BEE-MER: BIMODAL EMBEDDINGS ENSEMBLE FOR MUSIC EMOTION RECOGNITION

Pedro L. LOURO (pedrolouro@dei.uc.pt) (0000-0003-3201-6990) [1, *], **Tiago F. RIBEIRO** (tiago.f.ribeiro@dei.uc.pt) (0000-0003-1603-1218) [1, *], **Ricardo MALHEIRO** (rsmal@dei.uc.pt) (0000-0002-3010-2732) [1, 2], **Renato PANDA** (panda@dei.uc.pt) (0000-0003-2539-5590) [1, 3], and **Rui Pedro PAIVA** (ruipedro@dei.uc.pt) (0000-0003-3215-3960) [1]

[1] **University of Coimbra, CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra** , *Department of Informatics Engineering*, Coimbra, Portugal
[2] **Polytechnic Institute of Leiria School of Technology and Management**, Leiria, Portugal
[3] **Ci2 — Smart Cities Research Center, Polytechnic Institute of Tomar**, Tomar, Portugal
[*] These authors contributed equally to this work.

ABSTRACT

Static music emotion recognition systems typically focus on audio for classification, although some research has explored the potential of analyzing lyrics as well. Both approaches face challenges when it comes to accurately discerning emotions that have similar energy but differing valence, and vice versa, depending on the modality used. Previous studies have introduced bimodal audio-lyrics systems that outperform single-modality solutions by combining information from standalone systems and conducting joint classification. In this study, we propose and compare two bimodal approaches: one strictly based on embedding models (audio and word embeddings) and another one following a standard spectrogram-based deep learning method for the audio part. Additionally, we explore various information fusion strategies to leverage both modalities effectively. The main conclusions of this work are the following: i) the two approaches show comparable overall classification performance; ii) the embedding-only approach leads to a higher confusion between quadrants 3 and 4 of Russell's circumplex model; iii) and this approach requires significantly less computational cost for training. We discuss the insights gained from the approaches we experimented with and highlight promising avenues for future research.

1. INTRODUCTION

The most tackled problem in Music Emotion Recognition (MER) is single-label static emotion recognition, where the goal is to identify the predominant emotion in a song. Over the years, mostly audio-based systems have been proposed, ranging from feature engineering efforts to deep learning (DL) approaches. Beyond the known problems related to annotation protocols, dataset sizes, and lack of standardization [1], these audio-based approaches struggle to differentiate emotions with similar energy but different valence. On the other hand, lyrics-based MER (LMER) systems, although attaining relative success, face the opposite challenge, i.e., successfully differentiating valence while having difficulties with arousal [2].

Audio-lyrics bimodal systems, herein referred to only as bimodal systems, emerge as a natural solution. Previous work [3] has shown that simply fusing the information from feature learning architectures outperforms single-modality approaches, either based on audio or lyrics.

Considering the above, we defined the following research questions to guide our experimental process:

- RQ1: Are audio embeddings relevant for audio MER, in comparison with standard spectrogram-based deep learning approaches?
- RQ2: Does fine-tuning embedding models improve MER systems?
- RQ3: What is the best information fusion strategy for bimodal MER?

In this work, we tackle bimodal MER under the framework of James Russell's circumplex plane [4] (presented below), namely for classification into four emotion quadrants.

For evaluation purposes, we employed the recently proposed MERGE dataset [2].

We propose and compare two bimodal approaches. The first approach is based solely on embedding models, including audio and word embeddings, which is the primary focus of this article. The second approach uses a standard DL method that relies on spectrograms for the audio component. We also examine different information fusion strategies to effectively combine both modalities.

The key conclusions from our work are as follows: i) both approaches demonstrate similar overall classification performance; ii) the embedding-only approach tends to create more confusion between quadrants 3 and 4 of Russell's circumplex model; and iii) this approach significantly reduces the computational cost required for training.

We also provide some insights into the various embeddings and discuss the strengths of the various fusion strategies experimented with.

The document is structured into six sections. The present section introduces the context and the problem we propose to tackle, presenting the various approaches considered. Section 2 discusses some related work relevant to the present study. Each experimented methodology is discussed in detail in Section 3. In Section 4, evaluation details are presented, including optimization protocols and evaluation metrics. The obtained results are presented and discussed in Section 5. Section 6 concludes the present work and discusses possible future work from the insights gathered.

