

MERGE: A Bimodal Audio-Lyrics Dataset For Static Music Emotion Recognition

Pedro Lima Louro, Hugo Redinho, Tiago Filipe Rodrigues Ribeiro, Ricardo Santos, Ricardo Malheiro, Renato Panda and Rui Pedro Paiva

Abstract—The Music Emotion Recognition (MER) field has seen steady developments in recent years, with contributions from feature engineering, machine learning, and deep learning. The landscape has also shifted from audio-centric systems to bimodal ensembles that combine audio and lyrics. However, a lack of public, sizable and quality-controlled bimodal databases has hampered the development and improvement of bimodal audio-lyrics systems. This article proposes three new audio, lyrics, and bimodal MER research datasets, collectively referred to as MERGE (for Music Emotion Recognition next GEneration, the research project that funded its creation), which were created using a semi-automatic approach. To the best of our knowledge, this is the largest bimodal quality-controlled static MER dataset. Furthermore, the dataset is annotated both with discrete emotion categories (according to Russell's four emotion quadrants) and continuous arousal-valence values. To comprehensively assess the proposed dataset and establish a baseline for benchmarking, we conducted several experiments for each modality, using feature engineering, machine learning, and deep learning methodologies. Additionally, we propose and validate fixed train-validation-test splits. The obtained results confirm the viability of the proposed datasets, achieving the best overall result of 81.74% F1-score for bimodal classification.

Index Terms—music emotion recognition, bimodal datasets, music information retrieval, audio analysis, lyrics analysis, feature engineering, machine learning, deep learning.

I. INTRODUCTION

Music Emotion Recognition (MER) has attracted increasing awareness within the Music Information Retrieval (MIR) community. In fact, “music’s preeminent functions are social and psychological”, and so “the most useful retrieval indexes are those that facilitate searching in conformity with such social and psychological functions. Typically, such indexes will focus on stylistic, mood, and similarity information” [1]. Several studies on music information behavior highlighted emotional content as an essential criterion for music retrieval and organization [2]. This growing interest was also driven by the popularity of music streaming platforms and the need to organize and recommend music to their users.

Emotion in music can be understood in three ways: i) expressed, which refers to the emotion that the composer or performer intends to convey with the musical piece; ii)

induced, which relates to the emotion that is evoked in the listener as a reaction to the song; iii) and perceived, pertaining to the emotion that the listener identifies while listening to a song (which may differ from the composer’s original intention and what the listener personally feels in response) [3].

Over the years, several methodologies have been proposed, resulting in a substantial body of research on various aspects of MER, including emotion perception [4], induction [5], and expression [6]. Among these, the focus of this article is on *emotion perception*.

In terms of emotion perception, a significant portion of the studies address static MER (i.e., the identification of the predominant emotion perceived in a song), either regarding emotions as discrete labels [7], [8]) or continuous arousal-valence (AV) values [4]. Besides static MER, other problems have been studied, e.g., Music Emotion Variation Detection (MEVD), where emotion fluctuations throughout songs are analyzed [9], multi-label emotion classification, where each song is assigned several emotion tags [10], and emotion-based automatic playlist generation [11], among others. In this work, our focus is on *static MER*, both using discrete emotion labels (namely Russell’s emotion quadrants) and continuous AV values [12].

MER problems have been primarily tackled using audio as the only information source. Fewer works have employed song lyrics, either alone [13] or following bimodal audio-lyrics strategies [10]. It is well known that employing audio and lyrics enables the exploitation of synergies that result from the combination of the information conveyed by each source. For example, information regarding music arousal (i.e., the energy level of a musical piece) is captured mainly from the acoustic component of the music. On the other hand, the lyrical counterpart is instrumental in valence perception (i.e., how positive or negative the emotions conveyed in a song are). Although most MER systems are based on audio-only data, several psychological studies confirm the importance of lyrics in conveying semantic information. According to Juslin and Laukka [14], 29% of people mention that lyrics are essential to how music conveys emotions. Additionally, Besson et al. [15] have demonstrated that part of the semantic information of songs resides exclusively in the lyrics. Audio-lyrics bimodal systems, herein referred to simply as bimodal systems, emerge naturally, and several studies have confirmed that the combination of these information sources improves MER classification results [16].

To support the research on static MER, several datasets have been proposed over the years. Most current datasets employ audio data, and bimodal datasets are scarce. Moreover,

P. L. Louro, H. Redinho, T. F. R. Ribeiro, R. Santos, and R. P. Paiva are with the University of Coimbra, Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering, and LASI. E-mail: pedrolouro, redinho, tiago.f.ribeiro, ricardocorreia, ruipedro@dei.uc.pt.

R. Malheiro is with CISUC, LASI, and Polytechnic Institute of Leiria - School of Technology and Management. E-mail: rsmal@dei.uc.pt.

R. Panda is with CISUC, LASI, and C12 - Smart Cities Research Center, Polytechnic Institute of Tomar. Email: panda@dei.uc.pt.

quality and sizable datasets are crucial to fully fulfilling the promise entailed in bimodal MER approaches. Since emotion perception is inherently subjective, creating quality and sizable datasets is a challenging and labor-intensive task, prone to disagreement among annotators [2]. Hence, several limitations have been identified in current static MER datasets (either audio-only or bimodal): limited size, lack of diversity in the music genres and styles covered, noisy annotations, noisy samples, lack of public availability, and the use of emotion taxonomies without scientific support [7], [17]. These shortcomings, particularly regarding dataset scale, quality, and information source (typically audio-only), hinder the advancement of MER research.

To address the research gap stemming from the described limitations in the state of the art, we propose a new bimodal static MER dataset, called MERGE¹, containing 2216 bimodal samples, with quality-controlled annotations and sample validation. These bimodal samples were obtained from 3554 audio clips and 2568 lyrics, also annotated and available for single-modality experiments. The dataset is publicly available² and is annotated in two ways, based on Russell's circumplex plane [12]: i) four emotion quadrants; ii) and continuous arousal-valence (AV) values.

