Master Computer Science

Multi-modal Emotion Recognition

Name:	Michail Kolettis
Student ID:	s2689758
Date:	22/07/2022
Specialisation:	Computer Science: Data Science
1st supervisor:	Dr. Erwin M. Bakker
2nd supervisor:	Prof. Dr. Michael S. Lew

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Emotion recognition constitutes the task of detecting human emotions. Throughout the last decade, a lot of researchers have experimented and tried to take advantage of the technology and its applications, such as Affective Computing, where human non-verbal signs are analyzed by computers, to handle the task of human emotion detection. Different modalities, such as text, speech, or video are widely used in many projects trying to effectively classify human emotions. The goal of this project is to investigate the field of human emotion recognition and create a multi-modal emotion recognition system that will effectively manage to identify human emotions. To achieve that, three uni-modal systems, EmoT, EmoA, and EmoFH are created and evaluated using text, audio, and visual features respectively from the IEMOCAP dataset. Next, the three different modalities are fused and two multi-modal emotion recognition systems are proposed. Both, the TMER model, a novice Convolutional Neural Network (CNN) architecture, and the TMERT model, a transformer-based architecture refined the performance of the previous uni-modal systems and manage to efficiently classify the human emotions. Experiments showed that the TMERT model managed to reach competitive results with an accuracy score equal to 81.32%, which is almost 5% higher than the baseline [18], and a F1-score equal to 81.27%. Recently, after our studies, Abhinav Joshi et al. in [11] manages to better classify human emotions, by using a Graph Neural Network (GNN) based architecture, reaching an F1-score equal to 84.50%.

Contents

1 Introduction

Emotion recognition constitutes the dynamic process of human emotion identification. Recognizing human emotion is important because it may readily affect how people behave in day-to-day encounters. Furthermore, research areas like Human-Computer Interaction (HCI), Human Robot Interaction, etc. use the automatic human emotion recognition field and its applications to handle tasks such as social media analysis, human-machine dialogue generation, or even clinical diagnosis by using facial micro-expressions.

In general, people use many different ways to express their feelings both verbally and non-verbally, by using facial expressions, hand movements, expressive speech, text, etc. Recognizing the underline emotion is really useful for meaningful communication between people.

A human's accuracy at identifying another person's emotion may vary widely. Thus, the technology to recognize different human emotions becomes a very challenging and important research area.

The emotion recognition task constitutes a hot topic for the scientific community and many researchers try to create state-of-the-art systems to solve that problem. Different modalities, such as text, speech, or video are used in the task of human emotion recognition.

Over the years, scientists have created and assessed a wide range of different techniques for automatically identifying human emotions. Hundreds of different sorts of methods are now being proposed and evaluated in literature by combining methods from several disciplines, including speech processing, text mining, and computer vision.

Even though each of these areas greatly contributes to emotion recognition using one modality respectively, recognizing a person's emotion by taking into consideration only a single modality, may lead to false interpretations. Emotion detection accuracy is usually enhanced when expressions from multi-modal forms are combined. Therefore, a combination of multiple modalities may be analyzed to define a human's emotion accurately.

In this project, the different features from multiple modalities, such as text, audio, and video are extracted from IEMOCAP dataset, which consists of approximately twelve hours of recordings of ten actors in dyadic sessions. The IEMOCAP dataset is widely used in multi-modal emotion recognition researches, including [11], that reaches state-of-the-art results.

Specificaly in this project, language embeddings are used in the EmoT model, a BERT-based model that detects human emotion using text data, while vectors of Mel-frequency Cepstrum Coefficients combined with Log mel spectrograms and Chromagrams are used in EmoA, a Convolutional Neural Network (CNN) model, which classifies the human emotions using audio data. Finally, facial expressions together with hand movements are fed into the EmoFH model, which is another CNN model, that detects emotions by taking into consideration visual features.

After the uni-modal emotion recognition analysis, the three used modalities are fused to improve the performance of the uni-modal models mentioned above. Two multi-modal models, TMER and TMERT are created. TMER constitutes a CNN model, while TMERT is a transformer-based model. Both models managed to effectively classify human emotions, achieving competitive results in the task of human emotion recognition.

The outline of this research is as follows. In Section 2, the contribution of this project is mentioned and Section 3, presents the existing literature on the uni-modal and the multi-modal emotion recognition tasks. Next, in Section 4, fundamental concepts are analyzed. The description of the baseline method used in this project follows in Section 5. Furthermore, in Section 6, the proposed uni-modal and multi-modal models and their architecture are explained in detail. Then, the description of the dataset used in this project, the feature extraction methods, and the preprocessing of the data, are described in Section 7. The experimental setup together with the final results of each model compared with the baseline are shown in Section 8 and finally, Section 9 includes the major conclusions of this research.

2 Contribution

In this paper, the task of human emotion recognition is investigated, and three different modalities (text, audio, video) are analyzed to evaluate their effectiveness in emotion detection. The strategy of this project is based on investigating and creating an individual architecture for each modality and finally fusing them in the final layer for a more robust human emotion detection model.

Deep learning uni-modal models, which use text, audio, and visual features are created and features, that include contextual information (sentence embeddings) are fused with features from other modalities, such as audio and video, and they are used as an input in the multi-modal models.

The performance of each proposed model is evaluated against the baseline [18], which is also focused on classifying four different categories of human emotion using the IEMOCAP dataset, and resulted in competitive results in terms of accuracy in the multi-modal emotion recognition task using three different modalities, until May 2022, after our studies, where Abhinav Joshi et al. in [11] presented the GNN Contextual Multi-modal Emotion Recognition Model (COGMEN), the state-of-the-art model in the multi-modal emotion recognition task.

The contribution of this project:

- Each modality is analyzed in detail, by extracting informative features and applying them to multiple architectures.
- State-of-the-art models, such as BERT, that uses contextual relations between the words of the input text, are used and experimented with text data, while deep neural networks, like Convolutional Neural Networks (CNNs), are presented to handle the emotion classification on audio and visual data.
- A novel CNN architecture, TMER is firstly investigated and results on emotion detection using the combination of three modalities are obtained.
- Furthermore, another architecture, TMERT that uses transformers is also created to further experiment and finally enhance the emotion classification accuracy. In comparison with the existing literature at the period of the experiments, the proposed multi-modal systems manage to achieve competitive results in human emotion classification.

3 Related Work

For decades emotion recognition has been a very active research topic, with many recent state-of-the-art models, that efficiently classify human emotions, by using deep learning techniques. In this section, an overview of the related work in Emotion Recognition is given.

3.1 Unimodal Emotion Recognition

Most of the earliest attempts to address the emotion recognition challenge have been made using only one modality.
KHALFALLAH et al. in [13] developed a facial emotion recognition system to help students in their learning process. The "Clmtrackr" package was utilized, which follows a student's face and records the coordinate locations of the face model in an array. After identifying the face, the system will start to modify the facial model and measure the rate of various emotions including surprise, rage, sadness, and happiness.

Mustaqeem et al. [21] proposed a speech emotion recognition system. Speech signals often include noise, so to remove those, a dynamic adaptive threshold technique was used. The remaining signals were converted to spectograms, which were utilized by stride CNN architectures to learn the major and discriminative features resulting in a robust and effective model.

Tursunov et al. in [27] also studied the speech emotion recognition task and suggested a CNN model that learns deep frequency features by using a plain rectangular filter. They used speech spectrograms (images) to train their model and they achieved state-of-the-art results.

Text is another modality that attracted the interest of the research community and it has been widely used in the task of emotion detection and recognition. Batbaatar et al. in [2] developed a novel neural network architecture trying to recognise emotions from text by capturing the semantic/syntactic and emotional relationship between words. For capturing the semantic and contextual information they used a bidirectional Long Short Term Memory (BiLSTM), while a convolutional neural network (CNN) was utilized to extract the emotional features and understand the emotional relationship between the words.

Chiorini et al. in [4] investigated Bidirectional Encoder Representations from Transformers (BERT). They created a Bert-based (both uncased and cased version) architecture for classifying emotions based on Twitter data and they fine-tuned it, by adding a softmax layer on top of BERT, so that to classify the different categories of emotions.

3.2 Multi-modal Emotion Recognition

Multi-modal learning has proven to be more effective in emotion recognition than uni-modal approaches [16]. Features from different single modalities such as speech, text, facial expressions, and various combinations of those were used in many studies, trying to achieve systems with higher effectiveness on the emotion recognition task. Different approaches have been followed to address the multi-modal emotion recognition challenge.

3.2.1 Deep Neural Networks

Artificial neural networks such as deep neural networks have been utilized to recognize emotions and achieve state-of-the-art performance.

Soujanya Poria et al. in [23] proposed an LSTM-based recurrent neural network that extracts contextual features from the utterances of a video. First of all, they used audio, textual, and visual feature extraction methods to get context-independent uni-modal features. Then, they fed these features into an LSTM model allowing successive utterances in a video to provide important contextual information. The proposed method outperformed the existing state-of-the-art models showing significant improvement of the order of 10 %.

N. Majumder et al. in [18] followed a comprehensive fusion strategy in their emotion recognition research. At first, they extracted utterance-level features using three different modalities, text, video, and audio. Next, a feature fusion was applied for each bimodal combination, and only then were all three modalities fused. Recurrent Neural Networks (RNN) were also used to extract context-aware utterance features and pass them between the fully connected layers of their model. The final model outperformed the state-of-the-art models by 2.4%. This approach constitutes the baseline method of our project. However, the architecture of our multi-modal emotion recognition systems consists of state-of-the-art models, such as BERT that perceive the text data's contextual information and different features extracted to perceive important information to better distinguish the different categories of emotions. In addition, transformer attention mechanisms are also included in the TMERT model to better classify the four different human emotion categories.

Trisha Mittal et al. in [20] presented a multi-modal emotion recognition model, that can effectively sensor the noise in any of the different modalities such as facial expressions, speech, and text, that were used in their research. A data-driven multiplicative fusion method with LSTMs was implemented, resulting in learning their model to decide on a per-sample basis which modality to emphasize for making a prediction. In addition, they introduced a check step that follows the Canonical Correlation Analysis method, to distinguish efficient and inefficient modalities. Their model achieved one of the best results compared to other existing models.

Zexu Pan et al. in [22] studied the multi-modal attention network (MMAN), which is a hybrid fusion method for a speech emotion recognition task using both textual and visual cues. A cLSTM multi-modal attention mechanism was used for an early fusion of speech, visual and text features and three uni-modal cLSTM networks, one for each modality used for fusing the features in the final stage. The proposed hybrid approach managed to achieve state-of-the-art performance for the task of emotion recognition.