2. RELATED WORK

In this section, we briefly describe some of the approaches proposed for static MER over the years. We present standalone audio and lyrics static MER systems, followed by the audio-lyrics bimodal methodologies.

2.1 Emotion Models

Psychology researchers have long studied emotions, resulting in various taxonomies categorized into categorical (or discrete) models and dimensional models. Categorical models use distinct categories or descriptors, while dimensional models organize emotions along axes, as seen in Russell's circumplex model. The latter, which is the most widely used in MER and the target of the present work, proposes that emotional states stem from two neurophysiological systems: one for valence (pleasure vs. displeasure) and another for arousal (energy level). The arousal-valence (AV) plane defines four emotion quadrants: positive valence - high arousal (Q1: happy), negative valence - high arousal (Q2: tense), negative valence - low arousal (Q3: sad), and positive valence - low arousal (Q4: relaxed).

2.2 Unimodal Audio MER

Most of the studies dealing with MER are on static emotion recognition based on audio. Early approaches dealt with a small set of audio features known to be correlated with emotional states, either tackling classification [5] or regression [6]. More recently, Panda et al. [7] proposed a new set of audio features mostly related to underrepresented musical dimensions, such as articulation and musical texture, that achieved state-of-the-art results.

Beyond classical approaches (based on feature engineering and classical Machine Learning methodologies), Deep Learning (DL) saw increased interest after the multi-tag classification approach by Choi et al. [8]. Various other approaches built on this foundation, such as proposing music-theory-driven filters for the feature learning portion [9] or replacing it with representation learning directly from the audio waveform [10].

Audio embeddings were already proposed to classify emotion in music, as proposed by Koh et al. [11]. The approach employed the OpenL3 embedding model , trained on environmental sounds, and evaluated it on a dataset developed by our team , reporting a 72% F1-score. However, our efforts to replicate these results fell short, achieving a 55.70% F1-score [12]. Our intuition is that the data used to train this model was not well-suited for music-related tasks since all training datasets consisted of environmental sounds.

2.3 Unimodal Lyrics MER

As in the audio domain, early works in LMER relied on manual feature extraction techniques such as Bag-of-Words, Term Frequency-Inverse Document Frequency, and topic modeling (e.g., Latent Dirichlet Allocation) [13]. These methods, though computationally efficient, provided only a shallow representation of lyrics, thereby limiting their ability to capture the complex semantic and contextual nuances essential for accurate emotion representation.

Advancements in deep learning prompted a shift toward data-driven feature learning. Notably, static word embedding models such as Word2Vec [14] and GloVe [15] emerged, which generate fixed vector representations of words based on co-occurrences or context within a text corpus. These models facilitated a more refined encoding of semantic relationships, although each word is assigned a single representation that remains independent of the context in which it appears. In parallel, recurrent neural network architectures, particularly Long Short-Term Memory (LSTM) models, were introduced to model the sequential dependencies inherent in lyrical text, but the restricted context windows inherent to recurrent models continued to limit the possibility of capturing emotions across broader contexts.

More recently, transformer-based models have emerged as state-of-the-art in the LMER domain [16]. These models produce contextual embeddings where each word's representation dynamically adjusts based on its surrounding sentence context. Models such as BERTand RoBERTaemploy self-attention and pre-training mechanisms on large text corpora to capture long-range emotional contexts. Variants such as XLNet [17] have overcome limitations related to sequence length, enabling the processing of longer song lyrics and achieving significant results in lyrics-based MER [18].

2.4 Bimodal Audio-Lyrics MER

In this work, bimodal MER refers to systems that incorporate both audio and lyrics information to predict the predominant emotion in a song. Considering that audio is known to more accurately predict a song's arousal and lyrics a song's valence, it is natural to exploit together both modalities' capabilities [19].

Few audio-lyrics systems are found in the literature. Although some incorporate classical audio and lyrics' feature engineering techniques to perform classification [20], more recent DL-based approaches have achieved more interesting results [21].

The work from Delbouys et al. [22] thoroughly compares different fusion approaches with the best-performing unimodal model approaches, namely mid-level fusion, which concatenates the learned embeddings from both modalities and late fusion, which creates a voting ensemble from the above-mentioned models. We further explore this approach with more recent data representations.