Although this is not yet a large-scale dataset, to our knowledge, it is the largest publicly available and quality-controlled bimodal static MER dataset, as discussed in Section III-A. Moreover, besides emotion quadrants and AV values, the dataset provides annotations with multiple emotion tags and genre labels for each song, making it suitable for other problems besides static MER.

To create the dataset, we followed a semi-automatic protocol based on AllMusic³ emotion tags and manual validation. This considerably accelerates a fully manual annotation stage while promoting annotation quality.

To validate the effectiveness of the dataset, we performed a comprehensive set of experiments using state-of-the-art classical Feature Engineering (FE), Machine Learning (ML), and Deep Learning (DL) approaches. These experiments targeted both quadrant classification and AV regression problems, using audio-only, lyrics-only, and bimodal data. The attained results and analysis confirm the viability of the proposed datasets for benchmarking further static MER studies.

The main contributions of this article are the following:

- a novel, public, and quality-controlled bimodal audio-lyrics MER dataset, annotated with emotion quadrants, AV values, multiple emotion tags, and genre labels;
- a semi-automatic approach to construct the dataset;
- an extensive set of experiments comprising audio-only, lyrics-only, and bimodal classification and regression, to establish a baseline for benchmarking.

The document is organized as follows. Section II reviews the relevant background concepts for this work, namely emotion

taxonomies and data collection and annotation approaches. Section III discusses related work regarding available audio, lyrics, and bimodal static MER datasets. Section IV presents the proposed semi-automatic creation protocol, generation of Train-Validation-Test (TVT) splits, and contents of the dataset. Section V describes the methodologies followed for evaluating the proposed datasets and establishing a baseline for benchmarking. The results and insights obtained are discussed in Section VI. Finally, Section VII presents the main conclusions and final thoughts of this study.

II. BACKGROUND

The employed emotion model is a key factor in the creation of MER datasets. Therefore, we start this section with a review of common emotion taxonomies. We then provide an overview of the data collection and annotation approaches, setting the stage for the revision of related work, namely static MER datasets, in the following section.

A. Emotion Taxonomies

Psychology researchers have long discussed how emotions can be represented and classified, leading to the proposal of several emotion taxonomies over the last century. These can be grouped into two major paradigms: categorical (or discrete) models and dimensional models. The two paradigms have been the subject of active research in emotional psychology, and each has its strengths and weaknesses [19], as discussed in the following paragraphs.

1) Categorical Models:

Discrete emotion theories propose the existence of fundamental emotions that are universally shared among cultures. [19]. In the categorical paradigm, emotions are represented as a set of discrete categories or emotional descriptors, e.g., Ekman's basic emotions (comprising anger, disgust, fear, happiness, sadness, surprise) [20], Plutchik's emotion wheel (which presents eight primary emotions in a cyclical format: acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise) [21], or Hevner's adjective circle (where 67 emotion labels are organized into 8 clusters) [22], among others.

Categorical models might be organized into at least four groups: models based on basic emotions, music-specific models, data-driven models, and models defined by music platforms.

Models based on Basic Emotions. Models in this group are based on the theory that humans have a discrete and limited set of basic emotions that are universal and innate [20]. However, the exact number and nature of these basic emotions remain subjects of debate, and it has been argued that emotional responses often arise from the interplay of two or more core emotions [23]. Additionally, the emotions represented in some of these models may not be well-suited to music. For example, Ekman's model aimed to encode facial expressions, and some of the employed categories may not be adequate for the musical case (e.g., disgust). In contrast, others usually associated with music are absent (e.g., calm) [24].

¹MERGE is the acronym of "Music Emotion Recognition next GEneration", a research project funded by the Portuguese Science Foundation.

²Available at: <https://zenodo.org/records/13939205>

³AllMusic is "a popular music database that provides professional reviews and metadata for albums, songs and artists" [18]. URL: <https://www.allmusic.com/>

Music-specific Models. In contrast, music psychologists such as Kate Hevner have proposed categorical models to specifically represent emotions in music [22]. The argument in Hevner's adjective circle (mentioned above) is that it contains descriptors that are musically more plausible. However, the origin of such descriptors is less substantiated. Additionally, in the design of the model, only classical music was employed. Several authors have proposed updates to Hevner's adjectives circle, adding new terms and reorganizing the clusters [25].

Data-driven Models. Besides the previous models (proposed by psychology researchers), MER researchers have also made contributions. A notable example is the Music Information Retrieval EXchange (MIREX) emotion taxonomy [26]. The model was derived from the song metadata provided by AllMusic. At that time, a total of 179 emotion tags were available. The authors proposed a data-driven approach that resulted in a categorical emotion model comprising 29 emotion words grouped into five clusters. This model has also faced some criticism. Being purely data-driven, it lacks support from psychology studies [17]. Moreover, there is semantic and acoustic overlap between some clusters [17] (based on the analysis of the corresponding MIREX dataset, discussed in Section III-A).

Models defined by Music Platforms. Some music platforms propose their own lists of emotion tags. One example is AllMusic, which currently comprises 305 emotion adjectives⁴. This extensive list encompasses a wide range of emotional responses triggered by music. However, the employed emotion adjectives might be ambiguous [2]. Also, although these emotion labels were created (and assigned to the songs in the platform) by professional editors [26], it is unknown whether emotion psychology studies were employed to validate them.

In summary, categorical models offer a simple and intuitive approach to distinguishing between different emotions [10]. A key parameter in this paradigm is the number of emotion words represented. A limited set of words may not adequately capture the full range of emotional responses elicited by music [2]. However, using a more extensive emotion lexicon may not solve the problem, as the language used to describe emotions might be ambiguous and vary significantly between individuals [14].

2) Dimensional Models:

The dimensional perspective sustains that the fundamental components of emotions are a limited set of dimensions. Here, emotions are identified based on their positions on a hyperplane with typically two or three axes, e.g., valence or arousal [12].

In this paradigm, Russell's circumplex emotion model [12] has gained particular acceptance in the MER community. Supporters of this idea suggest that emotional states arise from the combination of two distinct neurophysiological systems: one for valence (pleasure-displeasure, i.e., the polarity of emotion in terms of positive and negative states, also known as pleasantness) and another for arousal or activity (aroused-not aroused, also known as activity, energy, or stimulation level).