Shamane Siriwardhana et al. in [26] used two pretrained "BERT-like" architectures. In their research speech and text data, which were represented by two self-supervised learning algorithms, were fine-tuned for the task of multi-modal emotion recognition. After multiple experiments, they demonstrated that a simple fusion mechanism leads to an overall simpler structure and improves complex fusion mechanisms.

Krishna et al in [14] introduced a new approach for the emotion recognition task, that combines cross-modal attention with a convolutional neural network based on raw waveforms. One-dimensional convolutional models were used to process raw audio, while attention processes for audio and text features were applied to acquire enhanced emotion classification results.

Although, text, audio, and image data were widely analyzed for the task of emotion recognition, other modalities, such as body physiological signals were also used for emotion classification tasks. Yongrui Huang et al. in [10] proposed two decision-level fusion methods using a sum rule or a production rule for both brain and peripheral signals. Neural network classifiers and two support vector machines (SVM) were used for detecting emotions through an electroencephalogram (EEG) and facial expressions respectively, reaching an accuracy score up to 92,5 %.

Juan Antonio Dominguez et al. in [5] tried to recognize the emotions of 37 volunteers by recording two biosignals, heart rates, and galvanic skin response, while they were watching video clips. These signals were analyzed in both the time and frequency domain to extract a set of features. Multiple techniques for feature selection and classification were applied and they managed to achieve high accuracy (up to 100%) by using a support vector machine as a classification model and a random forest for recursive feature reduction.

The novelty of our proposed systems, TMER and TMERT can be expressed in both the models used in the architecture of our multi-modal emotion recognition systems and the different features extracted to perceive important information to better distinguish the different categories of emotions. Specifically, state-of-the-art models, such as BERT are used to get the contextual information of the text data, CNNs are used to classify the emotions using different kinds of features, while transformers are also included in TMERT, to perceive semantic information from each modality, achieving an emotion recognition system with even better performance compared to TMER.

3.3 Emotion Recognition in conversation

3.3.1 Deep Neural Networks

Navonil Majumder et al. in [19] presented a new method for the task of emotion recognition during a conversation. It is based on recurrent neural networks and takes into account each person's state individually using three gated recurrent units. By using this information their model managed to recognize different emotions and it outperformed other models that used textual and/or multi-modal features.

Jingye Li et al. in [15] presented a multi-task learning network for Conversational emotion recognition (CER) with the assistance of speaker identification. Their goal was to collect improved information related to the speaker, so to achieve that they combine the information exploited from a BERT-based model with two hierarchical bidirectional gated recurrent (Bi-GRU) neural networks. Regarding their results, they managed to achieve the best performance compared with the previous state-of-the-art models in the task of emotion recognition in a conversation.

3.3.2 Graph Neural Networks

Dong Zhang et al. in [30] built a conversational graph-based convolutional neural network to model both speaker-sensitive and context-sensitive dependence not only in traditional two-speaker but also in multi-speaker conversations. In their proposed model, nodes represent the utterances and the speakers while edges represent the context-sensitive dependencies and the speaker-sensitive dependencies, managing to surpass several state-of-the-art models.

Changzeng Fu et al. in [7] applied augmentation techniques in audio samples and build an emotion-oriented encoder-decoder in order to enhance the performance of their model. In addition, they introduced a graph attention network-based decision-level fusion for their final multi-modal emotion recognition model.

Deepanway Ghosal et al. in [9] presented a graph neural network, which uses the interlocutors' self- and inter-speaker dependency for contextual modeling, managing to handle important issues, such as long-term contextual information propagation.

Abhinav Joshi et al. in [11] proposed the COntextualized Graph Neural Network based Multi-modal Emotion recognitioN (COGMEN) model. It is a Graph Neural Network (GNN) based architecture, that holds local information, such as inter-dependency between speakers and contextual information. COGMEN reaches state-of-the-art results, showing the importance of modeling both local and contextual information.

3.3.3 Commonsense knowledge

Contextual and commonsense knowledge, such as mental states, and causal relations is key to detecting the real emotion of people in conversations. Peixiang Zhong et al. in [31] proposed a knowledge-enriched transformer (KET) for identifying emotions. They used hierarchical self-attention to interpret contextual utterances, while a dynamic context-aware affective graph attention mechanism obtains the external commonsense knowledge. Their experimental results verified the benefits of exploiting the knowledge from both context and commonsense in the emotion recognition task.

Deepanway Ghosal et al. in [8] also presented a framework (COSMIC), which covers multiple elements of commonsense knowledge, including mental states, actions, events, and cause-effect relations. Based on this knowledge they tried to interpret the interactions between participants in conversations, managing to achieve state-of-the-art results for emotion recognition.

4 Fundamentals

This section includes general information about the models that are used in this project to process and analyze the text, audio, and visual data of the IEMOCAP dataset. The BERT model, which constitutes the base of the text emotion recognition system in this project, and the CNNs, which are used for detecting human emotions using audio and visual features, are analyzed in detail. In addition, definitions of the extracted audio features and the evaluation metrics of the proposed models are presented.

4.1 BERT model

BERT [4] constitutes a model that generates language embeddings. These embeddings can be used to tackle a lot of different NLP problems, such as Question Answering, Language Translation, Text Summarization, and Emotion Recognition. BERT's architecture consists of a bidirectional transformer, that processes the entire input sentence at once, instead of processing each word separately. Thus, BERT can perceive contextual information of the input text data, improving the performance of the related NLP task. In the case of emotion recognition, obtaining the context behind each sentence can give useful information about the hidden emotion, that is why BERT is used in this project's experiments.

When using the BERT model, a specific set of rules is used to represent the input text. Each input is a combination of position embeddings, segment embeddings, and token embeddings as Figure 1 shows.

Positional embeddings are used from BERT so that the position of each word in a sentence is expressed. Since the position of the words in a sequence determines their meaning, these embeddings are used to capture that useful information.

Sentence pairs can be used as input to BERT. Thus, BERT needs to learn a unique embedding for each sentence. These embeddings are called segment embeddings and distinguish the two

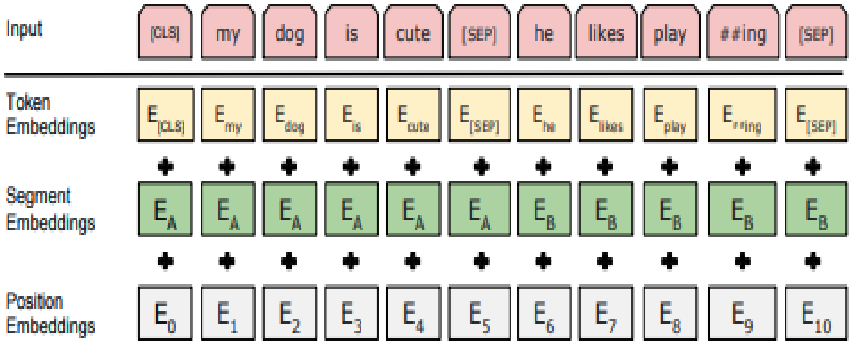

Figure 1: Representation of BERT inputs as depicted in [6]. $E_{token} \in \mathrm{N}$, $\mathrm{E}_{segment} \in \{0, 1\}$, $E_{position} \in \{1, 2, ..., maximum\ sequence\ length\}$. The input embeddings are the sum of the position embeddings, the segmentation embeddings and the token embeddings

sentences between each other. In Figure 1, all the tokens which are marked as E_A belong to sentence A, while all the tokens which are marked with E_B belong to sentence B.
Finally, the token embeddings for each token constitute the embeddings, that were learned from the WordPiece token vocabulary.
The input representation of each token corresponds to the sum of the respective token, segment, and position embeddings. The embeddings are added up instead of concatenated. By adding the embeddings, the information that is needed to accomplish the tasks of BERT's pre-training is obtained. During the pre-training process of the BERT model, two NLP tasks are implemented, the Masked Language Modeling, and the Next Sentence Prediction.
During the Masked Language Modeling, 15% of the words are randomly masked, and the model focuses on the context of words that exist on both the right and the left side of the masked word. During this process, BERT model acquires useful information regarding the relation of the words and their meaning. Since, the input text files are used directly for masking the words, while the true tokenized representations of the text are known, Masked Language Modelling is considered a self-supervised task where no labeled data have to be given. On the other hand, during the Next Sentence Prediction task, BERT learns the relationships between sentences.

BERT's pre-training constitutes a computationally expensive task. However, when the pre-training task is finished, then the model can be used and applied to various tasks, such as human emotion recognition. To use such a pre-trained model for a specific task, there is often the need to fine-tune the model, which constitutes a much easier and less computationally expensive task. BERT usually results in state-of-the-art performance, by making use of the contextual information of the words of the input text files.

4.2 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is one of the common deep learning neural networks and it is usually used for processing data with a grid pattern, such as images, but also for signal analysis tasks, such as Speech Emotion Recognition, which deals with audio signal analysis.

A typical CNN infrastructure is composed of three types of layers, namely convolution layers, pooling layers, and fully connected (dense) layers. Both convolution and pooling layers, extract features, whereas the dense layer, often use the extracted features to determine the final output, like a classification. The convolution layer, which often combines linear and nonlinear processes including convolution operations and activation functions, is the core element of the CNN architecture.

Convolution is a linear operation, typically used for feature extraction. The kernel, which is a small array of numbers, is applied across the tensor, which is an array of numbers. To obtain the output value, which is called a feature map, there is a calculation for every location of the tensor. This calculation is the sum of an element-wise product between the input tensor and every element of the kernel. By applying this process to multiple kernels, an arbitrary number of feature maps is formed, representing the input tensors' characteristics. Finally, a nonlinear activation function is used, where the outputs of the convolution operation are passed through it.

A CNN constitutes several stacked building blocks of convolution, pooling, and dense layers. The output of each layer is fed into the next layer, making the extracted features more and more complex. For the evaluation of the model's performance under specific parameters (kernels), and weights, the loss function is calculated through forward propagation on the training dataset. The kernels and weights are updated through a backpropagation algorithm with gradient descent, according to the loss function. Figure 2 shows an example of a CNN architecture.