3. MATERIALS AND METHODS

This section discusses the proposed methodologies. We begin by introducing the dataset employed in this study. Then, the proposed models are presented. This section concludes with details regarding the experimented information fusion strategies.

3.1 Datasets

MERGE [2] is a collection of datasets for static audio, lyrics, and bimodal MER research. Each modality provides a complete and balanced variant. For the purposes of this study, we only describe the bimodal datasets.

The MERGE Bimodal Complete dataset, herein referred to as MERGE-BC, contains a total of 2216 pairs of audio-lyrics samples annotated according to the four quadrants of Russell's Circumplex Model of emotion [4]. The Balanced variant, referred to as MERGE-BB, contains a total of 2000 pairs evenly distributed between each quadrant of the above-mentioned model, as shown in Table 1.

Table 1: Total audio-lyrics pairs in MERGE.

Dataset	Q_1	Q_2	Q_3	Q_4	Total
MERGE-BC	525	673	500	518	2216
MERGE-BB	500	500	500	500	2000

Alongside these samples, metadata for each sample pair is provided, from the AllMusic platform [23], as well as the aforementioned emotion categorical annotations. Each dataset also provides two train-validate-test splits, following a 70-15-15 and a 40-30-30 strategy. We employ the former to conduct the validation experiments on the proposed methodologies.

3.2 Embeddings-Only Model (EOM)

In this section, we describe the embeddings-only model (hereafter termed EOM), a proposed model based on audio and word embeddings, and discuss possible fusion strategies (early and late fusion).

3.2.1 Audio Embeddings

Due to the improvements observed by the introduction of word embeddings for lyrics emotion recognition, we explored the possibility of employing audio embeddings for bimodal classification. In the following paragraphs, we first present the main arguments for adopting audio embeddings, followed by a succinct explanation of the experimented embedding models: wav2vec2 [24] and MERT [25].

From our previous work, we have theorized that the Mel-spectrogram representation does not fully capture the same information that can be extracted using feature engineering approaches, such as the work from Panda et al. [7], explaining the consistently lower results from classical approaches.

Beyond this, the amount of data available severely limits the capabilities of DL models, a common problem in the MER field. Large general audio models pre-trained on several hundred hours of data may extract more relevant information useful for a variety of downstream tasks, circumventing the lack of domain-specific data.

Finally, the use of embeddings is particularly interesting due to the relatively lower computational cost of training a model for classification compared to models comprised of several convolutional layers.

wav2vec2

Starting with wav2vec2, this is a family of pre-trained models developed for speech recognition tasks. The available large version of the model comprises 960 hours of 16kHz speech data, which we adopt for this study. Given the task for which it was trained, we expect that these embeddings will extract information mostly related to acoustics and timbre.

The model pipeline can be described as follows. A multi-layer convolutional feature encoder first receives the raw audio signal, outputting audio features by passing the signal through a set of seven temporal convolutional blocks in 5-second increments with overlap. The resulting features are passed through a Transformer-like architecture, referred to as a context network, that uses dynamic convolutions to act as relative positional embeddings.

For our use case, the embeddings were extracted using the available 'Large' model on the wav2vec2's HuggingFace repository [26]. The extracted vector contains a set of 1028 values for each timestep of the 25 hidden layers of the context network. Since our task deals with static emotion, the information pertaining to timesteps is averaged, resulting in a more concise 1028 values for each of the 25 above-mentioned layers.

MERT

The acoustic Music undERstanding model with large-scale self-supervised Training (MERT) leverages the knowledge of two teacher models for acoustic- and musical-informed representation learning. We study the application of the v1-330M version, pre-trained on 160k hours of unlabeled music mined from the Internet. Beyond the information related to acoustics and timbre, the musical teacher is expected to condition the representation to include pitch and harmony information.

The backbone of MERT is the one-dimension convolutional network that encodes the raw audio signal, sampled at 24kHz, similarly to the feature encoder of wav2vec2, and the Transformer encoder, based on the HuBERT architecture, to obtain contextual representations from the encoded features.

In this study, we employ the available MERT-v1-330M model from its HuggingFace repository [27]. We follow the timestep averaging described for the wav2vec2 embeddings.

