and 2 textual emotion recognition models were implemented. The multimodal models combine the textual emotion recognition models with the audio emotion recognition models by taking the audio feature vector from the trained audio emotion recognition model and concatenating it with the textual representation of the sentence taken from the textual emotion recognition model as can be seen in Figure 1.Furthermore, 2 audio feature extraction and 2 text feature extraction methods were implemented.

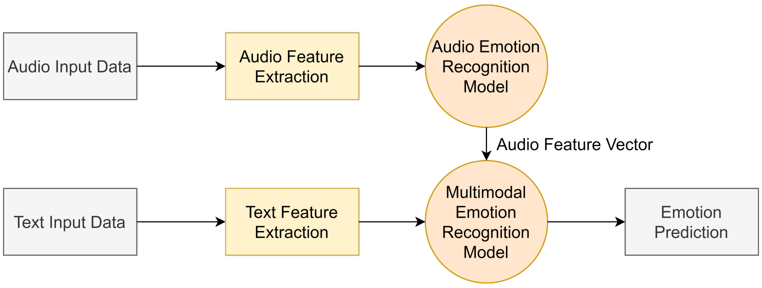

Figure 1: Multimodal emotion recognition pipeline

As said above, there are two audio feature extraction methods implemented:

- **MFCC only** - This feature extraction method uses only the Mel Frequency Cepstral Coefficients.

- **Collective Features** - This method is called collective features because there are several feature extraction methods that collectively create the feature vector. The vectors or scalars from these feature extraction methods are then concatenated and used as the feature vector.

[1] student of the master degree program Applied Sciences, field of study Software Engineering, e-mail: zemanm98@students.zcu.cz

The two mentioned text feature extraction methods are **Word2Vec** and word embeddings from the **BERT** pretrained model. To create the audio feature vectors used in multimodal emotion recognition, three audio models were implemented:

- **CNN1D** - This model is based on Convolutional Neural Networks and has only one dimensional kernels used for its convolution operation, there are a total of 4 convolution layers with a maximum over-time pooling between each layer and 3 linear layers.
- **CNN2D** - This model resembles more conventional convolutional models. There are two convolution layers and two linear layers.
- **MLP** - This model is the least complicated one and consists of only 3 linear layers.

The multimodal models combine the textual emotion recognition models with the audio feature vectors from the audio models. There are two implemented multimodal models:

- **LSTM** - This model consists of two LSTM layers and one linear layer for classification purposes. Both of the LSTM layers are bidirectional and the output from the first one is then passed through the Attention layer. The output from the Attention layer is concatenated with the audio feature vector and used as input into the second LSTM layer of the model.
- **BERT** - This model uses the pretrained *bert-base-uncased* model as a first layer and one linear layer as a classifier. It creates a representation of the text sentence. The output from the BERT layer is concatenated with the audio feature vector and used as input into the linear layer.

3 Results and Conclusion

Model	**Text feature extraction method**	**Audio feature extraction method**	**Train acc**	**Train F1**	**Test acc**	**Test F1**
ECF	audio	CNN1D + collective_features	70.5	45.6	40.8	18.7
	text	BERT	81.2	71.2	62.3	43.8
	audio + text	BERT + CNN1D + collective_features	73.4	51.8	**62.4**	**44.8**
RAVDESS	audio	MLP + collective_features	86.8	84.4	**55.3**	**51.6**
IEMOCAP	audio	CNN1D + collective_features	89.6	89.4	51.4	51.5
	text	BERT	93.7	92.7	60.5	55.8
	audio + text	LSTM + w2v + CNN1D + mfcc_only	90.4	90.4	**79.6**	**79.5**

Table 1: Best performing configurations for each modality and on each dataset

The maximum average F1 score for audio emotion recognition was achieved by the **CNN1D** model reaching up to **18.7%** on the **ECF** dataset and **51.5%** on the **IEMOCAP** dataset. The maximum average F1 score on the **RAVDESS** dataset was achieved by the **MLP** model with F1 score of **51.6%**. On the text emotion recognition task, the **BERT** model performed best and achieved an average F1 score of **43.8%** on the **ECF** dataset and **55.8%** on the **IEMOCAP** dataset. The additional information added by the audio feature vector proved to be beneficial for emotion recognition since both the multimodal **BERT** and **LSTM** models achieved better results with the audio feature vectors. Especially the **LSTM** model improved by over **15%** on accuracy and over **20%** on F1 score on the **IEMOCAP** dataset.