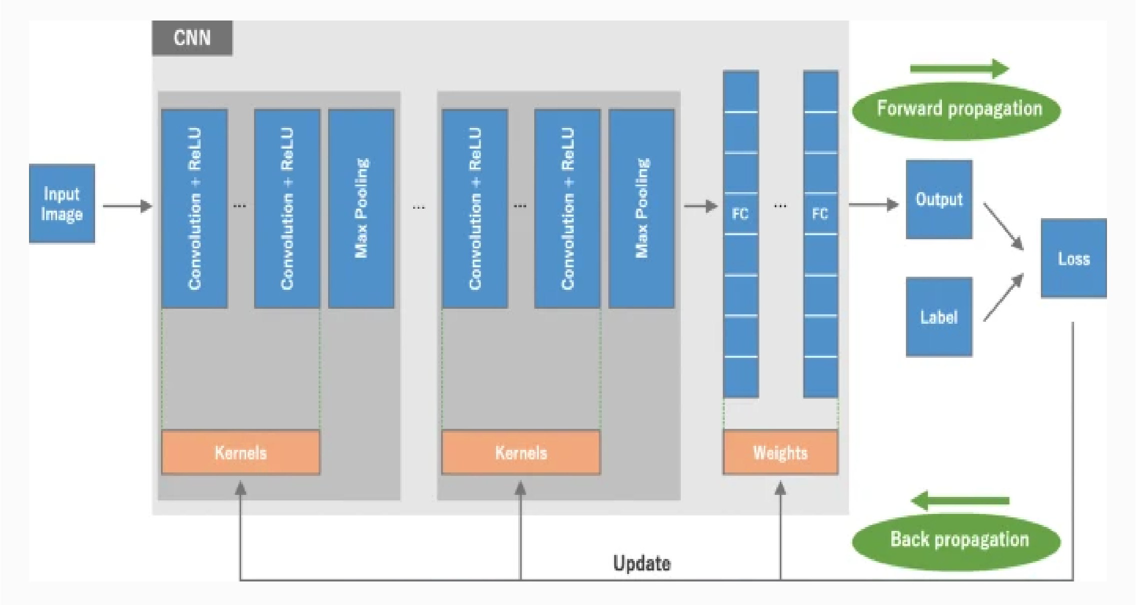

Figure 2: Example for image classification of a convolutional neural network (CNN) architecture as illustrated in [29]

4.3 MFCC, Log mel Spectogram and Chromagram

In this subsection definitions of the extracted frequency-domain audio features, that are used in the EmoA, TMER, and TMERT models are presented.

- **Mel-Frequency Cepstrum Coeffiecient**
  Mel-frequency Cepstrum Coefficient of a sound constitutes its short-term power spectrum representation. To obtain this kind of representation, the audio signal is transformed to mimic the human cochlea. Additionally, information regarding the rate changes in different spectrum bands is included in the MFCC. The Mel scale is really important because, in comparison with the linear scales, it better approximates human-based perception of sound.

- **Log mel spectrogram**
  The second feature that is extracted is the Log mel spectrogram. This is a representation of the frequency (time vs. log-mel frequency). The logarithmic form of mel-spectogram is useful for human emotion detection because human perceives the sound on a logarithmic scale.

- **Chromagram**
  The final feature that is extracted is the Chromagram, an audio's spectral energy representation at each of the 12 pitch classes.

4.4 Evaluation metrics

To evaluate the performance of the EmoT, EmoA, EmoFH, TMER, and TMERT, the accuracy and the F1-score are calculated.

Generally, accuracy is a metric for evaluating classification models and it is actually the fraction of predictions, each model got right. Here, is the formula to calculate the accuracy of a model:

$$Accuracy = \frac{Number\_of\_correct\_predictions}{Total\_number\_of\_predictions} \tag{1}$$

The F1-score is the second evaluation metric of this project's experiments. To calculate the F1-score, both precision and recall should be examined.

At this point, it is important to introduce the Confusion Matrix, which is useful for measuring precision and recall.

Confusion Matrix measures the performance of an ML classification task where the output classes can be two or more. It is a table that illustrates different combinations of both the predicted and the actual values. Table 1 shows a Confusion Matrix, where True Positive (TP) are the values that are correctly predicted as positive, True Negative (TN) are the values that are correctly predicted as negative, False Positive (FP) are the values that are incorrectly predicted as positive, and finally False Negative (FN) are the values that are incorrectly predicted as negative.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 1: Confusion Matrix

Definitions of Precision, Recall, and F1-score are analyzed below:

Precision is the fraction of the number of True Positives divided by the total number of True Positives and False Positives. It shows the percentage of the successful positive predictions among the examples that the model predicted as positive. The formula to calculate the precision is:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is the fraction of number of True Positives divided by the total number of True Positives and False Negatives. It actually shows the percentage of all the actual positive examples, the model correctly predicted to be positive. The formula to calculate the recall is:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score represents the harmonic mean of precision and recall for a more balanced assessment of the models. The formula to calculate the F1 score is:

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

5 Baseline method

N. Majumder et al. in [18] worked on a multi-modal emotion recognition task, presenting a novel strategy, the Context-aware hierarchical fusion (Figure 3), where features are fused in a hierarchical fashion. In [18], N. Majumder et al. are focused on classifying four different categories of human emotion using the IEMOCAP dataset, and at the time of our project's experiments, their method brought state-of-the-art results in terms of accuracy in the multi-modal emotion recognition task using three different modalities. These are the reasons why the performance of each proposed model in this paper, is evaluated against their research.
In their method, they first extracted utterance level features coming from three different modalities: text, audio, and video.
In the case of textual data, transcripts of videos were used. Each text utterance was represented as an array of 300-dimensional word2vec vectors that had been pre-trained, and a deep Convolutional Neural Network (CNN) [12] was applied for each utterance.

For the audio features, they used openSMILE to extract a number of Low Level Descriptors (LLD), such as voice intensity and pitch, while they also extracted a variety of statistical functionals of them, such as standard deviation, kurtosis, amplitude mean, etc.
Visual features were also extracted from each video frame and 3D-CNNs were used to model temporal features across frames.

N. Majumder et al. also considered the semantic dependence between the utterances in the videos. GRU models were added to their system, to perceive the context of the utterances. The extracted features of each modality were fed to GRUs, resulting in context-aware unimodal features. Next, they worked on the fusion of the three different modalities by combining each bimodal combination of the unimodal features, utilizing fully-connected layers. Here, the fused bimodal features were fed to GRUs again to get the semantic information of them. Finally, all extracted bimodal features were combined into a trimodal vector, resulting in a strategy that outperformed the state of the art models (2018) by 2.4 %.

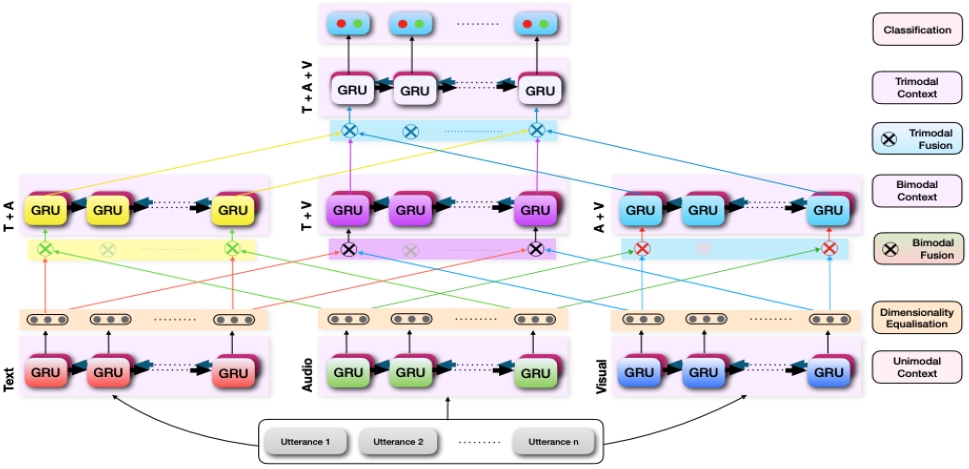

Figure 3: Context-aware hierarchical fusion architecture as depicted in our baseline method [18]

The proposed models of this paper follow a different approach. BERT model is used to get the contextual information of the text data by using sentence embeddings, while 12 MFC coefficients, 128 log Mel spectrogram bands, and 12 chromagram pitch classes consist the audio features. Motion-capture data are also used as visual features. In generall, TMER model employs three CNNs for the training of the text, audio and visual inputs independently. The last convolutional layers for each modality, are concatenated to form the final vector representation. This vector is then used as an input to an extra CNN with two convolutional layers, where the emotions are finally classified in four different categories.
On the other hand, TMERT model is a Transformer-based model, whose architecture is based on attention mechanisms. It uses text data, audio data, and visual data as an input to the transformer separately and only then they are fused and trained together using a CNN model. A detailed description of the multi-modal models follows in Section 6.

6 Proposed methods: TMER & TMERT

This section includes detailed information about the uni-modal emotion recognition models, that were investigated and created from scratch to recognize four different categories of human emotions, such as joy, sadness, anger, and neutrality. Furthermore, it is described how these uni-modal models are used to constitute the architecture for the multi-modal fusion method and the novel two multi-modal emotion recognition systems, TMER and TMERT.

6.1 Uni-modal emotion recognition models

In this subsection the uni-modal emotion recognition models, EmoT, EmoA, and EmoFH are introduced.

6.1.1 EmoT: Emotion Recognition Model using Text

Our EmoT model uses BERT to handle the emotion recognition task by processing text files from the IEMOCAP dataset. BERT [4] constitutes a model that generates language embeddings. These embeddings can be used to tackle a lot of different NLP problems, such as Question Answering, Language Translation, Text Summarization, and Emotion Recognition. BERT's architecture consists of a bidirectional transformer, that processes the entire input sentence at once, instead of processing each word separately.
For the EmoT architecture, a rather simple construction is used. The input data, which are transcripts of the dialogues of the actors in IEMOCAP dataset, are transformed into a special format (tokens), and then each token can be used from the pre-trained BERT model, to obtain the corresponding embeddings. To perform emotion recognition, fine-tuning the pre-trained BERT is needed, that is why some network is required to be laid on top of it to classify its output into four different categories of emotion. To do that, we added a BERT model from Hugging Face: `BertForSequenceClassification`, which uses a Linear Classifier to distinguish the emotion categories. As the number of the emotions is four, the number of output labels is set to four as well. By using this network, IEMOCAP dataset sentences can be provided to it and a number for each class will be outputted. High values of this number, correspond to higher possibilities of the model to have classified the corresponding label correctly. Thus, the predicted output of each sentence can be found by taking the argmax of this list. Position 0, corresponds to the "angry" label, 1 is "neutral", 2 is "happy", and 3 is "sad".

