big data and cognitive computing

IMPACT FACTOR 3.7

CITESCORE 7.1

Article

MM-EMOR: Multi-Modal Emotion Recognition of Social Media Using Concatenated Deep Learning Networks

Omar Adel, Karma M. Fathalla and Ahmed Abo ElFarag

Special Issue

Challenges and Perspectives of Social Networks within Social Computing

Edited by

Dr. Maria Chiara Caschera, Dr. Patrizia Grifoni and Dr. Fernando Ferri

MDPI

https://doi.org/10.3390/bdcc7040164

big data and cognitive computing

MDPI

Article

MM-EMOR: Multi-Modal Emotion Recognition of Social Media Using Concatenated Deep Learning Networks

Omar Adel *, Karma M. Fathalla and Ahmed Abo ElFarag

Department of Computer Engineering, Faculty of Engineering and Technology, Arab Academy for Science, Technology and Maritime Transport (AAST), Alexandria 1029, Egypt
* Correspondence: omar95adel@aast.edu; Tel.: +20-111-439-4765

Abstract: Emotion recognition is crucial in artificial intelligence, particularly in the domain of human–computer interaction. The ability to accurately discern and interpret emotions plays a critical role in helping machines to effectively decipher users' underlying intentions, allowing for a more streamlined interaction process that invariably translates into an elevated user experience. The recent increase in social media usage, as well as the availability of an immense amount of unstructured data, has resulted in a significant demand for the deployment of automated emotion recognition systems. Artificial intelligence (AI) techniques have emerged as a powerful solution to this pressing concern in this context. In particular, the incorporation of multimodal AI-driven approaches for emotion recognition has proven beneficial in capturing the intricate interplay of diverse human expression cues that manifest across multiple modalities. The current study aims to develop an effective multimodal emotion recognition system known as MM-EMOR in order to improve the efficacy of emotion recognition efforts focused on audio and text modalities. The use of Mel spectrogram features, Chromagram features, and the Mobilenet Convolutional Neural Network (CNN) for processing audio data are central to the operation of this system, while an attention-based Roberta model caters to the text modality. The methodology of this study is based on an exhaustive evaluation of this approach across three different datasets. Notably, the empirical findings show that MM-EMOR outperforms competing models across the same datasets. This performance boost is noticeable, with accuracy gains of an impressive 7% on one dataset and a substantial 8% on another. Most significantly, the observed increase in accuracy for the final dataset was an astounding 18%.

Keywords: classification; MobileNet; Roberta; multimodal; emotion; recognition; IEMOCAP; MELD; social media

Citation: Adel, O.; Fathalla, K.M.; Abo ElFarag, A. MM-EMOR: Multi-Modal Emotion Recognition of Social Media Using Concatenated Deep Learning Networks. *Big Data Cogn. Comput.* **2023**, *7*, 164. https://doi.org/10.3390/bdcc7040164

Academic Editors: Maria Chiara Caschera, Patrizia Grifoni and Fernando Ferri

Received: 30 August 2023
Revised: 11 October 2023
Accepted: 12 October 2023
Published: 13 October 2023

1. Introduction

Correctly perceiving different people's emotions is a key factor in proper communication. Emotional understanding makes social networking more natural by eliminating ambiguity and helps interpret the conveyed messages effectively. Due to the importance and complexity of emotions in conversations, emotion recognition has become one of the vital fields of study applying artificial intelligence (AI) techniques. In addition, the ubiquitous use of social networking services created further demand for automated content and emotion analysis using machine learning, as it has become infeasible to analyze them otherwise [1].

Emotion recognition has a wide range of applications in various fields such as robotics, security, healthcare, automated identification, customer support call review, and lie detection. It also plays an essential role in improving human–computer interaction (HCI) [2]. It can help the machines to take user feedback to enhance the user experience. The diversity of the applications and the availability of big data volumes stipulate the significance of developing novel approaches for emotion recognition. Emotional expression can be verbal and non-verbal. The human perception of emotion involves capturing and analyzing facial

expressions, speech text, and voice examination. Hence, automated emotion recognition can bridge the gap between humans and machines [3].

Speech recognition has gained a lot of focus over the last decades. It is an important research area for human-to-machine communication. Early methods focused on manual feature extraction and conventional techniques such as Gaussian mixture models (GMM) [4], and Hidden Markov models (HMM) [5]. More recently, neural networks such as recurrent neural networks (RNNs) [6], and convolutional neural networks (CNNs) [7] have been applied to speech recognition and have achieved great performance.

In addition, emotion extraction from text is of huge importance and is an uprising area of research in natural language processing. Recognition of emotions from text has high practical utilities for quality improvement like in the field of HCI. Most works on it have proposed solutions based on neural network techniques like LSTM [8], and CNN [9] with adequate results.

Despite the extensive development of unimodal learning models, they still cannot cover all the aspects of human interpretation of emotion. People present their emotions with various modes of expression. The combination of speech and text analysis carries the potential for improving emotion recognition [10]. Hence, in this study, a multimodal emotion (MM-EMOR) analysis system is proposed to examine audio and text data.

While the MM-EMOR system uses well-established technologies such as MobileNet for image data and ROBERTA for text data, we feel that its originality comes in the pragmatic merging of various modalities. This integration, while seemingly simple, is motivated by the desire for a more thorough knowledge of human emotions. Furthermore, our motivation goes beyond the technological aspects. We envision MM-EMOR as a versatile tool with far-reaching implications in disciplines such as mental health, human–computer interaction, and beyond. The capacity to effectively recognize and interpret emotions can improve our ability to communicate, empathize, and, eventually, improve the human experience in an increasingly digital environment.

The proposed MM-EMOR is a collaborative effort to process and synthesize information from a variety of modalities [11]. MM-EMOR uses a multimodal approach to bridge the gap between these modalities, ushering in a new era of sophisticated emotion recognition. The meticulously designed unimodal learning models are central to the architecture of MM-EMOR, each of which is tailored to perform well within its respective domain while still yielding competitive overall performance. Notably, the audio-based model stands out as a standout component, harnessing the subtle nuances of emotions via the complex interplay of Mel Spectrogram and Chromagram features. These audio representations give each emotion a distinct identity, adding depth to our understanding of emotional states. Furthermore, modalities are integrated using a seemingly simple yet effective concatenation technique that harmoniously blends the insights gained from audio and other modalities to achieve a robust multimodal emotion classification. The proposed MM-EMOR system stands as a testament to its efficacy in the grand tapestry of emotion recognition, as evidenced by its noteworthy performance that outperforms existing benchmarks in the field.

The rest of this paper is organized as follows. In Section 2, we clarify the related work which motivated and inspired the current study. In Section 3, we define the materials and methods used in this study. In Section 4, the results of MM-EMOR and its comparison with the state of the work are presented. Finally, conclusions are given in Section 5.

2. Related Works

Due to the importance of emotion recognition, many studies have been directed toward improving the performance of emotion recognition systems. The studies employed various deep learning methods and algorithms to distinguish human emotions [10,12,13]. Recently, Huddar et al. [14] proposed a cascaded approach for merging modalities, where textual features were extracted using CNN, audio features using the openSMILE toolkit [15], and visual features using 3D-CNN. They extracted the unimodal context using biLSTM and

merged two modalities at a time to get the bimodal features. Similarly, biLSTM extracts the bimodal context and merges it to get the trimodal features. The adopted approach makes the complexity of the algorithm high. The IEMOCAP dataset [16] was used to train and test their model, with four classes (happy, angry, sadness, and neutral).

Kumar et al. [17] made a combination of three kinds of acoustic features (Mel spectrogram, MFCC, and chroma vectors) and trained it by RNN. For lexical features, they used the Bert model (bidirectional encoder representations from transformers), which is a transformer-based machine-learning technique for natural language processing. After that, they concatenated these features and used drop-out and dense layers to make the prediction. The IEMOCAP dataset was used to verify the performance of their model, with four classes (happy merged with excited, angry, sadness, and neutral).

Guo et al. [10] used BLSTM for text modality and audio modality, where the weights between different modalities needed to be considered to obtain a richer comprehensive emotional representation. As such, they used a weighted fusion layer. The IEMOCAP dataset was used with four classes (happy, angry, sadness, and neutral).

Singh et al. [18] extracted 33 audio features (16 prosody-based, 16 spectral-based, and one voice quality-based feature), in addition to the MFCC feature. For text, they used ELMo, which is a state-of-the-art NLP framework developed by AllenNLP and trained on the one Billion Word Benchmark. After being concatenated, they used the DNN model for prediction. To test their model performance, they used the IEMOCAP dataset with classes (happy merged with excited, angry merged with frustration, sadness, and neutral).

Wang et al. [19] proposed a new method called multimodal transformer augmented fusion (MTAF) for SER. MTAF uses a hybrid fusion strategy that combines feature-level fusion and model-level fusion. A model-fusion module composed of three cross-transformer encoders was proposed to generate a multimodal emotional representation for modal guidance and information fusion. Specifically, the multimodal features obtained by feature-level fusion and text features were used to enhance speech features. Experiments were implemented on the IEMOCAP dataset and the MELD dataset.

Zaidi et al. [20] proposed a multimodal dual attention transformer for cross-language speech emotion recognition. The multimodal dual attention transformer combined two attention mechanisms: one for the acoustic features of the speech signal called Roberta, and one for the textual features called wav2vec. The proposed approach was evaluated on the IEMOCAP dataset.