Classification Procedure

For both audio embedding model's outputs, we perform classification using the dense network described in Section 3.4. To this end, we further reduce the dimensionality of these embeddings (a 25x1024 matrix) using a simple one-dimensional convolutional layer (outputting a 1x1024 vec-

tor). This compresses the embeddings into a single vector with the most relevant information for our problem. Details regarding the optimization process can be found in Section 4.1.

3.2.2 Word Embeddings

This section describes the experimental setup and methodology for drawing on word embeddings derived from two Transformer-based models: RoBERTa and ModernBERT. Their encoder-only design ensures computational efficiency without compromising representational quality, making them well-suited for resource-constrained environments.

The availability of pre-trained models, built on vast text corpora, provides a robust starting point for downstream tasks with minimal data, such as lyrics datasets in MER.

RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) refines the original BERT training strategy by eliminating the Next Sentence Prediction objective, incorporating dynamic masking, and scaling up training on extensive corpora.

For this study, we utilize the pretrained `RoBERTa large` configuration available from HuggingFace [28]. This model comprises 24 layers, 16 attention heads per layer, and a hidden size of 1024. Pre-trained on diverse datasets including BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories (aggregating approximately 160 GB of text), RoBERTa has demonstrated robust performance in linguistic meaning representation tasks—including in emotion recognition scenarios [29].

ModernBERT

ModernBERT represents a contemporary evolution of the BERT architecture, incorporating several advancements to improve both efficiency and performance. Key innovations include Rotary Positional Embeddings (RoPE) for enhanced positional encoding, GeGLU activation layers, a streamlined architecture with reduced bias terms, and an extra normalization layer for stabilized training. These improvements enable ModernBERT to support extended sequence lengths—up to 8,192 tokens—thus accommodating lengthy song lyrics without truncation.

In this work, we employ the pre-trained `ModernBERT large` model from HuggingFace [30].

Classification Procedure

The generated text embeddings were classified using a Support Vector Machine (SVM) with an RBF kernel. Prior to classification, the embeddings were standardized to ensure consistent feature scaling. The final SVM model was trained on the combined training and validation sets and evaluated on the test set. Data splits were aligned with the fine-tuning process to maintain comparability. Optimization and fine-tuning details are provided in Section 4.1.

3.3 Embedding Fusion Strategies

We explore different strategies to fuse the extracted audio and lyrics embeddings for bimodal classification, namely, early fusion and late fusion. The rest of this section describes the process used for each strategy.

3.3.1 Early Fusion

Our first approach is to fuse the embeddings obtained from audio and lyrics as a single feature vector and use it as input to the dense network classifier from the ASWEM model. We expect the model to automatically learn the most relevant features from each modality to perform classification. All embeddings are considered as possible input pairs to this methodology.

In particular, all combinations of one audio and one word embedding are evaluated (e.g., MERT+RoBERTa, wav2vec2+ModernBERT, etc.).

3.3.2 Late Fusion

For the decision-level fusion, we consider two possible ensemble approaches: majority and soft voting.

Majority voting receives a set of predictions for a given sample and chooses the class with the majority of the votes. We consider 2 neural networks to obtain the prediction for wav2vec2 and MERT, while 2 different SVMs are used to obtain predictions from RoBERTa and ModernBERT. In the event of a tie, the ensemble final prediction is given by the first model, which in our implementation is the one using MERT as its input.

Soft voting differs from the previous approach by taking the probabilities from the models and performing a weighted average to obtain the final prediction. For our purposes, the weights of each model are directly proportional to the F1-score obtained during the standalone experiments.

3.4 Audio Spectrogram + Word Embedding Model (ASWEM)

We propose another approach based on the more or less standard procedure for audio modeling using deep learning approaches, i.e., employing an audio spectrogram as input to a deep neural network. Moreover, giving the superior results attained by the RoBERTa word embedding for lyrics classification (as discussed later on in Section 5), we use this embedding for the analysis of the lyrics counterpart. Hereafter, this audio spectrogram plus word embedding model is termed ASWEM.

As such, the architecture comprises three different modules: an audio feature extractor, a lyrics feature extractor, and a dense classifier.