As mentioned before, the input data of EmoT model are converted into a special format. Specifically, `BertTokenizer` is used from the Hugging Face library[1]. This performs all the necessary preprocessing on the dataset so it can be processed by BERT. It actually, splits the raw text into tokens, which are numeric representation of words, and it is also used to encode the data.
Additionally, the tokenizer also lowercases each sentence of the dataset. This is necessary as the pre-trained model called "bert-base-uncased" [1] from Hugging Face is used. Lowercased English text from unpublished books and English Wikipedia was used to train this model, thus the IEMOCAP transcripts need to be lowercased as well.

Furthermore, an optimizer and a scheduler for the input data are created. AdamW optimizer [17] is the optimizer used. This optimizer works well with high volumes of data, and it is a stochastic gradient descent method that manages to update the weights of the network efficiently.
Additionally, the
`get_linear_schedule_with_warmup` scheduler from Hugging Face is used, to minimize the instability at the beginning of the training process. In this manner, the learning rate can be raised linearly from lower rates to a constant value.
Figure 4 illustrates the overall architecture of the EmoT model.

6.1.2 EmoA: Emotion Recognition model using Audio

The architecture of EmoA, the model that was created for recognizing the human emotion by using audio files from IEMOCAP dataset, is simple. The audio data are fed into a 1D

[1] https://huggingface.co/transformers/model$_{d}oc/bert.html$

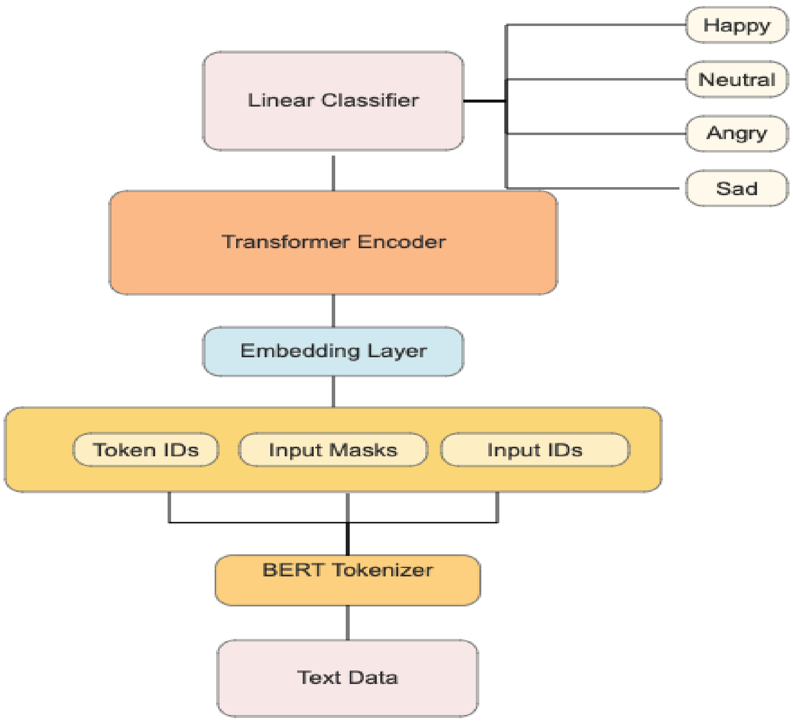

Figure 4: The EmoT model architecture

Convolutional Neural Network (CNN), which is composed of convolution layers and pooling layers, that extract features, and fully connected layers, that use the extracted features to determine the final output, namely the classification of the audio files into four different categories. Additionally, batch normalization and dropout layers are also included in the CNN to improve the learning process of the model.

The number of the hidden layers and the number of neurons added not only in EmoA but also to all the other models that are created in this research, were decided through several preliminary experiments. For the EmoA model, 3 Convolutional layers were used with 128, 64, and 32 neurons respectively. Three hidden layers are capable to learn complex representations, while the selected number of neurons manage to overcome the underfitting or the overfitting and the model performs the best.

The activation of the input layers is done using the Rectified Linear Unit (ReLU) function. When dealing with convolutional neural networks, the ReLU function is the default activation function. After experimenting with several others, such as the sigmoid function and the tanh function the most successful model in our preliminary experiments resulted when using the ReLU function. The softmax function is used for the activation of the output layer. This function assigns a probability to the classes of the network, in this project's case, the 4 emotion classes. The label that the model finally predicts is the one that is assigned the highest probability. The Adam optimizer with a learning rate of 0.0001 is used and the loss function is the categorical crossentropy, since this is a multi-class model. In addition, between the convolution layers of the EmoA model, batch normalization and dropout layers are added.

In addition, between the convolution layers of the EmoA model, batch normalization and dropout layers are added. Batch normalization is used so that the output of the previous layers is normalized. Dropouts are added to randomly switching some percentage of neurons

on the network. By switching off some of the neurons of the network, the incoming and outgoing connection to those neurons is also switched off. By using both batch normalization and dropout layers, the learning becomes more efficient and the model's overfitting is also diminishing.
The pooling layer provides a typical dimension reduction operation. It actually, reduces the number of the learnable parameters and the computation amount in the network. EmoA model uses a MaxPolling operation, which calculates the maximum value in each patch of each feature map.

The output of EmoA's final convolution is then flattened, i.e., transformed into a one-dimensional array of numbers, and connected to a number of dense layers or fully connected layers, in which each input and each output are connected by a learnable weight. The number of the output nodes of the final fully connected layer is usually the same as the number of classes, that is why four dense layers were added in EmoA, just like the number of the emotion classes. Figure 5 illustrates the overall architecture of the EmoA model.

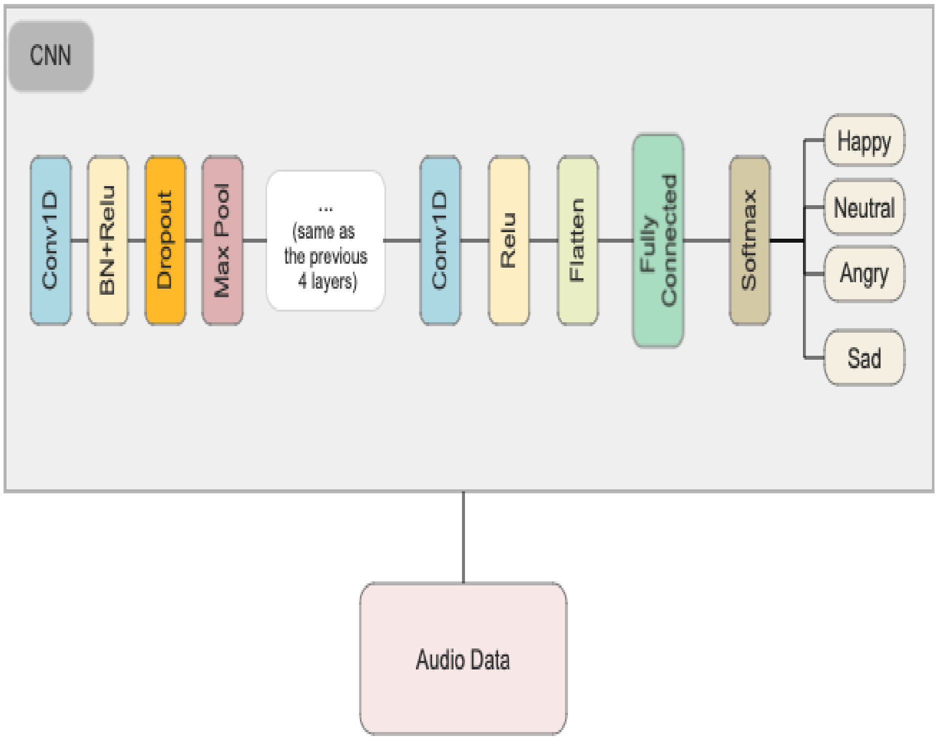

Figure 5: The EmoA model architecture

6.1.3 EmoFH: Emotion Recognition model using Facial expressions and Hand movements

EmoFH model is created to classify different human emotions using facial expressions and hand movements. The same model architecture with the EmoA is used to build EmoFH. The visual data are fed into a 1D Convolutional Neural Network (CNN), which is composed of convolution layers, and pooling layers, that extract features, and fully connected layers, that use the extracted features to determine the final output, namely the classification of the audio

files into four different categories. Additionally, batch normalization and dropout layers are also included in the CNN to improve the learning process of the model.
Specifically, 3 Convolutional layers are added in EmoFH model, while the activation of the input layers is done using the Rectified Linear Unit (ReLU) function. The output layer of EmoFH is activated with the softmax function, while Adam is the optimizer that is used with a learning rate of 0.0001 and the loss function is the categorical cross-entropy since this is a multi-class model.

Also, between the convolution layers of the EmoFH model, batch normalization and dropout layers were added. Similar to EmoA, EmoFH uses a MaxPooling operation, which calculates the maximum value in each patch of each feature map. Next, the output of the final convolution of EmoFH is flattened and connected to several fully connected layers, also known as dense layers, in which every input is connected to every output by a learnable weight. The number of the output nodes of the final fully connected layer is usually the same as the number of classes. In this project the number of classes, we want to distinguish the data is four, that is why four dense layers were added in EmoFH. Figure 6 illustrates the overall architecture of the EmoFH model.