Canal et al. [21] provided a significant contribution to the field of facial expression analysis. This study investigated novel ways of facial expression identification using state-of-the-art machine learning algorithms and computer vision methodology. The research looked at the development of deep learning models, namely convolutional neural networks (CNNs), for extracting relevant characteristics from facial photos and accurately detecting and classifying a wide range of emotional states. In addition, it covered the use of these techniques in a variety of domains, highlighting their potential impact on fields such as human–computer interaction, affective computing, and emotion-aware systems. The insights and approaches described in this research are a significant resource for researchers and practitioners interested in advancing the subject of facial expression analysis and its practical applications.

The outlined related work shows promising results in emotion recognition, but some limitations exist due to the nature of the used datasets like multimodal integration challenges. Combining information from different modalities (e.g., text, audio, video) into a unified model can be complex and may introduce additional challenges [22]. This limitation needs a powerful model to overcome this. Our approach involves training Roberta, a highly performant natural language processing model, to extract textual features. Simultaneously, for audio data, we applied a feature extraction process to obtain a set of powerful multiple features from the acoustic domain. These features are then merged to capture complex relationships between text and audio data.

Another possible limitation is overfitting of emotion recognition models, particularly deep learning models, which may overfit the training data, leading to high training accuracy but poor generalization to new, unseen data [23]. We tried to overcome this by adding a dropout layer and regularizer to avoid overfitting in our model.

Another characteristic challenge of the problem is continuous emotions where some emotions are continuous and can vary in intensity, making it difficult to assign discrete labels [24]. We used the Roberta model which can overcome this. Therefore, we propose MM-EMOR an effective approach that can overcome these limitations.

3. Materials and Methods

MM-EMOR aims to analyze multimodal speech for emotion recognition. It merges textual and audio modalities for this task. The proposed system's architecture is described in Figure 1, with textual- and audio-based emotion recognition modules. The system is explained in detail in the following sections.

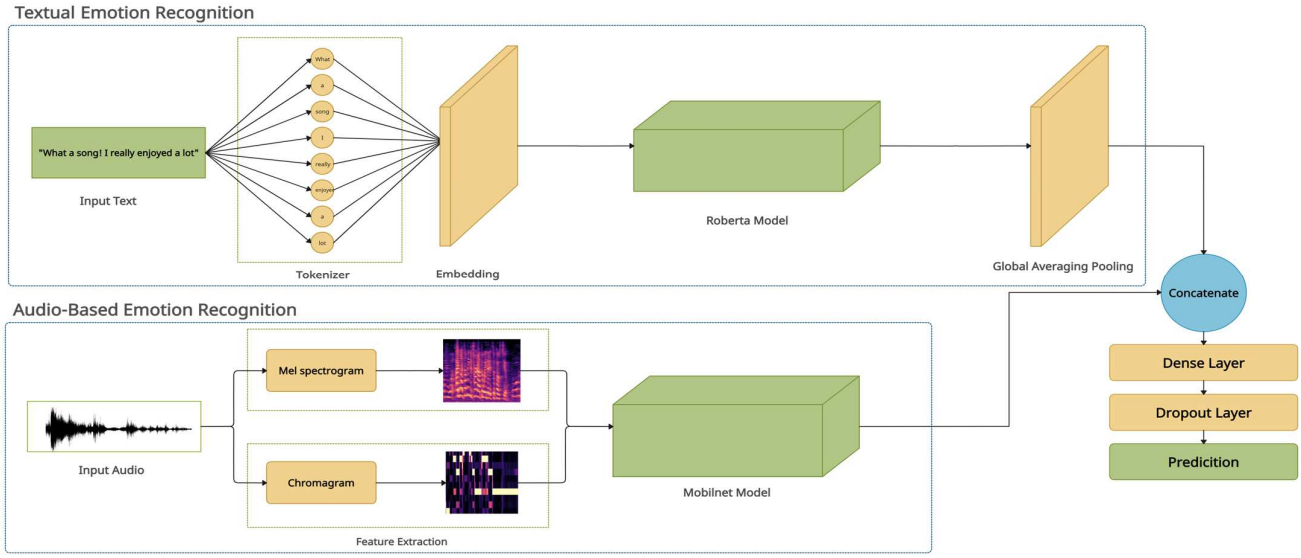

Figure 1. The proposed MM-EMOR system.

3.1. Textual Emotion Recognition

Roberta

In this part, the textual emotion recognition model is described. The core pivotal component of this module is the Roberta model. Facebook AI Research (FAIR) unveiled the robust language model known as Roberta in 2019 [25].

The Bidirectional Encoder Representations from the Transformers (BERT) model [26], another well-known NLP model, is referenced in the model's name since it incorporates many of BERT's advantages. Its bidirectional context allows it to comprehend text completely, and it employs transfer learning to rapidly adapt to varied NLP tasks. BERT's contextual embeddings efficiently capture nuances, and its large-scale architecture allows it to represent long-term dependencies.

To improve its functionality and sturdiness, Roberta is an improved version of the BERT model that has several advantages. It employs a better training methodology, a larger dataset, and data augmentation techniques. Roberta eliminates the next sentence prediction (NSP) assignment from BERT's training procedure and concentrates on successfully predicting masked tokens. Typically, the model is trained on a larger dataset, resulting in stronger contextual embeddings and increased performance on downstream NLP tasks. Roberta requires less fine-tuning and is more generalizable across domains. Its repeatability and consistent outstanding performance has made it a preferred choice for a wide range of NLP applications.

The training strategy is one significant difference between Roberta and BERT. BERT was trained with a masked language modeling (MLM) objective, whereas Roberta uses a version known as "dynamically masked language modeling" (DMLM). Instead of randomly masking words during training, DMLM trains the model on a vast volume of unlabeled text with no masks. This allows for improved generalization by exploiting the entire context of the text.

Roberta also has a larger training corpus than BERT, which includes both in-domain and out-of-domain data. It is trained using large amounts of text data from books, websites, and Wikipedia. Due to the thorough pretraining, the model can learn a wide range of linguistic patterns and semantic correlations. We used the Roberta-Base model which has 12 layers of the transformer model as shown in Figure 2. We selected a token length of 1000 due to the inherent characteristics of our datasets, characterized by the presence of lengthy text sentences. In addition, longer token lengths can lead to longer training times, so there is often a trade-off between training duration and model performance. A token length of 1000 tokens might strike a reasonable balance.

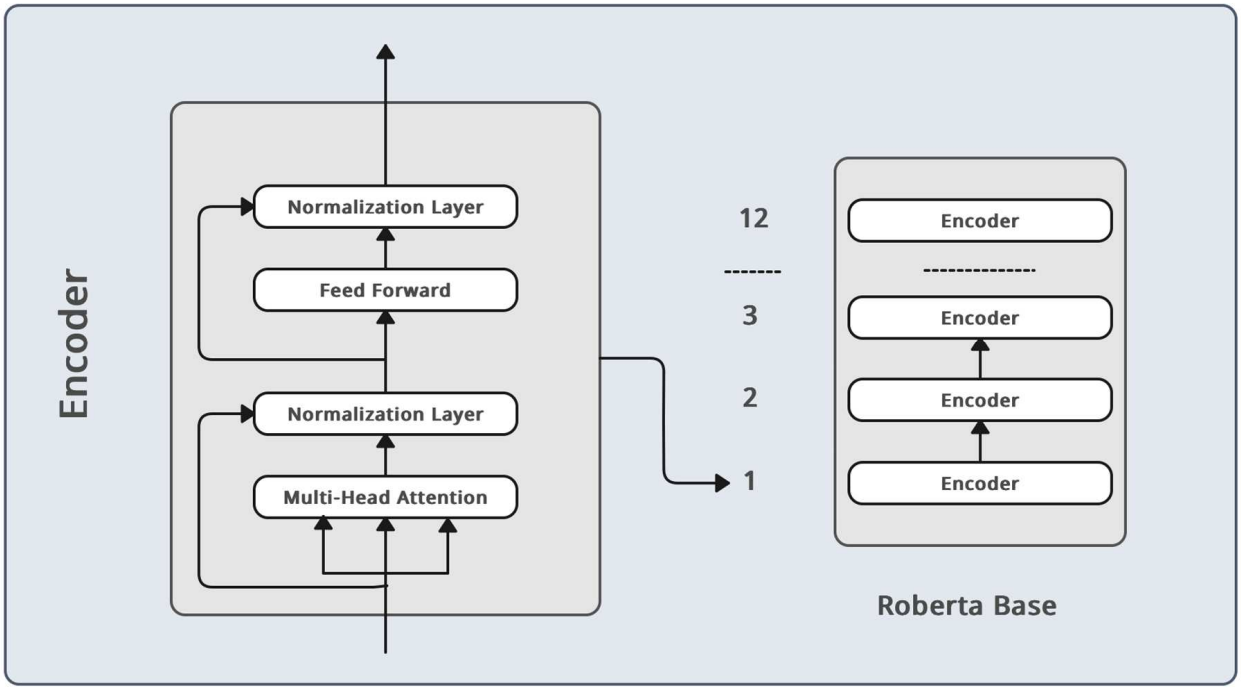

Figure 2. Roberta model architecture.

3.2. Audio-Based Emotion Recognition

3.2.1. Feature Extraction

Audio signals are dynamic and multidimensional; they are difficult to analyze and comprehend. They are comprised of temporal, spectral, and frequently harmonic data. Variations in pitch, timbre, rhythm, and dynamics, as well as the existence of background noise and transitory occurrences, add to the complications. Audio signals can also transmit emotions, intentions, and cultural nuances, adding another degree of complication to their interpretation. Two main feature representations were emphasized to capture different features of these signals: spectrograms and chromagrams.