The audio portion receives a Mel-spectrogram representation obtained from the raw audio signal. The most relevant patterns are extracted using a series of four two-dimensional convolutional blocks. The feature extraction architecture is adapted from Choi et al. [31].

The lyrics feature extractor operates in a similar fashion, receiving word embeddings obtained from the lyrics' text and extracting relevant information through a sequence of 4 one-dimensional convolutional blocks.

Each modality is trained using small dense networks. Therefore, we obtain an ASWEM audio-only model and an ASWEM lyrics-only model.

Dataset	Strategy	Embeddings	F1	Prec.	Rec.	Model Config.
COMPLETE	ASWEM	N.A.	79.21%	79.60%	79.22%	Batch Size=16, Optimizer=SGD, Learning Rate=$1e^{-2}$
	EOM: Early Fusion	wav2vec2-Large + RoBERTa Large	77.87%	79.30%	77.71%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
		wav2vec2-Large + ModernBERT Large	75.18%	75.56%	75.30%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
		MERT-v1-330M + RoBERTa Large	77.24%	78.20%	77.11%	Batch Size=128, Optimizer=Adam, Learning Rate=$4.28e^{-4}$
		MERT-v1-330M + ModernBERT Large	76.45%	76.82%	76.51%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
	EOM: Majority Voting	All	77.93%	78.07%	77.84%	N.A.
	EOM: Soft Voting	All	77.86%	78.89%	77.72%	N.A.
BALANCED	ASWEM	N.A.	78.41%	79.07%	78.33%	Batch Size=64, Optimizer=SGD, Learning Rate=$1e^{-2}$
	EOM: Early Fusion	wav2vec2-Large + RoBERTa Large	76.18%	77.91%	76.00%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
		wav2vec2-Large + ModernBERT Large	72.14%	72.95%	72.33%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
		MERT-v1-330M + RoBERTa Large	75.75%	77.98%	75.67%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
		MERT-v1-330M + ModernBERT Large	72.84%	73.70%	73.00%	Batch Size=128, Optimizer=Adam, Learning Rate=$4.28e^{-4}$
	EOM: Majority Voting	All	75.97%	76.10%	76.00%	N.A.
	EOM: Soft Voting	All	73.21%	73.27%	73.33%	N.A.

Notes: F1=Weighted F1-score, Prec.=Precision, Rec.=Recall, N.A.=Not Applicable.

Table 2: Bimodal Emotion Classification Results.

To train the bimodal model, these small dense networks (employed to train each modality separately) are removed. Then, the previous layers from each branch are frozen, and their corresponding outputs are concatenated. After concatenating the retrieved patterns, a dense network, comprised of alternating dense and dropout layers, performs classification.

This procedure reduces the computational resources necessary to obtain the final model and ensures rapid convergence to an optimal solution.

4. EVALUATION AND MODEL SELECTION

Below is the evaluation procedure followed for each methodology described in the previous section. The considered datasets are briefly described, followed by the model selection procedure, including the optimization strategy and evaluation metrics.

4.1 Model Selection and Fine-tuning

We briefly describe the optimization setup for the evaluated methodologies, as well as specific optimizations for each when necessary.

4.1.1 Fine Tuning of Word Embedding Models

We fine-tune pre-trained RoBERTa-large and ModernBERT-large, adapted to a four-class emotion task using tokenized lyrics (maximum lengths: 512 tokens for RoBERTa, 2048 for ModernBERT). Performance is monitored using macro F1-score on the validation set, with early stopping triggered after 5 epochs without improvement, saving the best model state.

Post-training, the classification head is replaced with an identity layer to generate embeddings from the full dataset, processed batch-wise.

4.1.2 Classification Models Optimization Protocol

We optimize Support Vector Machine classifiers through systematic hyperparameter tuning using Optuna. The process employs a search space for the critical parameters: (i) C (Regularization Param.): $[1e^{-3}, 1e^{3}]$ (log. scale), (ii) γ (RBF Kernel Coefficient): $[1e^{-4}, 1e^{1}]$ (log. scale).

Embeddings are standardized using the combined training and validation datasets. Hyperparameter optimization is performed by maximizing the macro F1-score via a Bayesian optimization approach, evaluated through 5-fold cross-validation over 50 trials.

The remaining approaches, which utilized neural networks as their foundation, employed validation accuracy maximization as their objective function. This difference is justified by the more stable training process.