Russell even claimed that valence and arousal are the "core processes" of affect, constituting the raw material or primitive of emotional experience [12].

The result, illustrated in Fig. 1, is a two-dimensional plane, referred to as the arousal-valence (AV) plane. There, the X-axis represents valence and the Y-axis represents arousal, resulting in four quadrants that can be defined as: 1) positive valence and arousal, i.e., happy and energetic emotions such as excitement or enthusiasm (Quadrant 1 - Q1); 2) negative valence and positive arousal, i.e., frantic and energetic ones such as anxiety, fear or anger (Q2); 3. negative valence and arousal, i.e., melancholic and sad emotions such as depression (Q3); 4) and positive valence and negative arousal, representing calm and positive emotions such as contentment or serenity (Q4).

Russell's circumplex model (and dimensional models in general) can be represented using a continuous or a discrete perspective, as follows.

Continuous Perspective. In the continuous paradigm, there are no specific discrete emotion tags. Instead, emotions are regarded as a continuum. Thus, each point in the plane can represent a different emotion. For this reason, it is argued that the continuous paradigm better captures the complexity of the emotional space and entails lower ambiguity since no subjective tags are employed [27]. Nevertheless, in Russell's circumplex model, some important aspects of emotion might be obscured, leading to information loss. For example, emotions such as anger and fear (or boredom and melancholy) are closely placed in the AV plane but have distinct meanings. For this reason, a third dimension — dominance (or potency) — was proposed to distinguish such cases (e.g., fear and anger are close in the 2D AV plane but have opposite dominance, i.e., the former is submissive while the latter is dominant) [2]. Another issue is that the continuous paradigm is less intuitive and increases the cognitive load on annotators, making it difficult to acquire consistent annotations. This inconsistency "heightens subjectivity, diminishes dataset reliability, and increases the potential for noise" [10]. This problem is amplified when dominance is employed, which further increases model complexity and cognitive load; therefore, this dimension is not usually used.

Discrete Perspective. In the discrete version of Russell's AV plane, different regions of the emotion plane represent different emotions, described by distinct emotion words (similarly to the previously described categorical emotion paradigm). Russell proposes several discrete emotions distributed across the AV plane (Fig. 1). Besides this representation, Russell's model may be simplified to comprise only four emotions, one for each quadrant, as previously mentioned. The relevance of this taxonomy was validated in a study by Laurier, where a semantic emotion space created from a community of users from the music social network Last.fm⁵ was summarized into these four basic emotions [17]. Although widely used in MER (mostly for its reduced operational complexity, e.g. [7]), the simplified 4-quadrant representation of Russell's AV plane has faced criticism, namely because it might neglect complex aspects of the emotional process and fail to capture

⁴<https://www.allmusic.com/moods>

⁵<https://www.last.fm/>

the full range of emotional responses elicited by music [10]. In summary, the discrete class of dimensional models entails the same advantages and drawbacks discussed above in the categorical emotion paradigm.

All in all, Russell's circumplex model provides a simple yet powerful representation of emotions. The two employed dimensions "provide a good balance between a parsimonious definition of emotion and the complexity of the study" [2]. This is the main reason why it has received broad support from music psychology studies [28] and has been adopted in several MER works, e.g., [7], [17], [29].

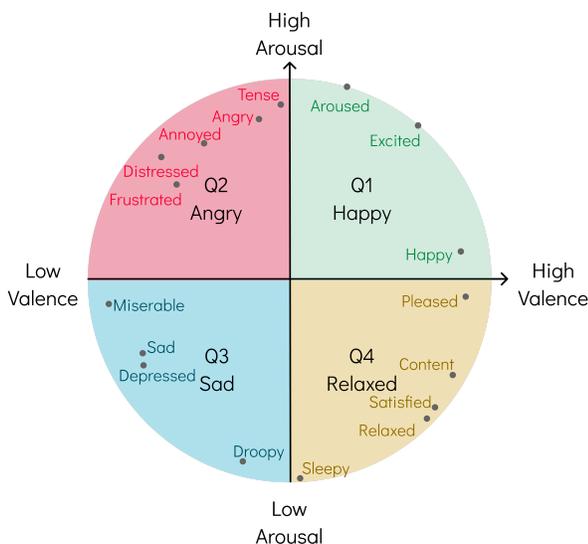


Fig. 1: Russell's circumplex model of emotion as seen in [30].

B. Data Collection and Annotation Approaches

After selecting an emotion model, data must be collected and annotated accordingly. This section starts with a brief overview of data collection tasks and then reviews data annotation strategies.

1) Data Collection, Pre-Processing and Cleaning:

Researchers typically acquire audio clips from music platforms, e.g., AllMusic, Spotify, YouTube, and song lyrics based on lyrics web crawlers, e.g., lyrics.com, ChartLyrics⁶, MaxiLyrics⁷ or MusixMatch⁸. Such platforms usually offer Application Programming Interfaces (API).

In audio MER studies, a musical piece can be an entire song, a structural section of a song (e.g., chorus, verse, bridge, as in [29]), or a fixed-length clip (e.g., 30-second song excerpt, as in [7]).

For static MER, clips with approximately 30-second duration are commonly used to provide excerpts that convey a consistent predominant emotion throughout and to reduce the workload on human annotators [18]. This duration is also used to comply with copyright issues.

⁶<http://www.chartlyrics.com/>

⁷http://www.lyricsmania.com/maxi_lyrics.html

⁸<http://https://www.musixmatch.com/>

Such clips can be manually collected, which is a time-consuming process, albeit more controlled. Thus, they are often automatically extracted from the beginning or middle of the song. In other cases, song previews from music platforms like AllMusic are obtained using the provided API. However, the rationale for creating those excerpts is unclear, as those song previews either correspond to sections from the beginning, middle, and end of the song (or even sections containing only applause in live performances [7]). These processes typically require manual validation of the excerpts, as previous automatic strategies can result in clips that do not accurately represent the predominant emotion or may contain variations (or even noise, such as applause). For this reason, such clips should be validated by humans, although this is not always done, as seen in some datasets, e.g., [10].