Mel spectrogram features are extracted from the audio signals. The extracted features are well suited to the task of emotion recognition [27,28]. Each frame of the spectrum in the Mel spectrum contains a short-time Fourier transform (STFT), which maps the signal from a linear frequency scale to a logarithmic Mel scale. After that, it is applied to the filter bank to produce the eigenvector, whose eigenvalues can be roughly described as the distribution of signal energy on the Mel-scale frequency. To extract audio features, we used Librosa [29] to extract the Mel spectrogram features, which were converted into a 2D image.

To train the convolutional neural networks (CNN) for recognition, the generated Mel spectrogram-based 2D images were input to the network. In Figure 3, we show some examples of Mel-spectrogram images of various emotions. As can be seen from the figure below, there are evident differences between various types of emotions. After representing the audio data as an image, MobileNet CNN was used to model the data and perform the emotion classification task.

Figure 3. Differences between classes Mel spectrogram.

The chromagram is a characteristic that is extensively utilized in audio signal processing, including AI and machine learning applications. It depicts the distribution of energy in different frequency bands or pitches in an audio signal over time.

A chromagram is a two-dimensional representation of an audio signal in which time is represented horizontally and frequency bands are shown vertically. It is frequently estimated using the short-time Fourier transform (STFT) approach, which analyzes an audio signal's spectrum over short overlapping time periods.

The chromagram's vertical axis is often quantized into a given number of bins, each representing a specific pitch or frequency range.

To generate a chromagram, an audio signal is separated into frames or windows, and the frequency content of each frame is analyzed using the STFT. The generated spectrum is then mapped to the correct chroma bins by grouping frequencies in the same pitch.

In AI and machine learning problems such as audio classification, chromagrams are frequently employed as a feature representation. Chromagrams can capture the underlying

musical structure, tonality, and chord progressions in an audio signal by describing it in terms of its harmonic content.

Once computed, the chromagram can be utilized as an input feature for machine learning methods such as convolutional neural networks (CNNs) as an image. These models can learn to recognize patterns and extract meaningful information from chromagram representations.

Figure 4 shows the difference between emotions. We used Librosa to extract chromagram features and merged them with the Mel-spectrogram to use it to train the MobileNet CNN model.

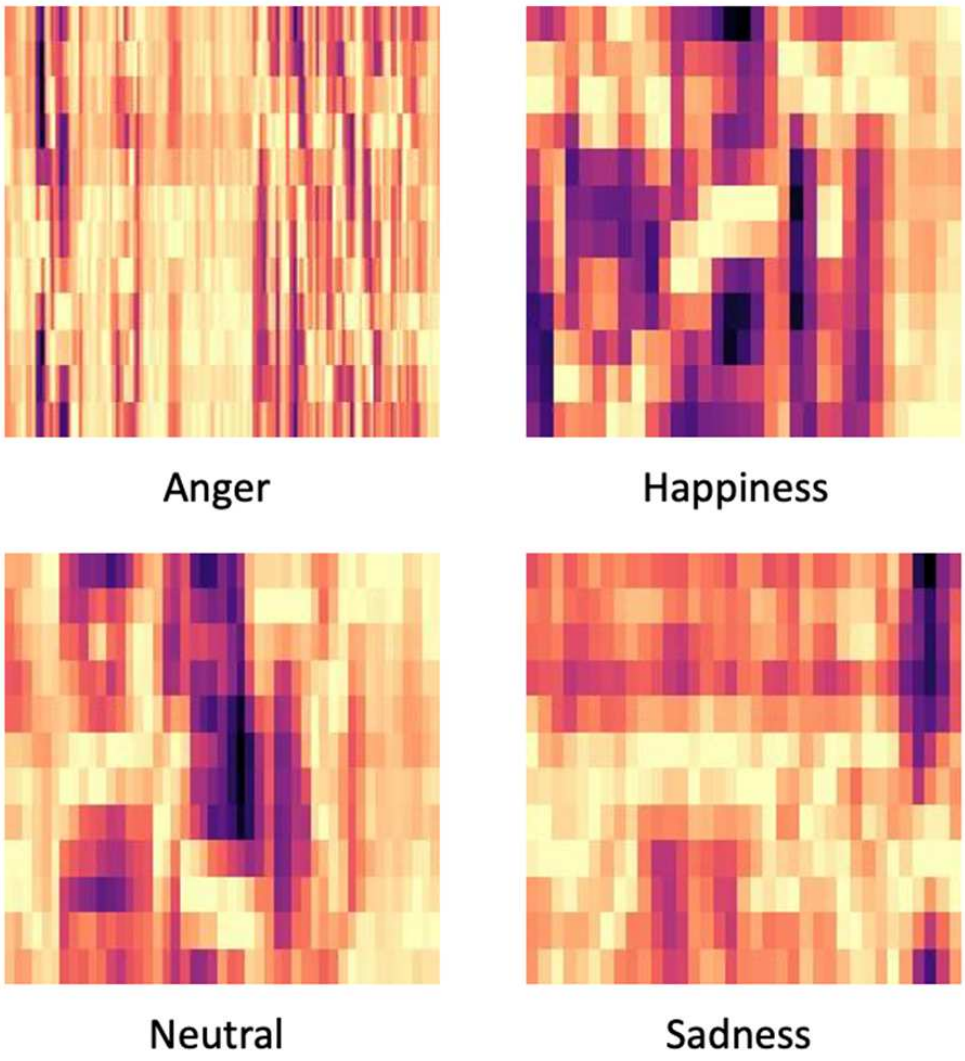

Figure 4. Differences between classes chromagram.

3.2.2. MobileNet

For the purpose of audio classification, we opted for a 2D image-based approach over 1D audio-based classification due to several compelling reasons. Firstly, 2D image-based classification is useful when working with audio data that can be converted into spectrograms. Spectrograms provide a comprehensive representation of essential temporal and frequency domain information, making them ideal for image-based classification applications. This approach is commonly used in speech recognition and sound analysis applications. Secondly, using 2D image-based techniques has the great advantage of utilizing pre-trained image classification models, particularly convolutional neural networks (CNNs). This not only saves time and computing resources but also takes advantage of the amount of knowledge contained inside well-established image models. It is important to

note that Solovyev et al. [30] conducted a comparative study and similarly advocated for the 2D image-based approach, achieving better results in comparison to 1D audio-based methods. We chose to use mobilNet model for audio classification.

MobileNet was first introduced in 2017 by Howard et al. [31], MobileNet is an efficient and portable CNN architecture, which is a sub-category of neural networks and is currently one of the most efficient image classification models. MobileNet is used to build lighter models by using depth-wise separable convolutions in place of the standard convolutions used in earlier architectures. MobileNet introduces two new global hyperparameters (width multiplier and resolution multiplier) that allow model developers to trade off latency or accuracy for speed and small size depending on their requirements.

The MobileNet model is primarily based on depth-wise separable convolutions that are a shape of standard convolution right into a depth-wise convolution, which is a type of convolution where we apply a single convolutional filter for each input channel and a 1×1 convolution known as a pointwise convolution. The pointwise convolution then applies a 1×1 convolution to combine the outputs of the depth-wise convolution as shown in Figure 5.

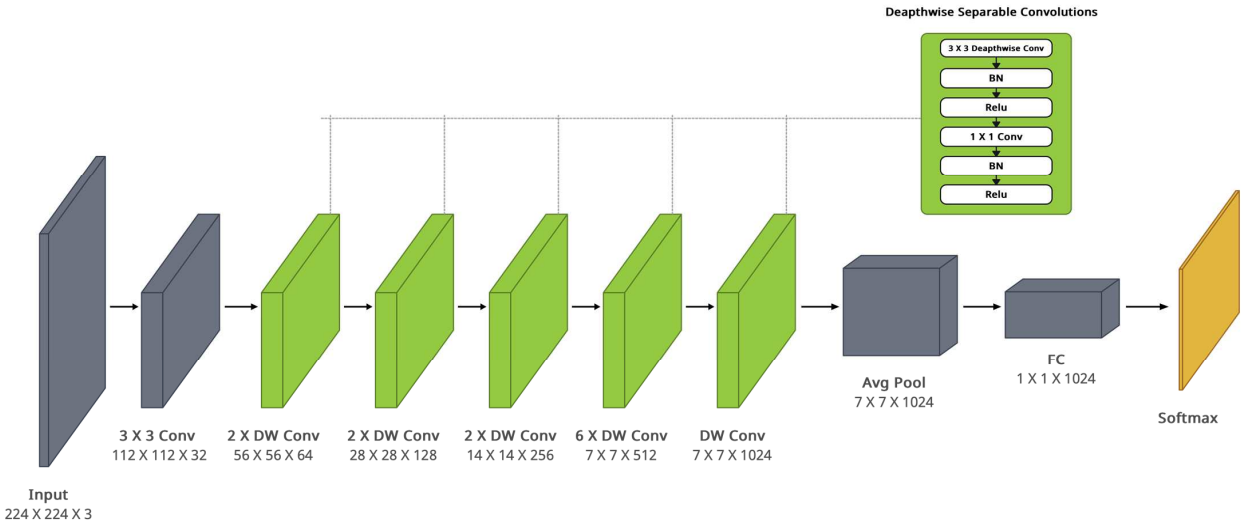

Figure 5. MobileNet architecture.