Following the search spaces defined in our own work for tuning audio-only and bimodal models, search spaces for these approaches were set as shown below: (i) Batch Size: $\{16, 32, \ldots, 128\}$, (ii) Optimizers: [SGD, Adam], (iii) Learning Rate: $[10^{-4}, 10^{-2}]$.

4.2 Evaluation Metrics

The proposed models are evaluated using standard metrics tailored for multi-class emotion classification (classes $Q1$, $Q2$, $Q3$, $Q4$), namely: F1-score, precision, recall and confusion matrices.

5. RESULTS AND DISCUSSION

The obtained results for the present study are presented and discussed in this section. We begin by discussing the best attained results and the gathered insights from comparing the various approaches.

P/A	Q1	Q2	Q3	Q4
Q1	75.0%	7.14%	5.95%	11.90%
Q2	7.07%	91.92%	0.00%	1.01%
Q3	4.69%	0.00%	81.25%	14.06%
Q4	8.24%	1.18%	23.53%	67.06%

Table 3: Confusion Matrix for the ASWEM Model — MERGE-BC

P/A	Q1	Q2	Q3	Q4
Q1	84.29%	5.71%	5.71%	4.29%
Q2	7.59%	87.34%	2.53%	2.53%
Q3	2.86%	2.86%	68.57%	25.71%
Q4	9.88%	0.00%	25.93%	64.20%

Table 4: Confusion Matrix for the EOM - Majority Voting — MERGE-BC

The best results were attained with the ASWEM approach (79.21% in the MERGE-BC), closely followed by the EOM majority voting methodology (77.93%, also in the MERGE-BC).

Despite the decrease of around 1% compared to the ASWEM model, the prediction accuracy for Q1 and Q2 is considerably higher, around 9% and 4%, as depicted in Table 4. However, the ASWEM model performs better for low arousal quadrants, particularly regarding Q3 (81.25% against 68.57% for EOM - majority voting). So, despite the ensemble being less computationally complex to optimize, the tradeoff should be considered.

Regarding the experimented audio embeddings, both outperform the ASWEM model's audio portion with an increase of 10% F1-score, as seen in Table 5. This increase is due to the higher discerning power of both Q3 and Q4, which are known to be difficult to accurately predict when using audio only. Despite this improvement, confusion is still high in comparison to the other classes. Moreover, MERT did outperform wav2vec2 as would be expected, but with only slight improvements, particularly an increase of 1.5% and 3% F1-score on MERGE-BC and MERGE-BB, respectively.

We observed that the best-performing early fusion approaches employed RoBERTa as their word embedding, achieving around 78% F1-score when paired with wav2vec2 audio embeddings. The difference with the MERT and RoBERTa pairing is very small at around 0.6% on MERGE-BC, however, we would expect MERT to introduce more relevant information considering it was specially trained for music-related tasks. As presented in Table 2, similar results are observed in MERGE-BB with MERT slightly above the wav2vec2 pairing by around 0.4%. The biggest drawback of this approach is the drop in Q4 prediction accuracy compared with lyrics' unimodal approaches. This points to more influence from the audio embeddings on the final prediction, considering that the increase in predicting Q3 is very small compared to the above-mentioned drop.

As for late fusion, the best-attained results from the newly experimented methodologies were attained using the majority voting strategy, as mentioned above. The differences between majority and soft voting are more noticeable on the balanced set, where a 3% F1-score difference can be observed, as presented in Table 2. Further analyzing the quadrant-specific performance, it is interesting to note that soft voting better predicts Q1 on MERGE-BC, but majority voting outperforms when using MERGE-BB. It is also worth noting that Q4 also has lower prediction accuracy compared to the lyrics-only methodologies, following the early fusion approaches. Q3 also suffers a large drop between MERGE-BC and MERGE-BB on both soft, 77.8% to 68.2% F1-score, and majority voting, 74.7% to 68.6% F1-score.

Overall, bimodal approaches continue to outperform audio- or lyrics-only methodologies. The achieved results on the experimented fusion strategies point to the current audio embedding models as a promising alternative to the Mel-spectrogram-based models. However, the fusion strategies still need further refining, as information from both modalities may not be used to its fullest potential. Moreover, fine-tuning the audio embedding models may contribute to better overall performance.