The collected audio clips are then pre-processed, e.g., to standardize audio samples in terms of sampling rate, frequency, bit depth, and number of channels, and to remove noisy clips.

Regarding song lyrics, it is essential to correct the text for orthographic errors and remove metadata and other descriptions inside the text (e.g., the name of the artist, title, and metadata about structural elements such as chorus) [13].

2) Data Annotation:

After data collection, pre-processing, and cleaning, song annotation must be performed. To this end, different approaches have been employed in the literature, namely:

- Direct manual annotations, which comprise manual annotations via surveys [27] and manual annotations via annotation games [31];
- Indirect annotations, e.g., where annotations are inferred from social tagging [32] and expert annotations provided by music platforms;
- Indirect annotations with human validation [7], [17];
- Automatic annotations, e.g., via pre-trained models [29] or emotion lexicons [33].

This is the most challenging part, as emotion perception is inherently subjective and annotation is time-consuming [27].

Manual Annotations via Surveys. Controlled manual annotations are typically performed via surveys. Each song is annotated by multiple subjects (typically more than 10) who are instructed to evaluate the emotional content of music in a controlled experimental setup. The process includes clearly defined emotion models, samples, and analysis of inter-annotator agreement [34]. Typically, the most prevalent opinion is selected. This process, coupled with an adequate quality-control protocol, has the potential to generate high-quality annotations. However, it requires hiring subjects, which can be expensive in terms of financial cost and time, besides being a tedious process, and, thus, error-prone [31]. These difficulties are amplified for continuous AV annotations due to the previously mentioned cognitive load on evaluators, making it difficult to acquire consistent annotations and increasing the potential for noise [10]. Moreover, for bimodal annotations, evaluators should make a mental effort to isolate the assessment of each source, e.g., ignoring the lyrics when annotating audio [7], [26] and vice versa, as they may convey different emotions

and affect the judgment of listeners. Several strategies can be implemented to reduce the impact of inaccurate annotations. These include discarding outlier evaluators and songs with low agreement among annotators using statistical tools such as average and standard deviation metrics [13] and inter-coder agreement metrics (e.g., Krippendorff's alpha [35]).

Manual Annotations via Annotation Games. To mitigate the effects of fatigue or lack of commitment associated with manual evaluation, an alternative method for annotating emotions in songs is through collaborative web games, also known as Games With A Purpose (GWAP), such as MoodSwings [31]. The idea behind these games is to increase the commitment and motivation of annotators within a gaming context.

Social Tagging. To tackle the difficulties with direct manual annotations, social tags obtained directly from music social networks such as Last.fm can be employed [17]. Compared to manual annotation, this method makes collecting ground truth data easier and faster. However, there are some problems with the obtained social tags: sparsity due to the cold-start problem and popularity bias, multiple spellings of tags, malicious tagging, or ad-hoc labeling techniques [31]. For example, when a subject uses the tag "hate" on Last.fm, this might either mean that the song is about "hate" or that the person hates the song. As the study by Laurier concludes, "[manual] validation (...) is a necessary step to ensure the quality of the dataset", because by "blindly follow[ing] the tags assigned by the community of Last.fm users, around 29% of errors, on average, would have been introduced" [17].

Expert Annotations from Music Platforms. Compared to the previous approach, a potentially more robust alternative is to employ expert annotations provided by music platforms such as AllMusic [16]. For example, through the AllMusic web service, we can obtain song clips (excerpts with a duration of around 30 seconds, as mentioned above) and their corresponding emotion tags. However, as previously discussed, these tags are not part of any scientifically supported taxonomy. In addition, the annotation process in AllMusic is unclear: all we know is that the employed tags were "created and assigned to music works by professional editors" [26]. Additionally, the audio clips provided by the platform often contain noise (e.g., applause or silence) or segments that do not align with the assigned emotion labels. Hence, the provided music excerpts and the associated emotion tags require post-processing and validation [7], as discussed above.

Annotations from Music Platforms with Human Validation. To partly overcome the described limitations, in [7], we proposed a semi-automatic data collection and annotation strategy based on AllMusic annotations. The basic idea was to map the provided multi-label annotations of each song to a single emotion quadrant (according to Russell's model). To increase the robustness of the annotations, the quadrant proposed for each song was manually validated. In this article, we extend our original approach to address the bimodal case and to provide AV annotations, as described in Section IV.

Automatic Annotations. Another possibility is to collect emotion annotations automatically. This strategy is applied in lyrics emotion datasets, where emotion lexicons such as ANEW [36] or Warriner et al.'s [37], which provide arousal-

valence-dominance (AVD) ratings of English words [33], are used. However, the fact that human validation is not performed is a point of concern. Other approaches utilize pre-trained classification models to collect emotion annotations [29]. However, it is questionable whether labels provided by machine learning systems can be regarded as actual annotations.

III. RELATED WORK

After reviewing the main emotion taxonomies and data collection and annotation approaches, we critically review current static MER datasets, highlighting the limitations that our proposed new dataset aims to address.

A. Current Static MER Datasets

Over the years, several MER datasets have been introduced in the literature to address different MER problems, e.g., static MER [7], multi-label MER [10], Music Emotion Variation Detection (MEVD) [34], and music emotion induction [38], among others.

In the following, we provide an overview of static audio, lyrics, and bimodal audio-lyrics MER datasets as summarized in Table I. We also review some multi-label MER datasets. Although these datasets do not directly address static MER, they could be adapted, for example, by mapping the employed emotion tags to Russell quadrants. For each dataset, we provide a brief overview of the data collection and annotation process, highlighting its strengths and shortcomings.

1) Audio-only Datasets:

Initial attempts to propose benchmarks involved challenge-related databases, such as the MIREX Audio Mood Classification (AMC) dataset [18] (for static MER). The MIREX dataset contains 600 audio clips (30 seconds each), equally distributed across the 5-cluster MIREX emotion taxonomy (previously described). The audio clips were carefully annotated by 15 human volunteers. However, several issues have been identified: i) the defined emotion taxonomy is not grounded in psychology studies; ii) some of the defined emotion clusters show semantic and acoustic overlap [17]. Finally, the dataset is not publicly disclosed and can only be accessed within the MIREX challenge.