A standard convolutional layer takes as input a $D_F \times D_F \times M$ feature map F and produces a $D_F \times D_F \times N$ feature map G where:

- D_F is the spatial width and height of a square input feature map.
- M is the number of input channels (input depth).
- D_G is the spatial width and height of a square output feature map.
- N is the number of output channels (output depth).
- The standard convolutional layer is parameterized by convolution kernel K of size $D_K \times D_K \times M \times N$ where:
- D_K is the spatial dimension of the kernel assumed to be square.
- M is the number of input channels.
- N is the number of output channels as defined previously.

Standard convolutions have the computational cost of $D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F$

Depthwise convolution has a computational cost of $D_K \cdot D_K \cdot M \cdot D_F \cdot D_F$

The combination of depthwise convolution and 1×1 (pointwise) convolution is called depthwise separable convolution, which was originally introduced in [32].

Depthwise separable convolution costs are $D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$ which is the sum of the depthwise and 1×1 pointwise convolutions.

By expressing convolution as a two-step process of filtering and combining, we get a reduction in the computation of:

$$\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2} \quad (1)$$

MobileNet uses 3 × 3 depthwise separable convolutions which use between 8 to 9 times less computation than standard convolutions.

3.3. Concatenation and Prediction

After taking the output from the Roberta model, we used a global averaging pooling 1D layer to preprocess the data to concatenate with the MobileNet model. After concatenation, we used a dense layer with a ReLU activation function.

We also added a regularizer to avoid overfitting in our model. regularizers are techniques used to prevent overfitting, which occurs when a model fits the training data too closely and fails to generalize well to new, unseen data. Regularization helps improve a model's ability to generalize by adding a penalty term to the loss function, discouraging overly complex or extreme parameter values. This makes the model more robust and less prone to overfitting.

After this, we added a dropout layer with a drop probability of 0.3 to avoid overfitting. The outputs of the dense and dropout layers were finally used to predict the probabilities of the emotion classes. The proposed model was trained using the Adam optimizer [33] and relies on categorical cross-entropy for the loss metric.

4. Results and Discussion

In this section, a detailed evaluation of MM-EMOR performance on different datasets is performed.

4.1. Dataset

In this study, publicly available well-studied datasets were chosen to test the performance of the presented multimodal approach, namely Tweeteval [34], Multimodal Emotion Lines Dataset (MELD) [35], and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [16]. The chosen datasets allow comparison with state-of-the-art performance.

4.1.1. Tweeteval Dataset

Tweeteval [34] is used to analyze people's emotions on social networks and determine the success of the proposed model to recognize emotions from textual data. Tweeteval [34] consists of seven heterogeneous tasks in Twitter, all framed as multi-class tweet classifications. The tasks include irony, hate, offensive, stance, emoji, emotion, and sentiment. All tasks have been unified into the same benchmark, with each dataset presented in the same format and with fixed training, validation, and test split. The number of labels and instances in training, validation, and test sets for each task is shown in Table 1.

Table 1. Tweeteval dataset stratification for the used classes.

Task	Number of Classes	Train	Val	Test
Emotion	4	3257	374	1421
Hate	2	9000	1000	2970
Irony	2	2862	955	784
Offensive	2	11,916	1324	860
Sentiment	3	45,389	2000	11,906
Stance	3	2620	294	1249
Emoji	20	45,000	5000	50,000

4.1.2. IEMOCAP Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [16] is an acted multimodal and multi-speaker database, recently collected at the SAIL lab at the University of Sothern California (USC). It contains approximately 12 h of audiovisual data, including video, speech, motion capture of face, and text transcriptions. Our experiments will focus on speech and text transcriptions modalities. The dataset contains 9 classes namely anger, happiness, excitement, sadness, frustration, fear, surprise, other, and neutral.

The most commonly used classes in the literature are anger, happiness, sadness, and neutral. Hence, they will be used in our experiments. In addition, due to the small number of happiness class records, some of the related studies merged happiness with excitement. As such, we will be reporting the merged results. The merging of emotion classes has been done based on Plutchik's wheel of emotions [36]. The number of records in each class is shown in Table 2.

Table 2. IEMOCAP dataset stratification for the used classes.

Classes	Happiness	Anger	Sadness	Neutral
Happiness, anger, sadness, and neutral	595	1103	1084	1708
Happiness + excitement, anger, sadness, and neutral	1636	1103	1084	1708

4.1.3. MELD Dataset

The Multimodal Emotion Lines Dataset (MELD) [35] has more than 1400 dialogues and 13,000 utterances from the Friends TV series. Multiple speakers participated in the dialogues. Each utterance in dialogue has been labeled by any of these seven emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. MELD also has sentiment (positive, negative, and neutral) annotations for each utterance. The number of records in each class is shown in Table 3.

Table 3. MELD dataset stratification for the used classes.

Classes	Surprise	Neutral	Fear	Joy	Sadness	Disgust	Anger
Train	1205	4710	268	1743	683	271	1109
Test	281	1256	50	402	208	68	345
Validation	150	470	40	163	111	22	153

The proposed model was evaluated on these datasets. For all training and testing purposes, networks from the Keras library for Python were implemented. For IEMOCAP, tests were performed using 5-fold cross-validation, for Tweeteval and MELD, we used test records on the dataset.

4.2. Performance Measures

The performance of MM-EMOR was evaluated using the following evaluation metrics: accuracy, precision, recall, and F1-score. The expression of these metrics is given as follows:

- Unweighted accuracy is just the proportion of correctly predicted observations to all observations. $\frac{TP+TN}{TP+FP+TN+FN}$
- Weighted accuracy takes into account the class-specific accuracy and assigns different weights to each class based on their importance or prevalence in the dataset. $\sum(W_i \times acc_i)/\sum W_i$
- Recall is a metric for how well a model detects true positives. $\frac{TP}{TP+FN}$
- Precision is the ratio of accurately anticipated positive observations to all actual class observations. $\frac{TP}{TP+FP}$
- F1-score is the weighted average of precision and recall, weighted. $2\frac{Precision \times Recall}{Precision+Recall}$

4.3. Performance Evaluation

The training of the models takes a maximum of 30 epochs. The results are reported as the best mode reached.

4.3.1. IEMOCAP Dataset

Figure 6 and Table 4 show the confusion matrix and the results of (anger, happiness, neutral, and sadness) classes on the IEMOCAP dataset with our performance metrics. Each metric for a given modality generates different values for every emotion. The anger emotion has the best value of precision of 90%. For recall, neutral has the best value of 82%. Otherwise, in the F1-score anger has the best value of 85%.

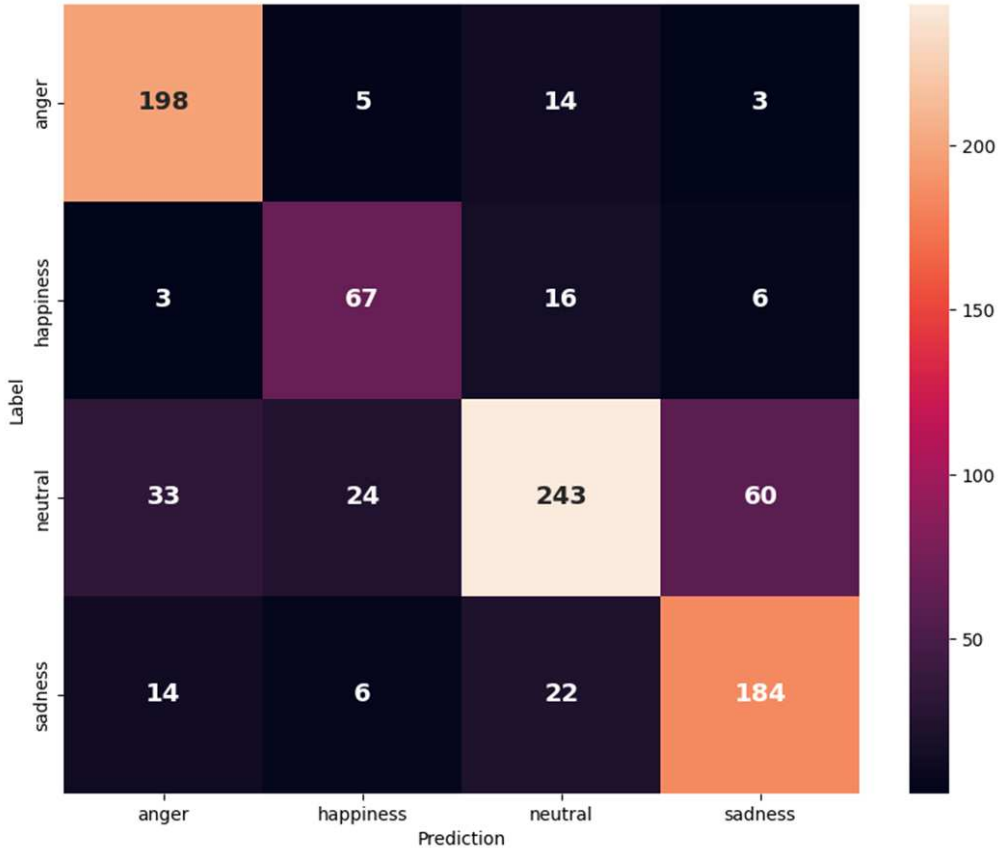

Figure 6. Confusion matrix for IEMOCAP (anger, happiness, neutral, and sadness) on the text and audio modalities.