6. CONCLUSION AND FUTURE WORK

Our study explored the application of audio and word embeddings, as well as spectrogram-based audio approaches, for static Music Emotion Recognition. To fully exploit the information provided by both audio and lyrics, two fusion strategies were explored: either concatenating the resulting embedded representations from the embedding models to a dense classifier or using an ensemble of expert models to obtain the final prediction.

Our best-attained results using majority voting did not outperform the ASWEM approach in terms of the overall score, but the reduced computational complexity and call for more research in this area. In addition, we observed that the EOM approach led to increased confusion between Q3 and Q4; on the other hand, the ASWEM led to higher confusion between Q1 and Q4. The experiments conducted on various combinations of early fusion strategies provide some insight into future lines of research, which are briefly discussed below.

The reported findings need to be further analyzed, considering the limited datasets considered in this study, which was an exploratory effort. For a fair comparison, it is also relevant to fine-tune the audio-embedding models employed in this study and retrain the ASWEM model with the tuned word embeddings.

Beyond the inclusion of more datasets for evaluating the developed methodologies, there are some interesting paths to pursue regarding these approaches. Regarding the actual embeddings, an interesting approach would be to ex-

Dataset	Model	F1	Prec.	Rec.	Model Config.
COMPLETE	ASWEM Audio-only	62.10%	63.03%	63.55%	Batch Size=150*, Optimizer=SGD, Learning Rate=$1e^{-2}$
	wav2vec2-Large	70.84%	71.42%	71.69%	Batch Size=128, Optimizer=Adam, Learning Rate=$4.28e^{-4}$
	MERT-v1-330M	72.37%	74.07%	72.59%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
	ASWEM Lyrics-only	72.33%	72.45%	72.59%	Kernel=Linear, C=2.09
	RoBERTa Large	76.47%	76.28%	75.49%	Kernel=RBF, C=0.10, $\gamma = 6.36e^{-5}$
	ModernBERT Large	76.13%	75.80%	75.14%	Kernel=RBF, C=5.77, $\gamma = 1.03e^{-3}$
BALANCED	ASWEM Audio-only	63.95%	64.01%	64.00%	Batch Size=150*, Optimizer=SGD, Learning Rate=$1e^{-2}$
	wav2vec2-Large	66.81%	66.65%	67.00%	Batch Size=32, Optimizer=Adam, Learning Rate=$4.91e^{-4}$
	MERT-v1-330M	70.05%	69.97%	70.33%	Batch Size=128, Optimizer=Adam, Learning Rate=$4.28e^{-4}$
	ASWEM Lyrics-only	69.67%	80.34%	70.0%	Kernel=RBF, C=1500, $\gamma = 3.35e^{-4}$
	RoBERTa Large	74.46%	75.01%	74.67%	Kernel=RBF, C=$4.40e^{-4}$, $\gamma = 1.09e^{-3}$
	ModernBERT Large	75.32%	75.93%	75.67%	Kernel=RBF, C=176.86, $\gamma = 1.13e^{-3}$

Notes: F1 = Weighted F1 score, Prec. = Precision, Rec. = Recall (in %). *ASWEM audio-only models were tested with fixed hyperparameters.

Table 5: Unimodal Classification Results.

tract embeddings with wav2vec2 from the vocal stem of an audio track and keep the proposed procedure with MERT, possibly fine-tuned on our bimodal MER datasets. It is also of interest to experiment with different numbers of audio and lyrics embeddings in the early fusion approach to control the influence of each methodology on the final predictions. Finally, applying explainability methods to the experimented fusion strategies could provide important insights, particularly helping to understand the different quadrant confusions observed in the two approaches.

Acknowledgments

This work is partially financed through national funds by FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the MERGE project - DOI: 10.54499/PTDC/CCI-COM/3171/2021; and projects UIDB/00326/2025 and UIDP/00326/2025. Pedro Louro was supported by FCT through the PhD scholarship with reference 2024.05205.BD. Renato Panda was supported by Ci2 - FCT UIDP/05567/2020.

7. REFERENCES

[1] J. Kang and D. Herremans, "Are we there yet? A brief survey of Music Emotion Prediction Datasets, Models and Outstanding Challenges," Jun. 2024.