Another challenge-related contribution is the Database for Emotional Analysis in Music (DEAM) [34], which resulted from the successive benchmarks for the 2013, 2014, and 2015 MediaEval Emotion in Music tasks. This dataset contains 58 full-length songs and 1,744 excerpts of 45 seconds (the ones considered for static MER) annotated with continuous arousal-valence (AV) values. Each audio clip was annotated by at least 10 subjects, and several quality-control procedures were implemented, including measuring annotator consistency and filtering out evaluators with low-quality annotations. Unlike the MIREX AMC dataset, DEAM is publicly available.

Apart from the above tentative benchmarks, Yang et al. presented one of the first public audio datasets [27]. The dataset comprised 25-second excerpts from 195 popular songs (representing the predominant emotion present, mainly the chorus) taken from Western, Chinese, and Japanese albums.

TABLE I: Static MER datasets (sorted by modality and year).

4Q = Russell's 4 emotion quadrants; AV(D) = continuous AV(D) values; Bimodal = audio + lyrics; Game = manual annotations via annotation games; Manual = direct manual annotations; MP = expert annotations from Music Platforms; n.a. = not available / not applicable; ST = Social Tagging; VL = Variable Length.

Dataset	Year	Modality	Availability	Size	Sample Length	Emotion Model	Annotation Method
MIREX AMC [18]	2007	Audio	Private	600	30 sec	MIREX 5 clusters	Manual
Yang et al. [27]	2008	Audio	Public	195	25 sec	AV	Manual
MagnaTagATune [39]	2009	Audio	Public	25,877	29 sec	27 labels	Game
MSD [32]	2009	Audio	Public (only audio features)	1,000,000	n.a.	Over 100 labels	ST
DEAM [34]	2013-2015	Audio	Public	1,744	45 sec	AV	Manual
MSD Last.fm split [40]	2016	Audio	Private	242,842	n.a.	At least 3 labels	ST
4QAED [41]	2018	Audio	Public	900	30 sec	4Q	MP with validation
EMOPIA [42]	2018	Audio (+MIDI)	Public	1,087	VL	4Q	Manual
MTG-Jamendo [43]	2019	Audio	Public	55,609	Full songs	59 labels	ST
MuSe [44]	2021	Audio	Public (only song IDs)	90,001	30 sec	AVD	Automatic
LED1 [13]	2016	Lyrics	Public (links to lyrics)	180	Full lyrics	AV	Manual
LED2 [13]	2016	Lyrics	Public (links to lyrics)	771	Full lyrics	4Q	Manual
MoodyLyrics [33]	2017	Lyrics	Public (links to lyrics)	2,595	Full lyrics	4Q	Automatic
Laurier et al. [45]	2008	Bimodal	Private	1000	30 sec Full lyrics	4Q	ST with validation
Hu & Downie [16]	2010	Bimodal	Private	5296	Full songs Full lyrics	18 emotion clusters	ST
Delbouys et al. [46]	2018	Bimodal	Private	18644	30 sec Full lyrics	AV	ST
Music4All [47]	2020	Bimodal	Upon request	109,269	30 sec Full lyrics	AV	Automatic
Popular Hooks [29]	2024	Bimodal (+Video+MIDI)	Public	38,694	VL Full lyrics	4Q	Automatic
Hu et al. [10]	2025	Bimodal	Public (only song IDs)	169,148	Full songs Full lyrics	12 labels	Automatic

The fully manual annotation process involved 253 subjects, who labeled ten random samples with AV values. The final AV values were obtained by averaging all annotations. The dataset quality was deemed acceptable based on the test-retest study, where the annotation process was repeated two months after the initial annotation. However, the dataset is very imbalanced, e.g., only 12% of all samples belong to the second quadrant.

The MagnaTagATune is a sizable music tagging dataset, aiming at multi-label classification and annotation through a

gamification process [39]. A total of 25,877 samples from 5223 full songs are provided, accompanied by 29-second audio clips, with 188 unique tags (freely assigned by the players) across them, ranging from common high-level descriptors such as genre, emotion, and era to instruments and specific performing techniques (e.g., plucking). The list contains 27 emotion tags, e.g., happy, relaxed, joyful, angry, romantic, nostalgic, etc. According to [40], [43], the dataset suffers from limitations, including noisy labels and an unbalanced

tag distribution. Moreover, there is no standard data split for benchmarking, leading to inconsistency in the results reported using this dataset [43].

The Million Song Dataset (MSD) is another example of a music tagging dataset. It was specifically designed to address the limited size of available datasets, and it remains the largest dataset in the MIR field. Data annotation is based on the collection of tags of multiple types, including emotion labels, provided by the users of the platforms such as Last.fm. A total of 522,366 tags are included. The exact number of emotion tags is unknown, but we estimate that it contains hundreds. A drawback of the MSD dataset is that it only provides audio features and metadata, so researchers typically collect song previews from external platforms. As different studies employ different song clips, distinct results are reported, making benchmarking difficult. Moreover, it suffers from the previously mentioned limitations of approaches based on social tags, namely the lack of tag validation and the associated ambiguities.

Choi et al. [40] created another dataset focused on music tagging, based on a subset of the MSD, hereafter referred to as the MSD Last.fm split. It contains 242,842 samples obtained using track identifiers from the MSD. The dataset only includes samples with metadata that contains at least 50 unique general-purpose tags. As for emotion labels, the complete list is not disclosed, although three tags are mentioned: sad, happy, and chill. Despite being supposedly publicly available, the audio files cannot be acquired from the original provider, 7digital, as their API is no longer operational [48].