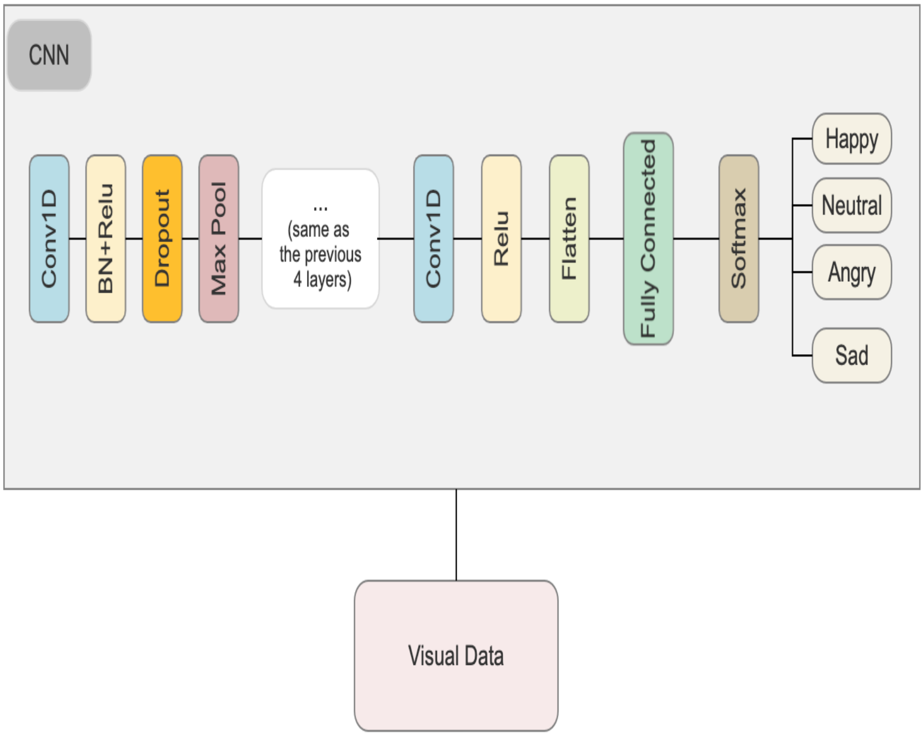

Figure 6: The EmoFH model architecture

6.2 TMER: Tri-Modal Emotion Recognition

A novel architecture called Tri-Modal Emotion Recognition (TMER), which combines features from three different modalities: text, audio, and video (facial expressions and hand movements), is presented and compared with the performance of the multi-modal emotion recognition systems of the baseline [18].

The architecture of TMER, is based on four Convolutional Neural Networks. Text, audio, and visual data are used as an input to three Convolutional Neural Networks separately. Then, the output features of the three CNNs, are combined and used as an input to the final CNN network, which classifies the data into four different emotion categories.

Sentence embeddings, which contain sequential textual information, MFCCs, Log mel spectrograms, chromagrams, which contain audio information, and motion capture data, which contains visual information, are all the extracted features that are fused and used to train the TMER. The audio and the visual features are the same as those used in the EmoA and the EmoFH model respectively.
After several preliminary experiments, the most successful model resulted when the preprocessed input data of each modality, are fed into a CNN separately, whose layers and parameters are also slightly different compared with the architecture of EmoA or EmoFH.

As mentioned before, the TMER model employs three CNNs for the training of the text, audio and visual inputs, independently. For each input, two convolutional layers with 128 and 64 nodes respectively are used. These nodes after the calculation of the weighted sum of the inputs, return activation maps, that identify important features. A built-in regularizer, L1 is also applied at the first convolutional layer. During the optimization, regularizers are used to apply penalties to the layer parameters, and finally, these penalties are summed into the loss function that the network optimizes.

Here is the computation of L1 regularization, where the learning rate, l1=0.0001 :

$$loss = l1 * reduce\_sum(abs(x)) \tag{5}$$

The activation of the input layers is done using the Rectified Linear Unit (ReLU) function. In addition, between the convolution layers, batch normalization and dropout layers were added. In that way, the learning becomes more efficient and the model's overfitting is also eliminated. In addition, TMER model uses a MaxPooling operation, which returns the most relevant features from the layer in the activation map.
Next, the features of each modality are fused and trained together. In this project, the feature-level fusion technique is used. Feature-level fusion is performed by combining the features of text, audio and visual modalities into a single feature vector. Here, the last convolutional layers for each modality, are concatenated to form the final vector representation, and this vector is then passed through two extra convolutional layers with 32 and 8 nodes respectively. Dropout layers are added between the convolution layers to avoid overfitting. The final output of TMER is flattened and then connected to four fully connected layers, as the number of the classes. The output layer of the TMER is activated with the softmax function, while Adam is the optimizer that is used with a learning rate of 0.0001. As a loss function, the categorical cross-entropy is used, since this is a multi-class model. Figure 7 illustrates the overall architecture of the TMER model.

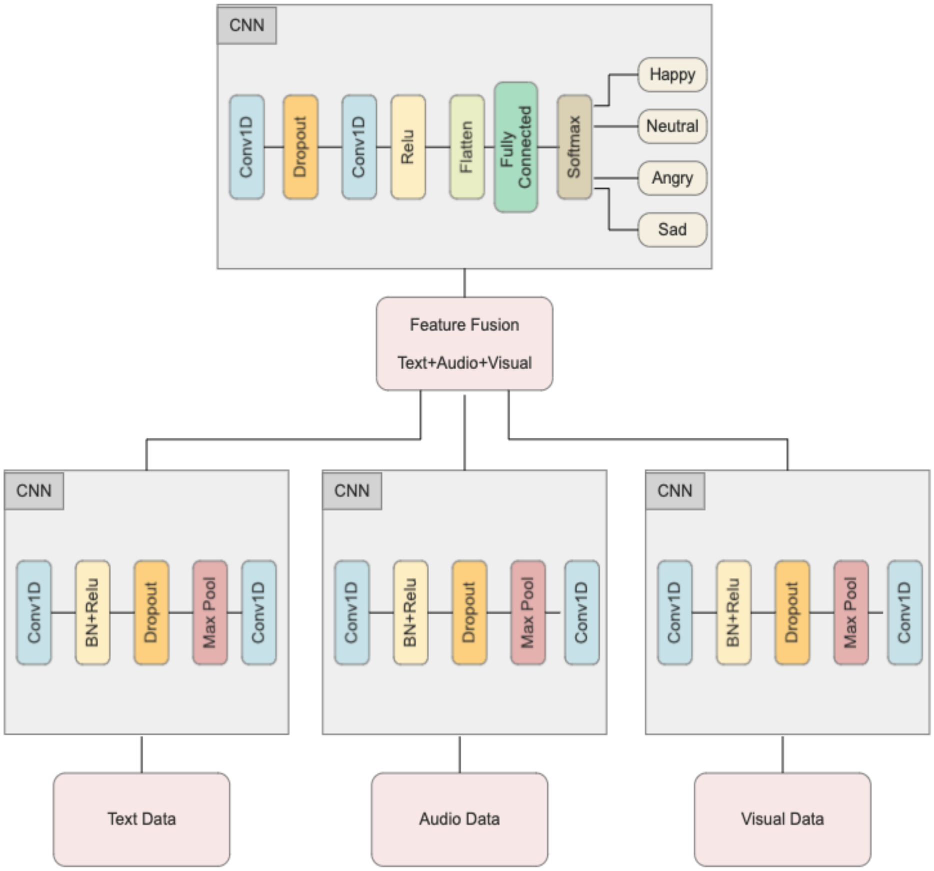

Figure 7: The TMER model architecture

6.3 TMERT: Tri-Modal Emotion Recogninion Transformer

In this section we propose a novel Tri-Modal Emotion Recognition Transformer, a multi-modal emotion recognition model using a transformer attention mechanism. In literature, many of the best performing models in human uni-modal and multi-modal emotion recognition tasks, make use of an encoder and a decoder through an attention mechanism. TMERT constitutes a transformer-based model, whose architecture is based on an attention mechanism as described by Vaswani et al. in [28]. Vaswani et al. used an encoder to map an input sequence of symbol representations $(x_1, ..., x_n)$ to a sequence of continuous representations $y = (y_1, ..., y_n)$. Then, an output sequence of symbols, $z = (z_1, ..., z_n)$ is generated by the encoder. The Transformer's architecture contains stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, as shown in the left and right part of Figure 8, respectively.

TMERT uses similar Transformer architecture as in [28], however, the TMERT model is applied to numerical data, since this is a time series classification task. Additionally, the TMERT architecture does not include the decoder part of the transformer. Overall, the input data are preprocessed and they are fed to the transformer separately to capture the contextual relationships of the data, through the multi-head attention layer. The output of the transformer is a set of encoded vectors for every input data. Then, the output data of each modality are combined and trained together for the task of emotion recognition.

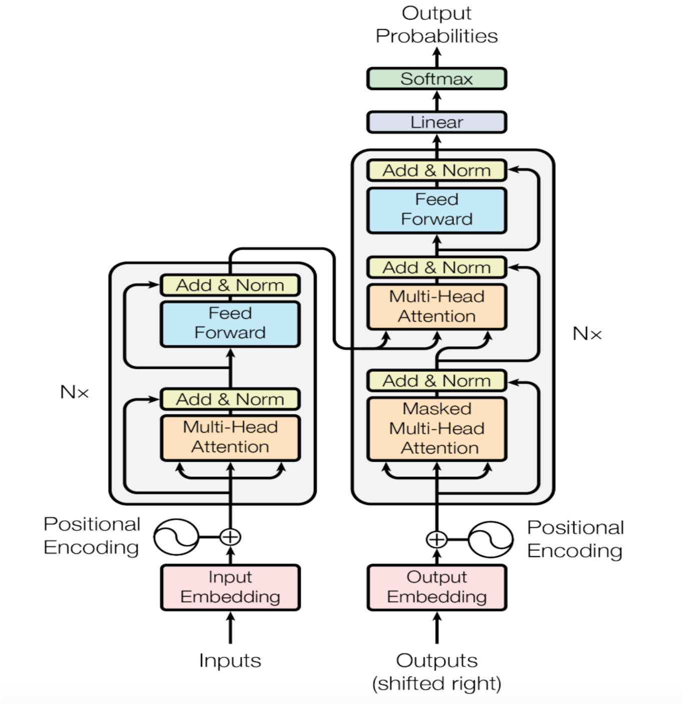

Figure 8: The Transformer - model architecture as depicted in [28]

TMERT processes a tensor of shape (batch size, sequence length, features), where the 'sequence length' corresponds to the number of time steps, while 'features' corresponds to each input time-series.

Residual connections, layer normalization, and dropout layers are included in the TMERT architecture and finally, the resulting layer is stacked multiple times. Specifically, a stack of four identical layers composes the encoder, while each layer contains two sublayers, a multi-head self-attention mechanism, and a position-wise fully connected feed-forward network. Residual connections followed by layer normalization are employed around each of the two sub-layer resulting in the output of the encoder. Figure 9 shows the architecture of the encoder of the TMERT model.

Next, the final Multi-Layer Perceptron classification head is added. Apart from a stack of Dense layers, the output tensor of the encoder needs to be reduced down to a vector of features for each data point in the current batch, so a GlobalAveragePooling1D layer is used for this purpose. Then, the features of each modality are combined into a single feature vector and they trained together. The activation of the final fused output of TMERT model is done using the Rectified Linear Unit (ReLU) function and batch normalization layers are also added. Next, the output is flattened and then connected to four fully connected layers, as the number of the classes. The output layer of the TMERT is activated with the softmax function, while Adam is the optimizer that is used with a learning rate of 0.0001. As a loss function, the categorical cross-entropy is used, since this is a multi-class model. Figure 10 illustrates the overall architecture of the TMERT model.