Table 4. Classification report results for the IEMOCAP (anger, happiness, neutral, and sadness) on the text and audio modalities in the function of (precision, recall, and F1-score).

Classes	Precision	Recall	F1-score
Anger	0.90	0.80	0.85
Happiness	0.73	0.66	0.69
Neutral	0.68	0.82	0.74
Sadness	0.81	0.73	0.77

Figure 7 and Table 5 show the confusion matrix and the results of (anger, happiness merged with excitement, neutral, and sadness) classes on the IEMOCAP dataset. Happiness and sadness emotions have the same value of precision of 84%. Recall anger has the best value of 82%. In the F1-score, anger also has the best value of 82%.

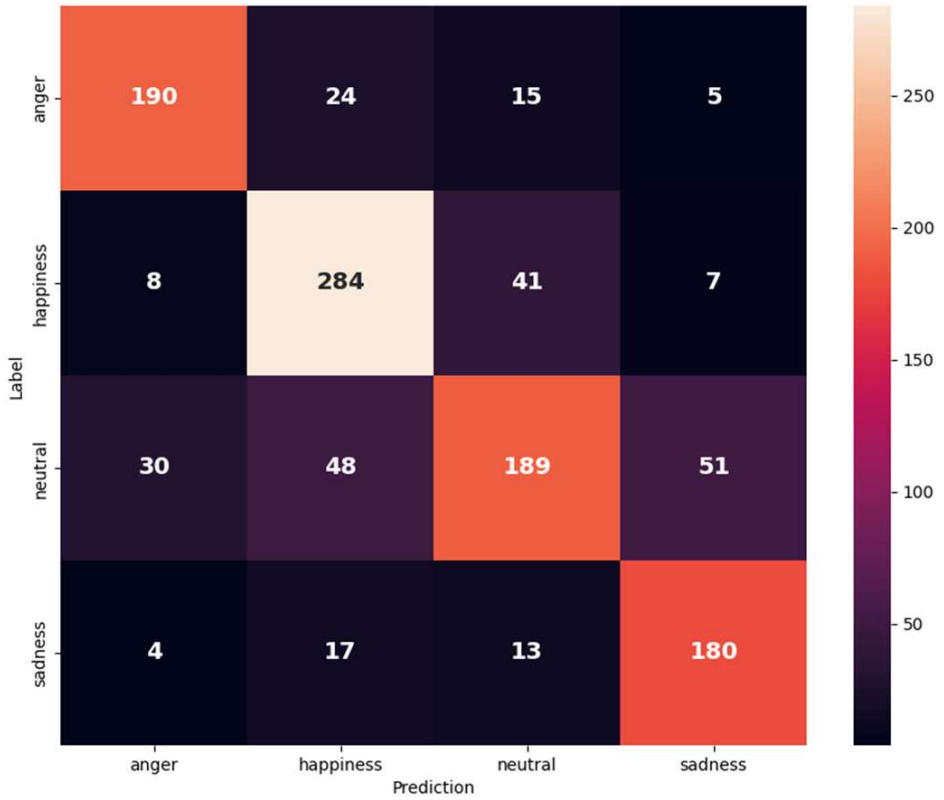

Figure 7. Confusion matrix of IEMOCAP (anger, happiness merged with excitement, neutral, and sadness) on the text and audio modalities.

Table 5. Classification report results for the IEMOCAP (anger, happiness merged with excitement, neutral, and sadness) on the text and audio modalities in the function of (precision, recall, and F1-score).

Classes	Precision	Recall	F1-score
Anger	0.81	0.82	0.82
Happiness	0.84	0.76	0.80
Neutral	0.59	0.73	0.66
Sadness	0.84	0.74	0.79

According to Table 6, the model has an accuracy of 77.06% in (happiness, anger, sadness, and neutral) classes, otherwise it has an accuracy of 76.22% in (happiness merged with excitement, anger, sadness, and neutral) classes.

Table 6. Accuracy results of the IEMOCAP dataset.

Classes	UA	WA
Happiness, anger, sadness, and neutral	0.7706	0.7792
Happiness + excitement, anger, sadness, and neutral	0.7622	0.7705

In Figure 8 the training accuracy versus epochs of the IEMOCAP dataset in using happiness, anger, sadness, and neutral classes is visualized. In Figure 9, the training accuracy versus epochs is illustrated, but for happiness and excitement, anger, sadness, and neutral classes. The figures elucidate improving accuracy with the increasing number of epochs, indicating the potential of the approach to model the data.

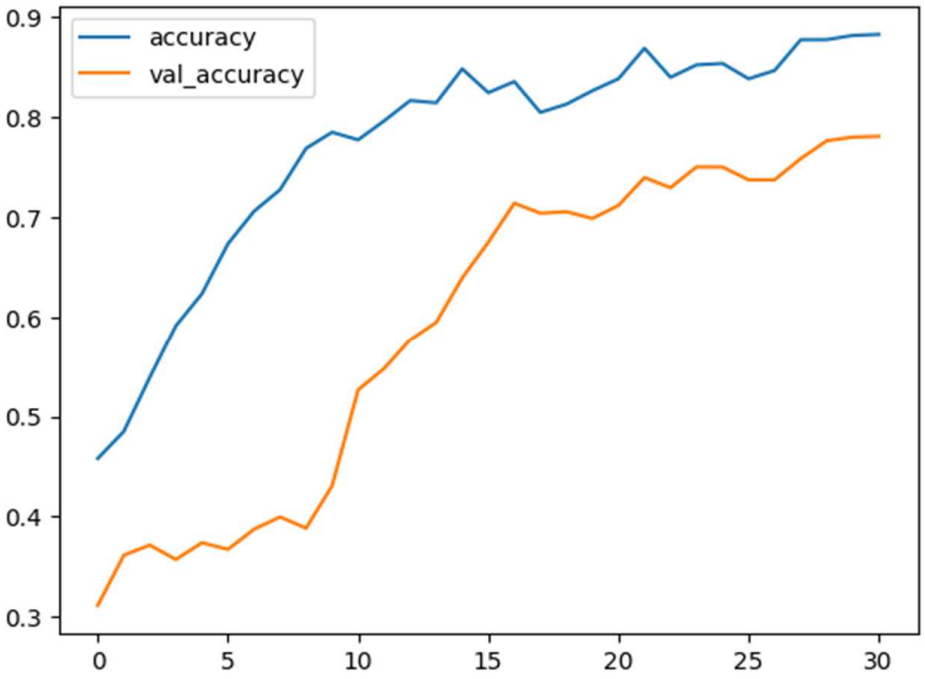

Figure 8. Training and validation accuracy versus epochs in happiness, anger, sadness, and neutral classes of IEMOCAP.

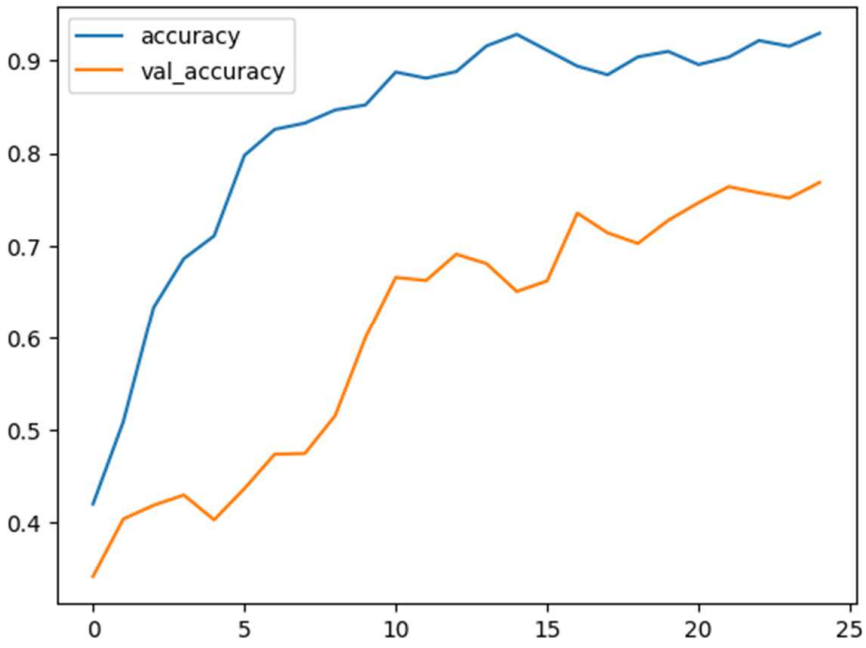

Figure 9. Training and validation accuracy versus epochs in happiness merged with excitement, anger, sadness, and neutral classes of IEMOCAP.

4.3.2. MELD Dataset

For the MELD dataset, Figure 10 and Table 7 show its confusion matrix and results. Neutral emotion has the best value of precision of 92%. Recall sadness has the best value of 92%. Otherwise, the F1-score for neutral has the best value of 74%, and the model has an accuracy of 63.33%. Figure 11 illustrates the training accuracy versus epochs on the MELD dataset, where the validation accuracy continues to rise with the training epochs.