[2] P. L. Louro, H. Redinho, R. Santos, R. Malheiro, R. Panda, and R. P. Paiva, "MERGE - A Bimodal Dataset for Static Music Emotion Recognition," 2024, preprint. [Online]. Available: http://arxiv.org/abs/2407.06060

[3] P. L. Louro, G. Branco, H. Redinho, R. Correia, R. Malheiro, R. Panda, and R. P. Paiva, "MERGE app: A prototype software for multi-user emotion-aware music management," in *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2024, Volume 1: KDIR, Porto, Portugal, November 17-19, 2024*, F. Coenen, A. Fred, and J. Bernardino, Eds. SCITEPRESS, 2024, pp. 159–166. [Online]. Available: https://doi.org/10.5220/0013067800003838

[4] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[5] Y. Feng, Y. Zhuang, and Y. Pan, "Music Information Retrieval by Detecting Mood via Computational Media Aesthetics," in *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, 2003, pp. 235–241.

[6] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[7] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, Oct. 2020.

[8] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional Recurrent Neural Networks for Music Classification," 2016.

[9] M. Won, S. Chun, O. Nieto, and X. Serrc, "Data-Driven Harmonic Filters for Audio Representation Learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 536–540.

[10] J. Lee, J. Park, K. Kim, and J. Nam, "SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification," *Applied Sciences*, vol. 8, no. 1, p. 150, Jan. 2018.

[11] E. Koh and S. Dubnov, "Comparison and Analysis of Deep Audio Embeddings for Music Emotion Recognition," 2021.

[12] P. L. Louro, H. Redinho, R. Malheiro, R. P. Paiva, and R. Panda, "A comparison study of deep learning methodologies for music emotion recognition," *Sensors*, vol. 24, no. 7, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/7/2201

[13] R. M. d. S. Malheiro, "Emotion-based Analysis and Classification of Music Lyrics," doctoralThesis, Universidade de Coimbra, Apr. 2017.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[15] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, 10 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[16] L. Schaab and A. Kruspe, "Joint sentiment analysis of lyrics and audio in music," 2024. [Online]. Available: https://arxiv.org/abs/2405.01988

[17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[18] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-based approach towards music emotion recognition from lyrics," in *Advances in Information Retrieval*, D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, Eds. Cham: Springer International Publishing, 2021, pp. 167–175.

[19] P. N. Juslin, "From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions," *Physics of Life Reviews*, vol. 10, no. 3, pp. 235–266, Sep. 2013.

[20] C. Laurier, "Automatic Classification of Musical Mood by Content-Based Analysis," PhD Thesis, Universitat Pompeu Fabra, 2011. [Online]. Available: http://mtg.upf.edu/node/2385

[21] K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Multi-Modal Song Mood Detection with Deep Learning," *Sensors*, vol. 22, no. 3, p. 1065, 2022.

[22] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music Mood Detection Based On Audio And Lyrics With Deep Neural Net," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 370–375.

[23] AllMusic, accessed: 2025-02-27. [Online]. Available: https://www.allmusic.com/

[24] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 12 449–12 460.

[25] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Y. Guo, and J. Fu, "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training," 2023.

[26] Facebook AI, "wav2vec2-Large-960h," 2021, accessed: 2025-02-11. [Online]. Available: https://huggingface.co/facebook/wav2vec2-large-960h

[27] Multimodal Art Projection, "MERT-v1-330M," 2024, accessed: 2025-02-15. [Online]. Available: https://huggingface.co/m-a-p/MERT-v1-330M

[28] Facebook AI, "RoBERTa large," 2024, accessed: 2025-02-27. [Online]. Available: https://huggingface.co/FacebookAI/roberta-large

[29] R. Kamath, A. Ghoshal, S. Eswaran, and P. Honnavalli, "An enhanced context-based emotion detection model using roberta," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2022, pp. 1–6.

[30] Answer.AI, "ModernBERT-large," 2024, accessed: 2025-02-27. [Online]. Available: https://huggingface.co/answerdotai/ModernBERT-large

[31] K. Choi, G. Fazekas, and M. B. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, M. I. Mandel, J. Devaney, D. Turnbull, and G. Tzanetakis, Eds., 2016, pp. 805–811. [Online]. Available: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/009_Paper.pdf