Regarding our team's efforts, the most recent public dataset is the 4-Quadrant Audio Emotion Dataset (4QAED) [7]. As previously mentioned, the creation of this dataset relied on collecting data and expert annotations from AllMusic. The main idea was to map the provided multi-label annotations of each song to an emotion quadrant (according to Russell's model) based on the ANEW English lexicon [36]. The dataset comprises 900 audio samples with accompanying metadata, evenly distributed across the four quadrants. The sample data consists of 30-second song excerpts, their respective categorical labels (emotion quadrants and the original AllMusic tags), and 1714 extracted features. As previously mentioned, the quadrant proposed for each song was manually validated.

MTG-Jamendo [43] is another large-scale music tagging dataset. It comprises 55,609 full audio tracks with 195 tags freely assigned by content uploaders of the Jamendo platform⁹. The list of tags comprises 95 genres and 59 emotions and themes, among others. Five random data splits (training, validation, and testing) are provided to permit benchmarking of the results. However, as in other large-scale datasets, the labels might be noisy and ambiguous since no manual validation was performed. Moreover, the dataset is significantly unbalanced.

The EMOPIA dataset [42] is a particular dataset in this category, since it contains 1,087 variable-length audio clips (from 387 full songs), paired with MIDI samples. The clips were carefully annotated by four subjects. Moreover, song

metadata and listener demographic information are provided. However, the dataset is restricted to pop piano music.

The previously described large-scale datasets, e.g., MagnaTagATune, MSD, MSD Last.fm split, and MTG-Jamendo, all addressed multi-label emotion classification in the context of music tagging. These datasets employed emotion tags freely assigned by users. More recently, in 2021, the Musical Sentiment Dataset (MuSe) [44] was introduced. This is a large-scale dataset specifically targeting static MER. The MuSe dataset includes arousal-valence-dominance (AVD) annotations for 90,001 songs, derived from user tags on Last.fm, incorporating 279 emotion labels from AllMusic as seeds. Similar to [7], the emotion tags are mapped to the AVD cube using the previously mentioned English lexicon by Warriner et al. [37], with the advantage that a weighted average of AVD values is calculated based on weights for each tag from Last.fm. However, audio clips are not directly available. Instead, the dataset provides Spotify IDs for 61,630 tracks, enabling the automatic retrieval of song previews, albeit with the limitations previously mentioned regarding the validity of the audio clips. Furthermore, according to Bogdanov et al. [49], only 41,021 30-second previews were accessible via the Spotify API (which was closed to public access by the end of 2024). Most importantly, no human manual validation has been conducted on the MuSe dataset. Therefore, the issues related to annotations based on social tagging remain unaddressed, which can lead to noisy annotations, as previously discussed. This is reflected in the low R2 scores obtained for arousal and valence (0.143 and 0.089, respectively), as discussed in [49], since the authors of the MuSe dataset did not provide a baseline.

2) Lyrics-only Datasets:

Compared to audio-based MER approaches, fewer works have employed song lyrics, as reflected in the smaller pool of available datasets. Here, we describe lyrics-only datasets. Other datasets containing song lyrics were created in the context of bimodal MER, and will be described afterwards.

The largest dataset consisting solely of song lyrics is MoodyLyricsPN [50], proposed by Çano and Morisio. It contains 5000 songs labeled as positive or negative based on Last.fm tags (thus, with limitations of approaches based on social tags). In the same article, the authors also proposed the MoodyLyrics4Q dataset, which contains 2000 songs labeled according to Russell's four quadrants and annotated based on Last.fm tags as well. The same authors also proposed the MoodyLyrics dataset [33]. It comprises 2595 lyrics annotated with Russell quadrants, based on the emotional weight of the words present in the text, using emotion lexicons such as ANEW [36] to map the aggregate AV values of all words in a song to each quadrant. It covers a large number of unique artists, and it is nearly balanced (although the first quadrant is over-represented). The construction of the three datasets was fully automatic. However, as previously mentioned, the absence of human validation is a point of concern.

Our team developed two lyrics-only datasets (termed LED, for Lyrics Emotion Datasets), as presented in [13], with a total of 942 lyrics. The first consisted of 180 manually annotated samples, and the second followed a semi-automatic annotation

⁹<https://jamendo.com/>

process akin to 4QAED, resulting in 771 samples. Some drawbacks of this dataset include its small size and slightly unbalanced quadrant distribution.

3) *Bimodal Audio-Lyrics Datasets:*

The first bimodal dataset we are aware of is the quality-controlled 1000-song dataset created by Laurier et al. [45]. It uses Last.fm as the source for audio and lyrics samples. The same platform provides the employed emotion tags, which are mapped to the four Russell quadrants and manually validated by 17 human subjects.

Hu [16] created one of the first bimodal audio-lyrics datasets, which aimed at multi-label emotion classification. It includes 5296 audio and lyrics retrieved from the Last.fm platform, along with their emotion tags. Similar tags were clustered into larger groups, resulting in 18 emotion categories that formed a data-driven emotion taxonomy (not validated by music psychology studies). Manual validation is not mentioned, and its quality cannot be assessed since it is private.

Another contribution is the sizable 18644-song dataset by Delbouys et al. [46], which used the MSD as its source. The dataset was annotated with continuous AV values by automatically mapping Last.fm tags into the AV space using Warriner et al.'s lexicon of English words and their corresponding AVD values [37]. 30-second excerpts were extracted from the complete songs (no exact details about the procedure were provided). The annotations were heavily biased towards audio, leading to possible conflicts with the emotional content of the lyrics, and, as was the case for Hu's dataset, its quality cannot be assessed because it is private. Adding to the lack of manual validation, the low regression results suggest that the samples and annotations are noisy.

In 2020, the Music4All dataset was released [47]. This large-scale dataset comprises 109,269 songs, annotated with AV values obtained automatically using the Spotify API (where the energy feature serves as a surrogate for arousal). The dataset is available upon request. Since the annotation relies on pre-trained models, this automatic process may introduce noise, as previously discussed and specifically confirmed for Spotify's valence and energy features [51]. For the audio content, 30-second excerpts were automatically extracted from the middle of the song without any human validation. Additionally, lyrics were automatically obtained via MusixMatch, but there was no human validation to verify whether the emotions perceived in the acoustic and textual components matched, as discussed earlier.