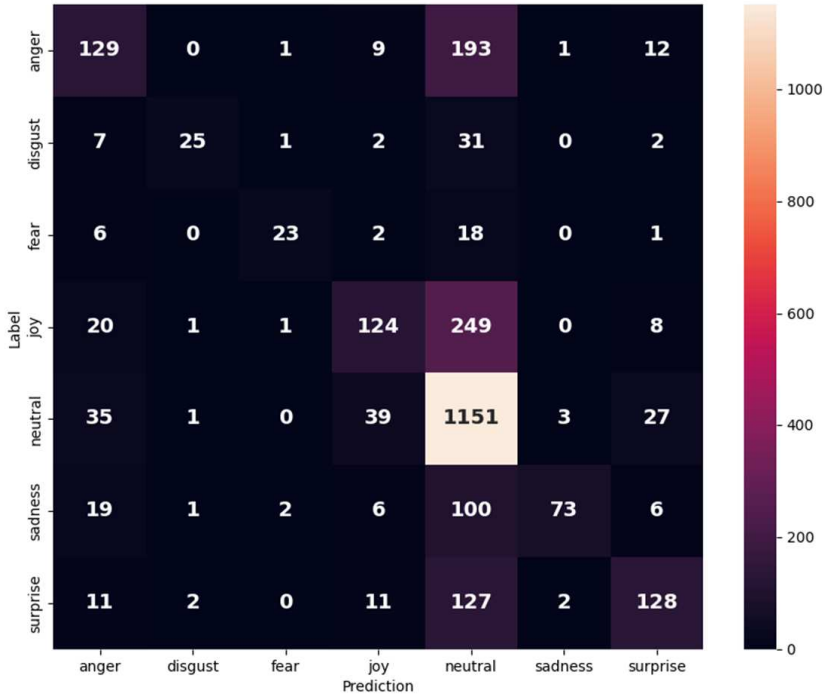

Figure 10. Confusion matrix of the MELD dataset.

Table 7. Classification report results for the MELD on the text and audio modalities in the function of precision, recall, and F1-score and accuracy of the model.

Classes	Precision	Recall	F1-score	Accuracy
Anger	0.37	0.57	0.45	
Disgust	0.37	0.83	0.51	
Fear	0.46	0.82	0.59	
Joy	0.31	0.64	0.42	0.6333
Neutral	0.92	0.62	0.74	
Sadness	0.35	0.92	0.51	
Surprise	0.46	0.70	0.55	

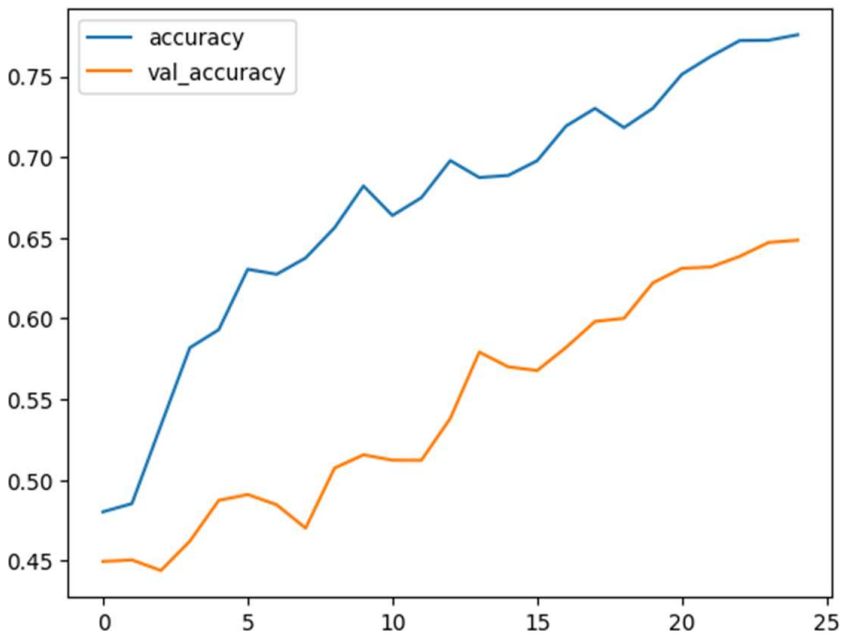

Figure 11. Training and validation accuracy versus epochs in MELD.

4.3.3. Tweeteval Dataset

For the Tweeteval dataset, Figure 12 shows the confusion matrix for each task in the dataset. (a) Refers to emotion which includes four classes (anger, joy, optimism, and sadness); (b) hate which includes two classes (non-hate and hate); (c) irony also has two classes (non-irony and irony); (d) offensive includes two classes (non-offensive and offensive); (e) sentiment which includes three classes (negative, neutral, and positive); (f) stance also has three classes (none, against, and favor); and (g) emoji which includes twenty classes (red heart, smiling face with heart eyes, face with tears of joy, two hearts, fire, smiling face with smiling eyes, smiling face with sunglasses, sparkles, blue heart, face blowing kiss, camera, united states, sun, purple heart, winking face, hundred points, beaming face with smiling eyes, Christmas tree, camera with flash, and winking face with tongue).

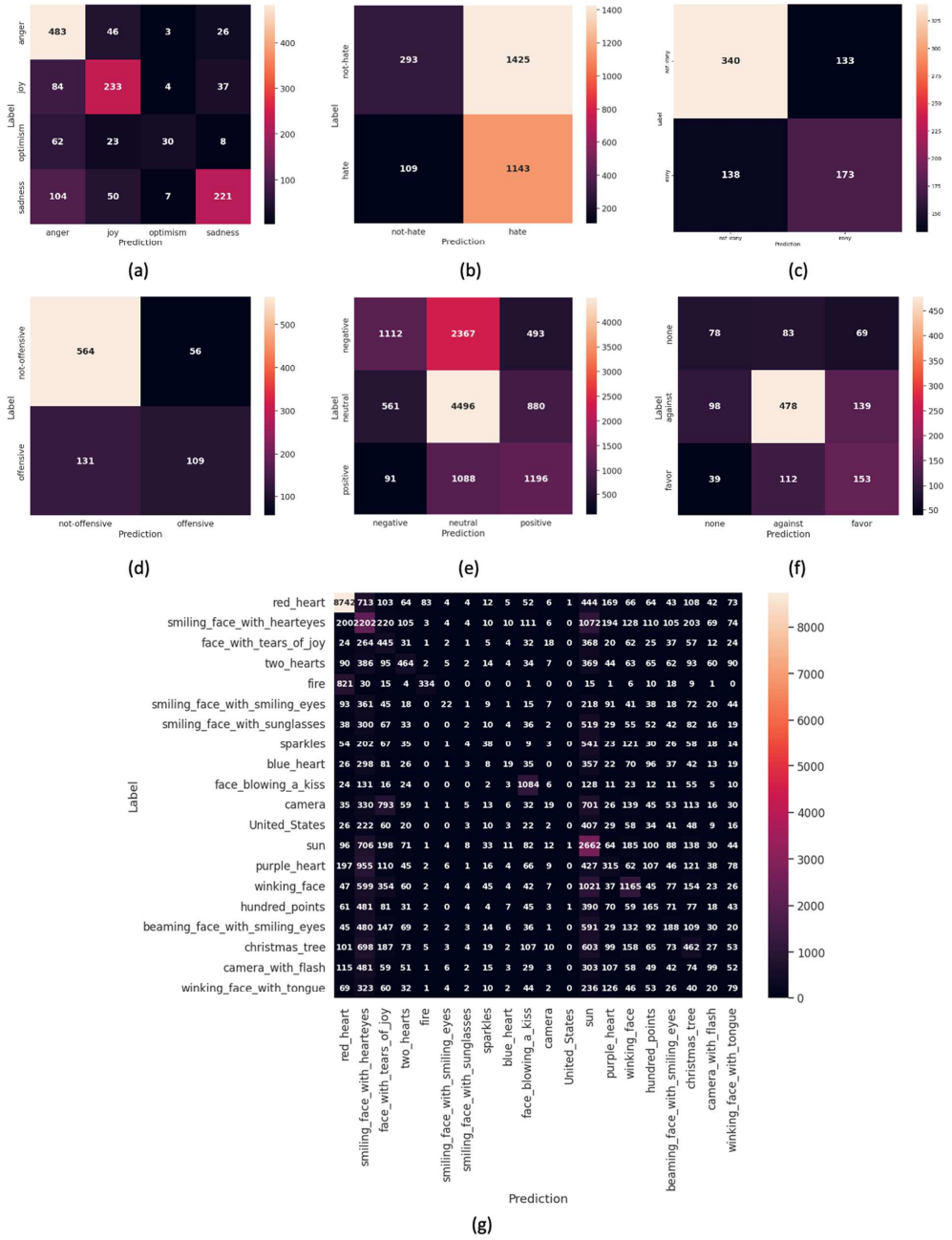

Figure 12. Confusion matrix for (**a**) emotion, (**b**) hate, (**c**) irony, (**d**) offensive, (**e**) sentiment, (**f**) stance, and (**g**) emoji in the Tweeteval dataset.

Table 8 shows the classification report for each task in the Tweeteval dataset. The offensive has the best average precision of 68%, 74% for average recall, and 70% for average F1 score. For accuracy, offensive task also has the best value of 78.26%. All results depend on text-only modality due to the data available in the Tweeteval dataset.

Table 8. Classification report results for the Tweeteval on the text and audio modalities in the function of precision, recall, and F1-score and accuracy of the model.

Task	Average Precision	Average Recall	Average F1 Score	Accuracy
Emotion	0.59	0.69	0.61	0.6806
Hate	0.54	0.59	0.44	0.4835
Irony	0.64	0.64	0.64	0.6543
Offensive	0.68	0.74	0.7	0.7826
Sentiment	0.51	0.56	0.51	0.5539
Stance	0.5	0.5	0.5	0.5677
Emoji	0.22	0.27	0.21	0.3701

In Figure 13, we visualize the training accuracy versus epochs of the Tweeteval dataset in each task. (a) Refers to emotion, (b) to hate, (c) to irony, (d) to offensive, (e) to sentiment, (f) to stance, and (g) to emoji.