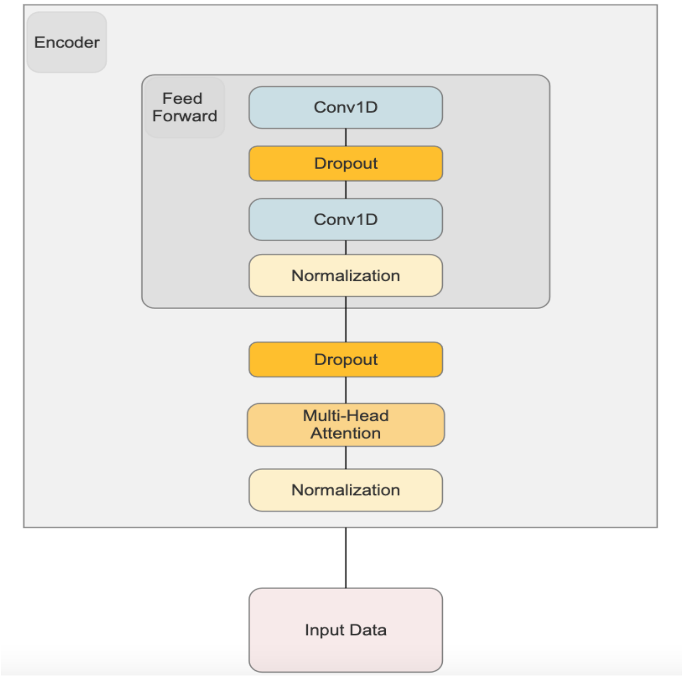

Figure 9: The TMERT's encoder architecture

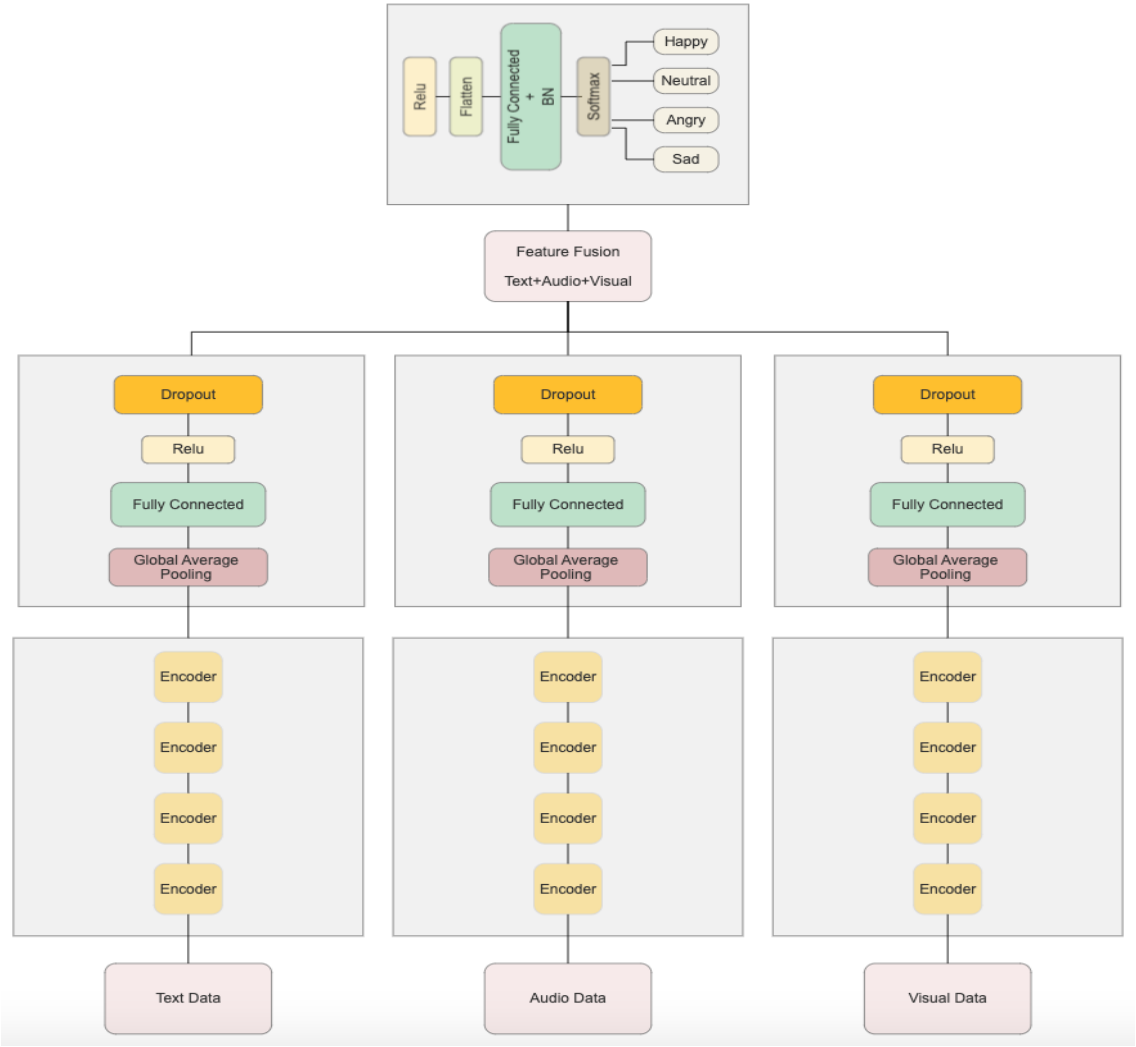

Figure 10: The TMERT model architecture

7 Dataset

7.0.1 IEMOCAP

Human emotion recognition can be a challenging task. During human interactions, there are interesting paralinguistic messages expressed through both speech and gestures. During a natural human conversation, the facial expressions, hand, and head movements, in conjunction with the energy and the tone of the speech, are all combined, indicating many different human emotions. These communicative characteristics need to be efficiently taken into account to develop and implement robust human emotion recognition models.
"Interactive emotional dyadic motion capture database" (IEMOCAP), collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC), is a widely used dataset in multi-modal emotion recognition tasks. This study uses the IEMOCAP dataset to run the experiments and compare the uni-modal models (EmoT, EmoA, EmoFH) and the multi-modal models (TMER and TMERT) with the baseline method.

IEMOCAP consists of approximately twelve hours of recordings of ten actors in dyadic sessions. In total there are 5 sessions, each containing about 30 dialogues (150 dialogues in total). Next, each dialogue is split into single sentences (10039 sentences in total). During the recordings of the sessions, markers on the face, head, and hands of one of the actors were placed, providing this way detailed information, regarding their facial expressions and hand movements. Certain types of emotions, such as happiness, anger, sadness, and neutrality, were elicited by the actors, who improvised hypothetical scenarios and performed selected emotional scripts.
Table 2 illustrates statistics regarding the total number of dialogues and sentences in different sessions.

	Session 1	Session 2	Session 3	Session 4	Session 5
# of Dialogues	28	30	32	30	31
# of Sentences	1819	1811	2136	2103	2170

Table 2: Total number of dialogues and sentences across all sessions in the IEMOCAP dataset.

Each sentence has been evaluated categorically by at least three annotators, among the following ten emotions: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other. In this study, only emotions such as anger, happiness, sadness, and neutral were used in the experiments, in order to be comparable with the baseline method. Table 3 shows the total number of samples per each emotion.

	# of Samples
angry	1090
neutral	1704
happy	586
sad	1077

Table 3: Number of samples of the main four emotions in IEMOCAP dataset, that are used in the experiments.

7.0.2 Train Test Split

4457 sentences (5498 in baseline) are obtained, after the preprocessing of the IEMOCAP dataset. The reason for selecting the exact sentences is that their type of emotion belongs to categories, such as angry, neutral, happy, and sad, which is the same category of emotions used in the baseline method [18]. Thus, a fair comparison between the models of each paper is achieved. In addition, these sentences are chosen so that they contain not only their corresponding audio files and transcripts but also full data regarding the provided facial expressions and hand movements from the IEMOCAP dataset. To create a training and test set, this set of sentences is divided hierarchically as seen in Table 4. The test set constitutes a random 15% of the entire set. This value was chosen because it allowed for a sufficient number of samples for each label in the test set.

Label	Data type	Sentences
angry	train	927
	test	163
neutral	train	1448
	test	256
happy	train	498
	test	88
sad	train	915
	test	162

Table 4: Label distribution after splitting the dataset into training (train) and test set.

7.1 Uni-modal Feature Extraction

In this subsection the extracted features of each modality, that are used in the proposed uni-modal and the multi-modal emotion recognition systems are introduced.

7.1.1 Textual Feature Extraction

In the case of the uni-modal emotion recognition system, language embeddings are obtained from the text input data. Raw sentences from the IEMOCAP dataset are tokenized, and lowercased and then each token is converted into an encoded form, where we obtain the input ids, attention masks, and token labels, that are used to train the EmoT model.

In the case of the multi-modal emotion recognition system, a different approach is followed, so that the multi-modal models achieve better performance. First of all, the text is pre-processed by converting all the letters to lowercase and replacing punctuation with space or multiple spaces with only one space. Next, sentence embeddings are obtained from the text data, and used to train the model. Sentence embeddings were extracted using Sentence-BERT (SBERT) model. SBERT was firstly introduced by Nils Reimers and Iryna Gurevych in [25] and can be expressed as BERT's model modification. Regarding the architecture of SBERT, there is a pooling operation, which is added to the output of BERT to obtain a sentence embedding with a fix size. SBERT uses siamese and triplet networks to derive sentence embeddings with useful semantic information. It manages to create better sentence embeddings than BERT because it takes into account the context of many previous steps and the dependencies between the words in the text data.

The procedure uses a SentenceTransformer model, `all-MiniLM-L12-v2`, which was trained on a large dataset with more than 1 billion training pairs, to map each sentence to embeddings. It maps the sentences to a 384-dimensional dense vector space and finally, the sentence embeddings are computed.

7.1.2 Audio Feature Extraction

Audio features can be distinguished into two main categories, namely time-domain features and frequency-domain features. Zero crossing rate, maximum amplitude, and energy of the signal constitute representative examples of the time-domain features. These kinds of features can be easily extracted and used in audio signal analysis. On the other hand, Mel-Frequency Cepstral Coefficients (MFCCs), chroma coefficients, spectral entropy, etc. belong to the frequency-domain features. These features can be obtained by the conversion of the time-based signal into the frequency domain, revealing deeper patterns, which can be extremely useful to identify the underlying emotion of an audio file.