Popular Hooks, a fully accessible multimodal dataset, was introduced [51]. The dataset comprises 38,694 hooks (i.e., memorable sections of songs) accompanied by synchronized audio, lyrics, music videos, and MIDI files. In addition to emotion annotations based on Russell's quadrants, a notable feature of this dataset is the inclusion of musical metadata such as song structure and tonality, as well as genre information. The emotion quadrant annotations were generated automatically using pre-trained models (one for each data source), and their accuracy was assessed through a user study involving 80 randomly selected samples from the dataset. Despite the authors' efforts to validate their automatic annotation approach, the

small size of the validation sample might not be sufficient to ensure the quality of the annotations. Furthermore, no baseline was provided using the entire dataset. It is also important to note that the dataset is heavily unbalanced, with approximately 85% of the data falling under Q1 and Q2.

Recently, Hu et al. [29] developed a large-scale bimodal dataset to address multi-label emotion classification. This dataset includes 169,148 complete songs from the NetEase Cloud Music platform, a freemium music streaming service based in China. A notable limitation of the dataset is that it only provides song IDs, which poses a challenge since the platform is exclusively available in China. In terms of emotion classification, the platform uses 12 predefined emotion tags: exciting, fresh, healing, happy, lonely, missing, nostalgic, quiet, romantic, relaxing, sentimental, and touching. The emotion labels for each song are gathered from user playlists. Although this process is automatic and lacks human validation, it appears reliable as it reflects the collective emotional consensus of multiple users. However, the employed taxonomy is limited in scope, as it does not cover a significant portion of the emotional spectrum; particularly, none of the tags fall within the second quadrant of Russell's circumplex model. The lyrics for the songs were obtained using the song IDs. Unlike other bimodal datasets, the authors took care to clean the data by removing irrelevant information, such as timestamps. However, similar to other datasets, there was no human verification to confirm if the emotions perceived in the acoustic and textual components aligned, as discussed previously.

4) *Discussion:*

Based on the analysis of the previous MER datasets, we highlight the following key points:

- There is a significant diversity in the employed emotion taxonomies. However, Russell's circumplex model is the most common approach, whether it is represented through discrete quadrants or continuous AV values. All other models lack validation from music psychology studies.
- Most of the reviewed datasets exhibit a good variety of genres and styles.
- Several large-scale datasets have been proposed; however, all of them are annotated using social tagging or automatic methods without human validation, resulting in a lack of quality control.
- As expected, datasets that include quality control are generally smaller due to the need for thorough human validation. One of the largest datasets that incorporates such quality control is DEAM, which contains 1,744 audio clips.
- None of the bimodal datasets performs human validation of the annotations.
- Furthermore, none of the bimodal datasets verify whether the emotions perceived in the acoustic and textual components align with each other. It remains unclear whether participants annotated songs based on audio, lyrics, or both. As previously mentioned, audio and lyrics should be annotated separately to evaluate their specific contributions to music emotion recognition.

- While most of the reviewed datasets are publicly accessible, some of the large-scale datasets only provide song IDs. Even if it is possible to acquire the audio clips, validation is still necessary, as discussed earlier.
- Some reviewed datasets, particularly recent large-scale datasets, do not provide baselines for assessing feasibility and establishing a basis for future comparison.

In light of this analysis, we propose a new dataset in the following section to address some of the limitations identified in current static MER datasets.

IV. PROPOSED DATASET

The proposed MERGE dataset comprises audio, lyrics, and bimodal modalities, enabling both single- and bimodal research. Each modality includes two variations: i) complete, i.e., all songs without any balancing; ii) balanced, focusing on even distribution across both quadrant and genre, following the protocol established by Panda et al. [7]. The datasets, along with their metadata and extracted features (audio and lyrics), are publicly available¹⁰.

The remainder of this section describes the building process and contents of the dataset. Before that, we present a set of requirements to consider during the dataset creation process.

A. Requirements for MER Datasets

After reviewing the MER datasets, we have established the following requirements for the MERGE dataset:

R1. Validated taxonomy: The dataset should be based on psychologically validated taxonomies. For simplicity, a reduced set of consensual emotional terms should be employed. From the discussion in Section II, Russell's circumplex model appears as the most promising approach, whether through discrete quadrants or continuous AV values. Although the previously described study by Laurier [17] concludes that Russell quadrants may adequately summarize emotion tags employed by Last.fm users, this simplified approach might fail to capture the complexity of the emotional space. Therefore, we propose using a taxonomy that combines both discrete Russell quadrants and continuous AV values.

R2. Variety and balance: Datasets should be varied, balanced, and not limited to a single musical genre, style, or era.

R3. Careful annotation: As discussed, datasets indirectly annotated from social tags or automatic systems should be validated by humans. In addition, noisy samples (audio and lyrics) should be discarded or cleaned.

R4. Reduced ambiguity: At least good annotator agreement should be achieved, minimizing the mentioned ambiguity issues. This would lead to datasets with reasonably clear emotions, a key need at the current stage of MER research.

R5. Separate annotation between audio and lyrics: When creating bimodal MER datasets (containing audio and lyrics), care should be taken to isolate the two sources

in the annotation process so that the impact of each modality can be properly assessed.

R6. Public availability: It is necessary that the datasets be public to permit a comparative analysis of different methods.

R7. Large size: sizable datasets are required to exploit ML and DL solutions better.

R8. Provision of a baseline: A baseline should be provided to assess the feasibility of the dataset and set the stage for future comparative analysis.

We also defined two additional secondary requirements:

S1. Metadata for a wide range of research works:

Besides emotion annotations, datasets should provide metadata such as genre, artist, album, year, and complete emotion tags. These would make the dataset relevant for the broader Music Information Retrieval (MIR) field and might be useful for later, more advanced tasks such as multi-label emotion classification.

S2. Semi-automatic construction process: Probably, the main difficulty with the previous primary requirements is that at least part of the annotation process must involve manual human validation. This calls for semi-automatic construction approaches, reducing the resources needed to build a sizable dataset, as discussed below.