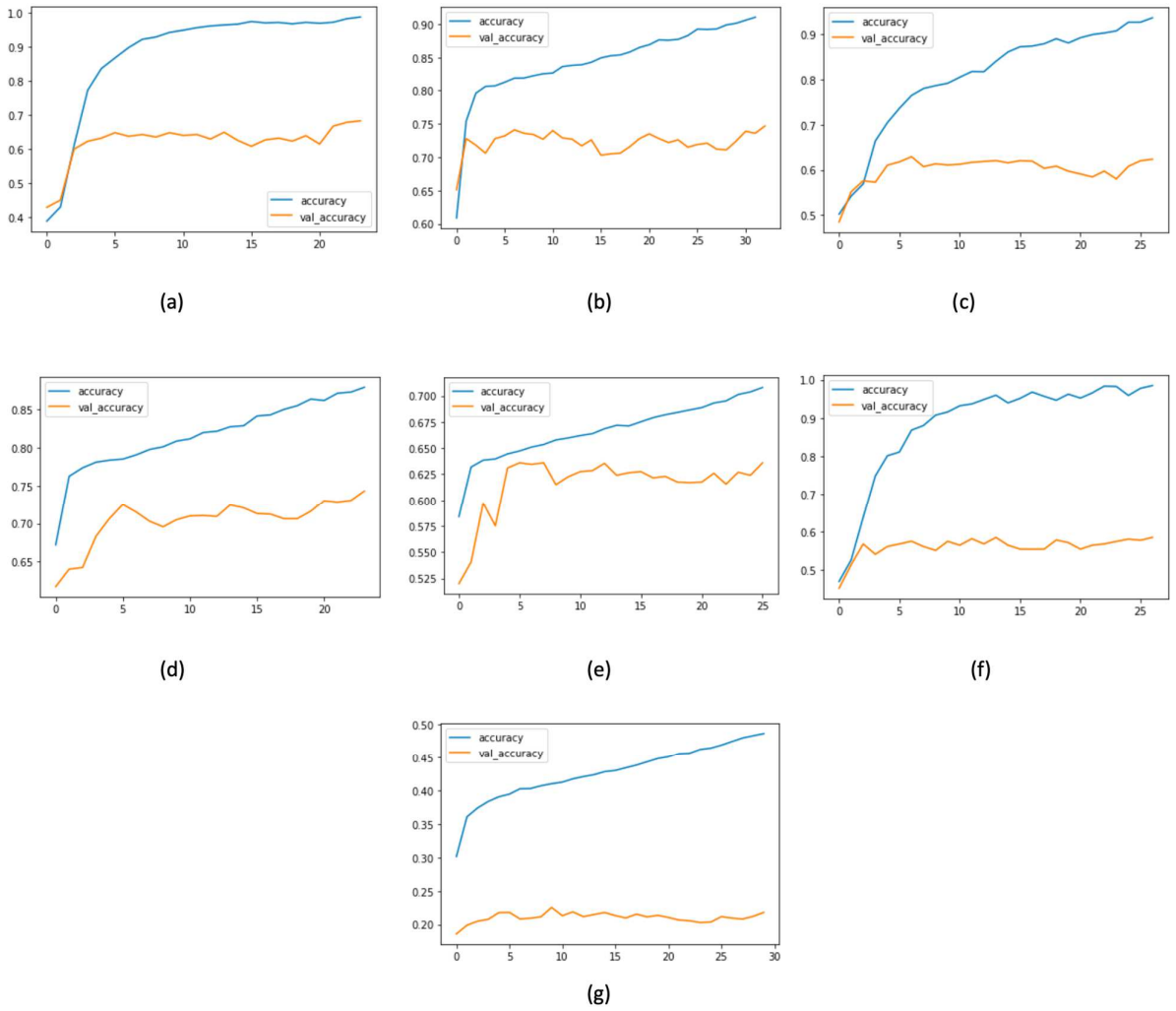

Figure 13. Training and validation accuracy versus epochs in (**a**) emotion, (**b**) hate, (**c**) irony, (**d**) offensive, (**e**) sentiment, (**f**) stance, and (**g**) emoji tasks in the Tweeteval dataset.

4.4. Comparison with State-of-the-Art Methods

We conducted an insightful investigation into the effectiveness of different feature representations within our model. We encountered a challenge in this regard due to the necessity of converting audio data into an image format suitable for training with the MobileNet model. This conversion was essential to align with the model architecture's requirements. As a result, we were unable to freeze the entire feature extraction process as it was integral to this conversion. As such, we examined the Mel spectrogram feature in isolation and compared it to the chromagram feature, both of which are commonly used representations for audio analysis. And presented in state-of-the-art comparison tables as Ours (Mel spectrogram) and Ours (chromagram). Our experimentation unveiled interesting findings. While both representations yielded promising results, the Mel spectrogram exhibited a slight advantage in terms of performance, signifying its greater discriminatory power in capturing essential acoustic characteristics. Intriguingly, when we merged these two features within our model, the outcomes were notably enhanced. This merging harnessed the complementary strengths of the Mel spectrogram and chromagram, resulting in a richer and more detailed feature set for training. Our findings highlight the significance of feature selection in optimizing model performance for audio-based tasks, emphasizing the benefits of leveraging a combination of feature representations to enhance the discriminative capacity of our model.

To further verify the competitiveness of MM-EMOR, it was compared to state-of-the-art methods. Our approach achieved higher accuracy, as shown in Table 9, for anger, happiness, neutral, and sadness classes in the IEMOCAP dataset. The increase in accuracy was up to 7%. Table 10 also shows that our approach achieved higher accuracy, up to 7%, for anger, happiness merged with excitement, neutral, and sadness classes. The consistent performance of MM-EMOR across different scenarios proves the strength and robustness of our multimodal emotion recognition system.

Table 9. Comparison of our proposed approach with state-of-the-art (anger, happiness, neutral, and sadness) classes on the IEMOCAP dataset.

Methods	UA	WA
H.Xu et al. [12]	0.709	0.725
Guo et al. [10]	0.725	0.719
Zaidi et al. [20]	0.756	-
Ours (Chromagram)	0.761	0.769
Ours (Mel spectrogram)	0.766	0.775
Huddar et al. [14]	0.766	-
Wang et al. [37]	**0.77**	0.765
Ours	**0.771**	**0.779**

Table 10. Comparison of our proposed approach with state-of-the-art (anger, happiness merged with excitement, neutral, and sadness) classes on the IEMOCAP dataset.

Methods	UA	WA
S. Sahoo et al. [38]	0.687	-
H. Feng et al. [39]	0.697	0.686
S. Tripathi et al. [13]	0.697	-
Ours (Chromagram)	0.744	0.753
Kumar et al. [17]	0.750	0.717
Ours (Mel spectrogram)	0.759	0.767
Setyono et al. [40]	**0.76**	0.76
Ours	**0.762**	**0.771**

We also compared our proposed approach with other state-of-the-art methods for the MELD dataset in Table 11. MM-EMOR offered higher accuracy with a difference ranging from 0.1% to 8.5%.

Table 11. Comparison of our proposed approach with state-of-the-art MELD dataset.

Methods	Accuracy
Wang et al. [19]	0.481
Guo et al. [10]	0.548
Wang et al. [41]	0.558
Ours (Chromagram)	0.588
Ours (Mel spectrogram)	0.592
Lian et al. [42]	0.62
Ho et al. [43]	0.632
Ours	**0.633**

For the Tweeteval dataset, we compared our model with other state-of-the-art methods. The results in Table 12 show that our approach achieved higher accuracy in emotion, emoji, and offensive tasks. The best UA improvement was obtained with the emoji task with an outstanding increase of 17.94%, followed by the offensive task scoring a 4.5% improvement. A similar performance was attained for the irony task. However, a performance gap exists for the hate, sentiment, and stance tasks. To this end, further tuning needs to be applied to enhance the performance.

Table 12. Comparison of our proposed approach with state-of-the-art Tweeteval dataset.

Methods	Emotion	Hate	Irony	Offensive	Sentiment	Stance	Emoji
Li et al. [1]	0.6770	**0.5950**	**0.6667**	0.7371	**0.6143**	**0.6756**	0.1907
Ours	**0.6806**	0.4835	0.6543	**0.7826**	0.5539	0.5677	**0.3701**

The state-of-the-art comparison shows that the IEMOCAP dataset achieves great results compared to other models, and Tweeteval too. The MELD dataset also achieved great performance compared to others but, in general, it achieved low accuracy because of the data collected from TV-shows which includes audience laughing and cinematic processing, so the data is not clear. In the case of Tweeteval, the observed accuracy is slightly lower, which can be attributed to the nature of social media data, known for its prevalence of human errors and incorrect text. This observation can be attributed to the inherent characteristics of the dataset at hand, which is exclusively comprised of text data and thus qualifies as an unimodal dataset. Unimodal datasets inherently possess a certain degree of simplicity in comparison to their multimodal counterparts, where the fusion of diverse data modalities can introduce added complexity.