After experimentation with frequency-domain features, it was realized, that the emotions of the different audio files of the dataset can be easily distinguished. The mean across each band over time of the MFCCs, Log Mel Spectograms, and Chromagrams, for audio files with the same content, is calculated and the different ranges of the obtained values approve that these features can be used to detect different human emotions. Thus, these are the features that are decided to be extracted for this research.

Overall 152 features were extracted, consisting of 12 MFC coefficients, 128 log Mel spectrogram bands, and 12 chromagram pitch classes. These features are used as audio data input in the EmoA, TMER, and TMERT models.

7.1.3 Visual Feature Extraction

Since emotion is related to landmarks and the head together with the hand movements, these are the features, that were considered and used for the visual emotion classification task. Different patterns can be identified in the correlation between head and hand movements and the emotion, that is hidden behind them. For example, when people feel neutral, the movement of landmarks is relatively small, while when people feel sad, they tend to lower their heads.

Motion-capture data were collected in IEMOCAP dataset and they were used in this project. During the recordings of the sessions, fifty-three markers were placed on the face of one of the actors per session, providing this way detailed information, regarding their facial expressions. Keeping the markers far from each other incremented the accuracy of the motion information. MPEG-4 standard's feature points were followed during the placement of most of the facial markers as illustrated in Figure 11. In addition, hand movement information is also obtained from the wristbands with two markers each, and the extra markers, that were placed in each hand.
The motion capture system samples at a rate of 120 frames per second. The subjects were asked to sit during the recording in order to prevent gestures outside the volume indicated by

the common field of view of the VICOM cameras. However, they were told to make as many natural gestures as they could without covering their faces with their hands.
A VICON motion capture system, which contains eight cameras is used for recording the trajectories of the markers' data, as shown in Figure 12. 120 frames per second, is the sample rate used from the motion capture system. Some details were also taken into account, to obtain qualitative data. Specifically, the actors remained seated during the whole recording, so that all of the gestures were inside the field of view of the cameras used. Also, the actors were asked to gesture naturally, avoiding covering their faces with their hands, and display neutral poses at the beginning of each session, intending to use this information to define neutral poses.

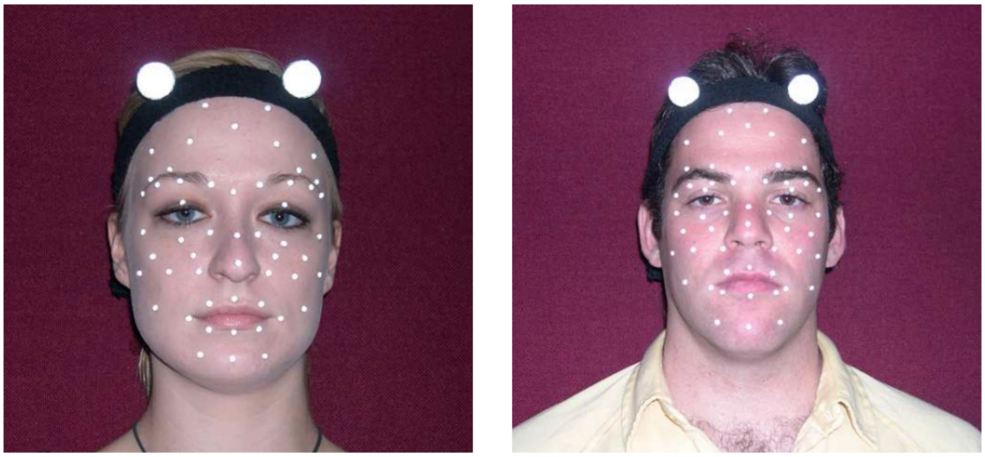

Figure 11: Markers placed on the face and headband of two of the actors, who participated in the recordings as depicted in [3]

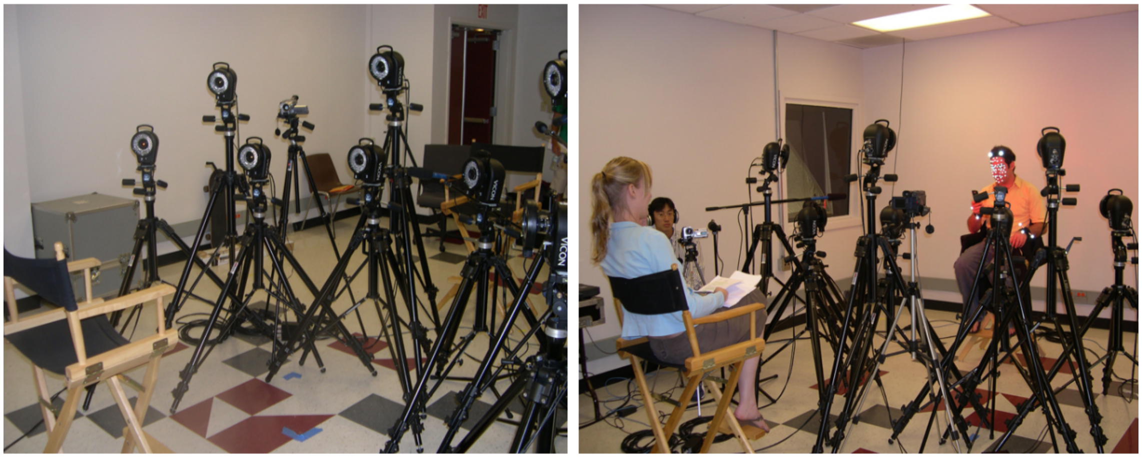

Figure 12: VICON motion capture system as illustrated in [3]

In IEMOCAP dataset the MoCap data exist in both dialog format and sentence format in the corresponding folders. Each sentence can be found in text files, in which each line is a frame number and each column is a marker coordinate. In this project, the sentence format is used, while for each sentence the mean of all the values of each marker is calculated and fed into the human emotion recognition systems.

7.2 Data normalization

Data normalization is the process of transforming data to appear similar across all records. By normalizing the data, the model training becomes less sensitive to the scale of features and it converges to better weights, having as a result higher accuracy.
After observing the values of the features used in this project, it is clear that many of the attributes are on a different scale, thus a normalization technique is required.
Min-Max is the normalization technique applied to the data before they are converted to vectors and fed into each model. It is a common and simple method, in which the data is scaled to a fixed range usually 0 to 1. The general formula for min-max normalization is given as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{6}$$

Data normalization is applied to the features used in the uni-modal and multi-modal emotion recognition systems. However, in the case of the multi-modal emotion recognition systems, Principal Component Analysis (PCA) is applied to the whole dataset after its normalization, so that the final feature vectors of each modality will have the same dimensions and will we suitable to be merged and fed into the multi-modal system.
PCA constitutes a dimensionality reduction method, in which a large set of variables is transformed into a smaller set, which still contains most of the large set information. In general, PCA consists of five steps, namely normalization, covariance matrix computation, computation of the eigenvectors and eigenvalues of the covariance matrix for the principal component identification, feature vector matrix creation and finally recasting of the data along the axes of the principal component.
The first step, normalization is already applied to the features as described at the beginning of this section. The reason why normalization is critical before the PCA is that variables with large ranges between the ranges of the initial variables dominate over those with small differences. So, normalization can enhance the performance of the model by preventing biased results.
Next, the covariance matrix is computed, which summaries the correlations between all the possible pairs of variables, and after this step, the principal components are identified by computing the eigenvectors and the eigenvalues of the covariance matrix. The feature vector is a matrix of vectors, that includes the most important (of high eigenvalues) components. The last step of PCA is recasting the data along the axes of the principal component. This step aims to use the feature vector to reorient the data from the original axes to the ones represented by the principal components. To do this, the transpose of the original data set is multiplied by the transpose of the feature vectors.

8 Experiments

In this section, the experimental setup and the performance of the created uni-modal and multi-modal emotion recognition models will be presented and compared to the baseline, using the evaluation metrics, which were presented in Section 4.

8.1 Experimental setup

In this project different experiments are performed, to evaluate the performance of each model in the human emotion recognition task. Two main categories of experiments can be defined, the experiments for uni-modal emotion recognition and the experiments for multi-modal emotion recognition.

The initial goal of this work is to make research on human emotion recognition by using data coming from only one modality. Thus, experiments based solely on one modality were made. EmoT, EmoA, and EmoFH were trained and tested using features, that were extracted from text, audio, and facial-hand movement data, as described in Section 7.1, respectively.
After investigating and creating those models, multi-modal emotion recognition experiments are performed. The TMER model is trained using the fusion of features from three different modalities. In addition, another model, TMERT, a model with a transformer-based attention mechanism as described in Section 6.3 is evaluated.

8.2 Experimental results

The results of the uni-modal and multi-modal models, compared with the baseline models, are presented in Table 5. (N. Majumber et al. in [18] obtained results not only by combining all the different modalities but also for each modality separately.) In general, EmoA and EmoFH outperformed the baseline by a margin of about 12-13%. On the other hand, EmoT performed worse than the baseline reaching an accuracy of about 69%, while the baseline model achieved an accuracy of 73.6%. In addition, both of the created multi-modal models, TMER and TMERT achieved better results than the baseline.

Modality Combination	Model	Accuracy	F1 Score
Text (T)	Baseline	**73.6%**	-
Text (T)	EmoT	69.35%	69.09%
Audio (A)	Baseline	53.3%	-
Audio (A)	EmoA	**65.77%**	63.93%
Video (V)	Baseline	57.1%	-
Video (V)	EmoFH	**69.05%**	68.90%
A+T+V	Baseline	76.5%	76.8%
A+T+V	TMER	81.01%	80.94%
A+T+V	TMERT	**81.32%**	**81.27%**

Table 5: Results for the different models compared with the baseline, using the two evaluation metrics. Accuracy constitutes the primary metric. The highest results for each metric are displayed in boldface.

Based on the obtained results, two general observations can be made. First of all, the features extracted in this project are superior to the features used in [18], and secondly, the current fusion method is more effective than the fusion method of the baseline.

- Uni-modal Experiments

  Regarding the uni-modal emotion recognition systems, it is obvious that the features that were extracted from the audio files, the facial expressions, and the hand movements that were used in EmoA and EmoFH respectively, manage to improve the performance of the baseline. The facial expressions and hand movements performed best over all the uni-modal models with a classification accuracy of 70.85%. However, in the case of text modality, EmoT, which uses the BERT model to classify the emotions performs worse than the baseline. This could be explained because, in the baseline method, the test data comes from the same domain as the training data, meaning that the word2vec model will encounter fewer tokens that do not appear in the training vocabulary. In addition, if the vocabulary size is small, the word2vec model captures relationships between fewer tokens, which can lead to a model that performs better than BERT. The main reason is that BERT as a subword model loses the relationships between fixed tokens in the text data. Instead it tries to generalize relationships across more than 30K subword tokens, as in the case of the bert-based-uncased model, which could lead to noise.

- Multi-modal Experiments

  Literature [24] refers that multi-modal analysis outperforms the unimodal analysis. In this project, the same trend is observed, as both the trimodal emotion recognition systems outperform the unimodal systems.
  The TMER model employs three CNNs for the training of the text, audio and visual inputs independently. The two convolutional layers, that were used on each input managed to point out the most important features, which then are all fused, by the feature-level fusion technique. In addition, the TMERT model, which uses a transformer and is based completely on attention, manages to capture the most important information of the data used during the training process and achieved a great performance in the human emotion classification.

  Both trimodal emotion recognition systems, achieve an accuracy score of about 81%. Specifically, TMERT achieves the best result improving the baseline by almost 5%. The accuracy is enhanced from 76.5% to 81.32% and the F1 score rises from 76.8% to 81.27%.

  In Figures 13 and 14, the confusion matrices of the results of the TMER and TMERT on the test set are presented. The vertical axis illustrates the actual labels of emotions, while the horizontal one shows the labels predicted by each model respectively. Overall, both TMER and TMERT achieve competitive scores, with a few outliers, especially in the case where the label is equal to 2 (happy). However, this can be explained by the size of the respective data, which is significantly smaller compared to the other three categories.

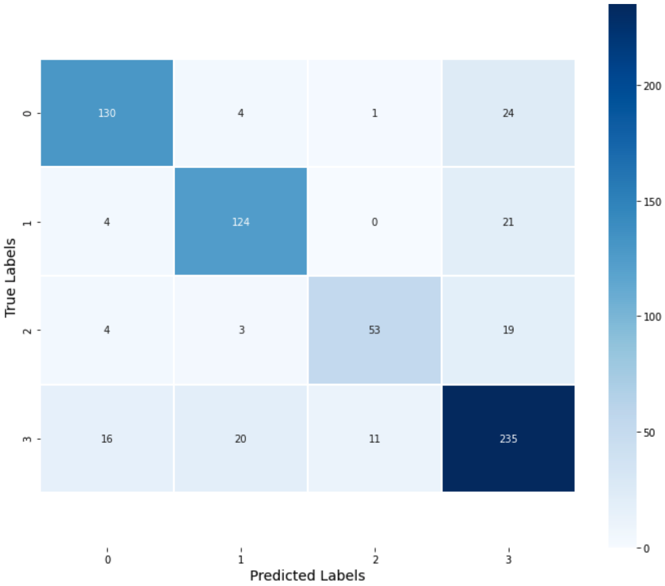

Figure 13: Confusion matrix of the results of TMER model on the test set. The vertical axis shows the actual labels, while the horizontal one shows the predicted labels.

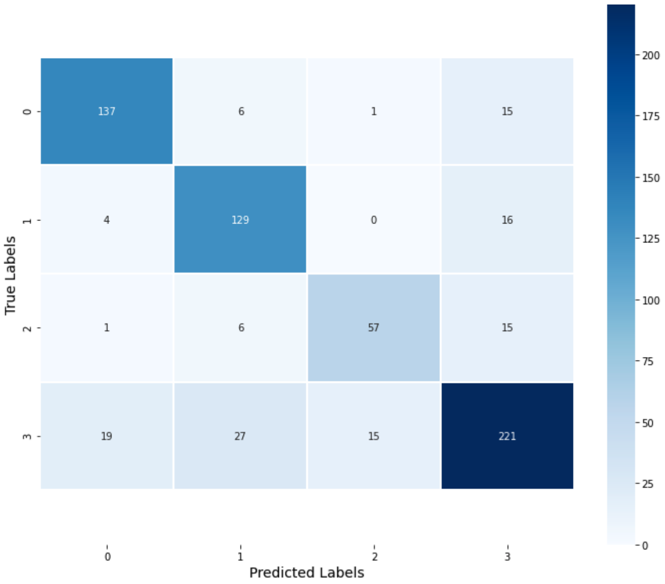

Figure 14: Confusion matrix of the results of TMERT model on the test set. The vertical axis shows the actual labels, while the horizontal one shows the predicted labels.

9 Conclusion

Emotion recognition constitutes a hot topic and attracts the interest of the research community. Several scientific projects on human emotion classification have been made in the last years taking into consideration multiple modalities. Text, audio, facial gestures, physiological signals, and several combinations of those are taken into account to better estimate human emotions.

In this project, features from each modality were extracted and tested in the EmoT model, the EmoA model, and the EmoFH model using deep learning. Overall, EmoA and EmoFH outperformed the baseline significantly. The extracted features and the architecture of those models, manage to achieve great performance on each modality respectively. On the other hand, EmoT, which uses BERT to classify human emotions, did not enhance the performance of the baseline, which could be attributed to the fact that BERT tries to generalize relationships across more than 30K subword tokens, as in the case of the bert-based-uncased model. Thus, it could lose the relationships between fixed tokens in the text data. Contextual information is really useful in the task of human emotion recognition, and without them, the training process of an emotion recognition system will be less effective.

Multi-modal learning is way more efficient compared with the unimodal [24], thus features from different modalities such as speech, text, facial expressions, and hand movements are fused to provide qualitative information that helps to greatly recognize the human emotions.

Two models were created to handle the task of emotion detection. TMER model consists of multiple layers of CNNs, while TMERT is a transformer-based model, which is based on

attention mechanisms.
Indeed the results of the experiments made in this research are still in line with the earlier findings in literature. Both tri-modal emotion recognition systems manage to capture the most important information of the IEMOCAP data and obtain competitive results, reaching an accuracy score of about 81%. However, recently, after our studies, Abhinav Joshi et al. in [11] manages to better classify human emotions, by using a Graph Neural Network (GNN) based architecture, reaching an F1-score equal to 84.50%

10 Future research

Even though our research was successful and we were able to achieve competitive results during the period of our experiments, some improvements could still be made in further research.
Our research regarding the uni-modal emotion recognition system that uses text data may have certain limitations due to the system's poor verbal text data adaptation. English texts from books and Wikipedia were used to train the pre-trained model. Results could be significantly improved by utilizing a BERT model that has been pre-trained on a large corpus of verbal text data. We suspect this because there will be a lot of specialized language that was not present in the dataset used to train "bert-base-uncased."
Furthermore, including data that belong to other emotion categories would be an interesting addition as a future research. Currently, the proposed uni-modal and multi-modal emotion recognition systems were trained to distinguish four different categories of emotions. By using data that belongs to all the categories of emotions, that the IEMOCAP dataset entails, the results of the proposed human emotion recognition systems could be differentiated.

References

[1] Model: bert-base-uncased. https://huggingface.co/bert-base-uncased. Accessed: 2021-01-08.

[2] Batbaatar, E., Li, M., and Ryu, K. H. Semantic-emotion neural network for emotion recognition from text. *IEEE Access* (2019), 111866–111878.

[3] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation 42* (12 2008), 335–359.

[4] Chiorrini, A., Diamantini, C., Mircoli, A., and Potena, D. Emotion and sentiment analysis of tweets using bert. In *EDBT/ICDT Workshops* (2021).

[5] De La Hoz, E. J., Martinez Santos, J. C., Contreras-Ortiz, S., and Dominguez, J. A. A machine learning model for emotion recognition from physiological signals. *Biomedical Signal Processing and Control 55* (01 2020), 101646.

[6] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.

[7] Fu, C., Liu, C., Ishi, C. T., and Ishiguro, H. Multi-modality emotion recognition model with gat-based multi-head inter-modality attention. *Sensors 20*, 17 (2020).

[8] Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 154–164.

[9] Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. F. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *CoRR abs/1908.11540* (2019).

[10] Huang, Y., Yang, J., Liao, P., and Pan, J. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational intelligence and neuroscience 2017* (2017).

[11] Joshi, A., Bhat, A., Jain, A., Singh, A., and Modi, A. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, United States, July 2022), Association for Computational Linguistics, pp. 4148–4164.

[12] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1725–1732.

[13] Khalfallah, J., and Slama, J. B. H. Facial expression recognition for intelligent tutoring systems in remote laboratories platform. *Procedia Computer Science 73* (2015), 274–281. International Conference on Advanced Wireless Information and Communication Technologies (AWICT 2015).

[14] KrishnaD., N., and Patil, A. Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *INTERSPEECH* (2020).

[15] Li, J., Zhang, M., Ji, D., and Liu, Y. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *CoRR abs/2003.01478* (2020).

[16] Lian, Z., Li, Y., Tao, J., and Huang, J. Investigation of multimodal features, classifiers and fusion methods for emotion recognition, 2018.

[17] Loshchilov, I., and Hutter, F. Decoupled weight decay regularization. In *ICLR* (2019).

[18] Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., and Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems 161* (2018), 124–133.

[19] Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A. F., and Cambria, E. Dialoguernn: An attentive RNN for emotion detection in conversations. *CoRR abs/1811.00405* (2018).

[20] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence 34* (04 2020), 1359–1367.

[21] Mustaqeem, and Kwon, S. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors 20* (12 2019), 183.

[22] Pan, Z., Luo, Z., Yang, J., and Li, H. Multi-modal attention for speech emotion recognition. *ArXiv abs/2009.04107* (2020).

[23] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, Canada, July 2017), Association for Computational Linguistics, pp. 873–883.

[24] Poria, S., Chaturvedi, I., Cambria, E., and Hussain, A. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (2016), pp. 439–448.

[25] Reimers, N., and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 3982–3992.

[26] Siriwardhana, S., Reis, A., Weerasekera, R., and Nanayakkara, S. Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition. *ArXiv abs/2008.06682* (2020).

[27] Tursunov, A., ., M., and Kwon, S. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors 20* (09 2020), 5212.

[28] Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *ArXiv abs/1706.03762* (2017).

[29] Yamashita, R., Nishio, M., Do, R., and Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging 9* (06 2018).

[30] Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., and Zhou, G. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI* (2019).

[31] Zhong, P., Wang, D., and Miao, C. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 165–176.