A noteworthy finding concerns the length of the training process across the datasets. The training and feature extraction duration on the IEMOCAP dataset was 242 min, indicating the comprehensive nature of the model's learning process. On the MELD dataset, a similar pattern emerged, with the training phase lasting 325 min. Notably, the Tweeteval dataset deviated from this pattern, necessitating a 49-min training time. Furthermore, the execution time required to process a single instance is an important metric of efficiency. This metric was low at 19.9, 16.9, and 0.51 milliseconds for IEMOCAP, MELD, and Tweeteval, respectively, thus highlighting the computational efficiency that characterizes the MM-EMOR system's real-time functionality. To train our model, we used the capabilities of cloud computing infrastructure. The hardware requirements used in this cloud-based environment were Nvidia V100 GPU, which is well-known for its performance in deep learning applications, to expedite model training, and 52 gigabytes of RAM, which provided adequate memory resources to meet the complex needs of our training datasets. We used an 8-core CPU to support these computational operations.

5. Conclusions

In conclusion, the MM-EMOR system emerges as a promising forerunner of advanced emotion recognition capabilities in the context of multimodal data processing. The inge-

nious concatenation of preprocessed audio data, harnessed via the Mobilenet Convolutional Neural Network, with textual insights gleaned from the Roberta model is a key component of this system. This approach has been thoroughly examined and validated across three distinct benchmark datasets, each representing a distinct aspect of the emotion recognition landscape. The interactive emotional dyadic motion capture (IEMOCAP) dataset, the multimodal emotion lines dataset (MELD), and the Tweeteval dataset are among these. Notably, the last dataset only contains textual modality but includes seven experimental scenarios, showing the MM-EMOR system's versatility. MM-EMOR consistently outperformed its state-of-the-art counterparts by a significant margin across this broad spectrum of evaluation. In the case of the IEMOCAP dataset, MM-EMOR achieves improved accuracy, with gains ranging from 0.1% to an impressive 7%. Similarly, the MELD dataset sees an increase in accuracy ranging from 0.1% to 8.5%. The most impressive progress is seen in the Tweeteval dataset, where MM-EMOR achieves an astonishing 18% increase in accuracy. These findings highlight the approach's profound potential in the domain of social media analysis, implying that it should be expanded to include facial gesture recognition within videos, a path that promises to usher in a more comprehensive understanding of human emotions.

Author Contributions: Formal analysis, O.A. and K.M.F.; conceptualization, O.A. and K.M.F.; methodology, O.A., K.M.F. and A.A.E.; software, O.A.; validation, O.A.; investigation, O.A.; resources, O.A.; data curation, O.A. and K.M.F.; writing—original draft preparation, O.A.; writing—review and editing, K.M.F. and A.A.E.; visualization, O.A., K.M.F. and A.A.E.; supervision, K.M.F. and A.A.E.; project administration, K.M.F. and A.A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We used the IEMOCAP dataset available at: https://www.kaggle.com/datasets/jamaliasultanajisha/iemocap-full accessed on 23 September 2022, MELD dataset available at: https://affective-meld.github.io/ accessed on 26 September 2022, and Tweeteval available at: https://github.com/cardiffnlp/tweeteval accessed on 8 December 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, J.; Mishra, S.; El-Kishky, A.; Mehta, S.; Kulkarni, V. NTULM: Enriching social media text representations with non-textual units. *arXiv* **2022**, arXiv:2210.16586.
2. Pablos, S.M.; García-Bermejo, J.G.; Zalama Casanova, E.; López, J. Dynamic facial emotion recognition oriented to HCI applications. *Interact. Comput.* **2015**, *27*, 99–119. [CrossRef]
3. Makiuchi, M.R.; Uto, K.; Shinoda, K. Multimodal emotion recognition with high-level speech and text features. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; IEEE: Piscataway, NJ, USA; pp. 350–357.
4. Kandali, A.B.; Routray, A.; Basu, T.K. Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In Proceedings of the TENCON 2008—2008 IEEE Region 10 Conference, Hyderabad, India, 19–21 November 2008; IEEE: Piscataway, NJ, USA; pp. 1–5.
5. Nwe, T.L.; Foo, S.W.; De Silva, L.C. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623. [CrossRef]
6. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA; pp. 6645–6649.
7. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
8. Breuel, T.M. High performance text recognition using a hybrid convolutional-lstm implementation. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 1, pp. 11–16.
9. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep structured output learning for unconstrained text recognition. *arXiv* **2014**, arXiv:1412.5903.
10. Guo, L.; Wang, L.; Dang, J.; Fu, Y.; Liu, J.; Ding, S. Emotion Recognition with Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information. *IEEE MultiMedia* **2022**, *29*, 94–103. [CrossRef]

11. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394. [CrossRef]
12. Xu, H.; Zhang, H.; Han, K.; Wang, Y.; Peng, Y.; Li, X. Learning alignment for multimodal emotion recognition from speech. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 15–19 September 2019; pp. 3569–3573.
13. Tripathi, S.; Tripathi, S.; Beigi, H. Multimodal Emotion Recognition on IEMOCAP Dataset using Deep Learning. *arXiv* **2018**, arXiv:1804.05788.
14. Huddar, M.G.; Sannakki, S.S.; Rajpurohit, V.S. Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. *Multimed. Tools Appl.* **2021**, *80*, 13059–13076. [CrossRef]
15. Eyben, F.; Wollmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
16. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *J. Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
17. Kumar, P.; Kaushik, V.; Raman, B. Towards the Explainability of Multimodal Speech Emotion Recognition. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 1748–1752. [CrossRef]
18. Singh, P.; Srivastava, R.; Rana, K.P.S.; Kumar, V. A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowl. Based Syst.* **2021**, *229*, 107316. [CrossRef]
19. Wang, Y.; Gu, Y.; Yin, Y.; Han, Y.; Zhang, H.; Wang, S.; Li, C.; Quan, D. Multimodal transformer augmented fusion for speech emotion recognition. *Front. Neurorobotics* **2023**, *17*, 1181598. [CrossRef] [PubMed]
20. Zaidi, S.A.M.; Latif, S.; Qadi, J. Cross-Language Speech Emotion Recognition Using Multimodal Dual Attention Transformers. *arXiv* **2023**, arXiv:2306.13804.
21. Canal, F.Z.; Müller, T.R.; Matias, J.C.; Scotton, G.G.; de Sa Junior, A.R.; Pozzebon, E.; Sobieranski, A.C. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Inf. Sci.* **2022**, *582*, 593–617. [CrossRef]
22. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: A survey. *Multimed. Syst.* **2010**, *16*, 345–379. [CrossRef]
23. Huang, J.; Li, Y.; Tao, J.; Lian, Z.; Wen, Z.; Yang, M.; Yi, J. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 2017; pp. 11–18.
24. Stappen, L.; Baird, A.; Christ, L.; Schumann, L.; Sertolli, B.; Messner, E.M.; Cambria, E.; Zhao, G.; Schuller, B.W. The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, Virtual Event, 24 October 2021; pp. 5–14.
25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
27. Venkataramanan, K.; Rajamohan, H.R. Emotion recognition from speech. *arXiv* **2019**, arXiv:1912.10458.
28. Hao, M.; Cao, W.H.; Liu, Z.T.; Wu, M.; Xiao, P. Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. *Neurocomputing* **2020**, *391*, 42–51. [CrossRef]
29. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; pp. 18–25.
30. Solovyev, R.A.; Vakhrushev, M.; Radionov, A.; Romanova, I.I.; Amerikanov, A.A.; Aliev, V.; Shvets, A.A. Deep learning approaches for understanding simple speech commands. In Proceedings of the 2020 IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), Kyiv, Ukraine, 22–24 April 2020; IEEE: Piscataway, NJ, USA; pp. 688–693.
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
32. Sifre, L. Rigid-Motion Scattering for Image Classification. Ph.D. Thesis, CMAP Ecole Polytechnique, Palaiseau, France, 2014.
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
34. Barbieri, F.; Camacho-Collados, J.; Neves, L.; Espinosa-Anke, L. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv* **2020**, arXiv:2010.12421.
35. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
36. Plutchik, R. The Nature of Emotions. *J. Storage (JSTOR) Digit. Libr. Am. Sci. J.* **2001**, *89*, 344–350.
37. Wang, Y.; Shen, G.; Xu, Y.; Li, J.; Zhao, Z. Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 4518–4522. [CrossRef]
38. Sahoo, S.; Kumar, P.; Raman, B.; Roy, P.P. A Segment Level Approach to Speech Emotion Recognition using Transfer Learning. In Proceedings of the 5th Asian Conference on Pattern Recognition (ACPR), Auckland, New Zealand, 26–29 November 2019; pp. 435–448.
39. Feng, H.; Ueno, S.; Kawahara, T. End-to-end Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 501–505. [CrossRef]

40. Setyono, J.C.; Zahra, A. Data augmentation and enhancement for multimodal speech emotion recognition. *Bull. Electr. Eng. Inform.* **2023**, *12*, 3008–3015. [CrossRef]
41. Wang, N.; Cao, H.; Zhao, J.; Chen, R.; Yan, D.; Zhang, J. M2R2: Missing-Modality Robust emotion Recognition framework with iterative data augmentation. *IEEE Trans. Artif. Intell.* **2022**, *4*, 1305–1316. [CrossRef]
42. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 985–1000. [CrossRef]
43. Ho, N.H.; Yang, H.J.; Kim, S.H.; Lee, G. Multimodal approach of speech emotion recognition using multi-level multihead fusion attention-based recurrent neural network. *IEEE Access* **2020**, *8*, 61672–61686. [CrossRef]