B. Creation Protocol

We guided the creation of the new dataset by the above requirements. Algorithm 1 describes the dataset creation procedure (adapted and improved from our previous work [7]). We outline the key concepts of the proposed algorithm in the following sections.

After gathering audio clips from AllMusic using the provided API¹¹, a key step of our approach is the mapping of AllMusic emotion tags (curated by AllMusic experts) to Russell's quadrants. To this end, we employ Warriner's adjectives list [37], which contains 13915 emotion adjectives (in English) with affective ratings in three dimensions: arousal, valence, and dominance (AVD). We then map each song to a point in the AV plane by averaging the emotion tags based on Warriner's scores. To minimize ambiguity, songs located near the center of the plane, specifically in the interval $[-0.2, 0.2]$ (on a $[-1, 1]$ scale), are excluded. Additionally, we maximize genre variability in each quadrant.

Following the audio collection, their corresponding lyrics are retrieved from platforms such as lyrics.com, Chart-Lyrics, MaxiLyrics, and MusixMatch. In this process, lyrics could not be found for some of the audio samples. The lyrics were then cleaned through a series of operations. These include correcting spelling errors, eliminating lyrics that are not in English, removing lyrics with less than 100 characters, and removing metadata text, among others. Additionally, the lyrics were complemented according to the corresponding audio. This means that repetitions of the chorus in the audio are added to the lyrics. Similarly, metadata defined in the lyrics (e.g., [Chorus x2]) implies adding one more instance of the

¹⁰Available at: <https://zenodo.org/records/13939205>

¹¹Available audio samples and corresponding metadata were retrieved through <https://tivo.stoplight.io/docs/music-metadata-api>.

chorus to the lyrics. After making these additions, the lyrics are then checked for any remaining cases of these patterns and eliminated. This process is described in greater detail in [13].

A crucial step in our approach is the manual validation of the acquired songs, specifically in terms of the assigned quadrants and the quality of the audio clips and lyrics. To achieve this, we conducted a blind inspection of the candidate set. Participants received sets of randomly distributed audio clips and song lyrics and were asked to annotate them according to Russell’s quadrants. If any sample was of poor quality (such as being noisy, containing claps, or being silent), it was discarded. A song was retained if the annotated emotion quadrant aligned with the quadrant derived from mapping the AllMusic emotion tags; otherwise, it was discarded. Similar to the method described in [17], we consider a song valid if at least one annotator confirms the tag. This step is crucial because it reduces the cognitive burden associated with a fully manual annotation process. By ensuring that expert annotations are carefully obtained, we can validate these with minimal human resources. A total of 8 participants were involved in the validation process.

The bimodal dataset is created from validated audio and lyrics datasets, including songs that correspond to the same audio and lyrics quadrants. These datasets are referred to as the “complete” datasets. To ensure balance, the “balanced” audio, lyrics, and bimodal datasets are formed by removing samples from the quadrants that are more heavily represented. This approach results in equally represented quadrants while also maintaining genre balance. Moreover, AV values are also provided (on a [-1, 1] scale).

In addition to the procedure described, the original 4QAED and LED datasets served as the foundation for creating the MERGE dataset. Whenever possible, lyrics were obtained for the audio-only samples from the 4QAED dataset, while audio was retrieved for the lyrics-only samples from LED.

The following paragraphs will discuss the resulting number and distribution of samples across quadrants for each dataset.

C. Dataset Description

The resulting datasets are hereafter termed *MERGE Audio* (AC for the complete variation and AB for the balanced variation), *MERGE Lyrics* (LC and LB), and *MERGE Bimodal* (BC and BB) and are summarized in Table II.

TABLE II: Datasets used for evaluation with respective sample distribution.

Dataset	Q1	Q2	Q3	Q4	Total
MERGE Audio Complete	875	915	808	956	3554
MERGE Audio Balanced	808	808	808	808	3232
MERGE Lyrics Complete	600	710	621	637	2568
MERGE Lyrics Balanced	600	600	600	600	2400
MERGE Bimodal Complete	525	673	500	518	2216
MERGE Bimodal Balanced	500	500	500	500	2000

In short, MERGE AC contains 3554 samples, while MERGE AB includes 3232 (808 per quadrant). For lyrics, MERGE LC contains 2568 samples, while MERGE LB has

600 samples per quadrant (2400 samples in total). Finally, MERGE BC comprises 2216 samples, whereas MERGE BB contains exactly 2000 samples (500 samples per quadrant). As can be observed, the audio sets are larger since, as mentioned previously, retrieving lyrics from the corresponding songs was not always possible. The bimodal dataset is smaller, as the annotated audio and lyrics quadrants do not always align.

In addition to audio clips and lyrics, each dataset includes individual metadata and train-validation-test (TVT) splits. The metadata file contains attributes such as the song identifier, title, artist, year, and genre tags, along with annotated quadrant and arousal-valence values. By providing this additional information, our datasets can be utilized for related tasks, such as music tagging (which employs audio and emotional labels) or era recognition. Regarding the provided AV values, it is important to bear in mind that these values were not manually validated; instead, they were obtained automatically from the mapping of the original emotion labels into the AV plane using Warriner’s dictionary.

The TVT splits provided for each dataset come in two configurations (for training, validation, and testing): 70-15-15 and 40-30-30. These splits were created using a method designed to maximize quadrant balancing and genre distribution across each set. In addition to experiments that use k-fold cross-validation, we encourage researchers to use the proposed TVT splits rather than create their own. This approach ensures the reproducibility of results.

Finally, as we will show in the following sections, among the proposed datasets, experiments with MERGE Bimodal Complete (BC), using the 70-15-15 TVT split, typically yielded the highest F1-scores (which were comparable to cross-validation F1-scores). Also, this data split is well-suited for optimizing and quickly validating MER systems. For this reason, we propose the MERGE BC dataset and the 70-15-15 TVT split as the *main profile* for future benchmarking.

V. BASELINE METHODOLOGIES AND EVALUATION STRATEGY

Following the description of the MERGE dataset and the process involved in its creation, our aim is to assess its feasibility. Therefore, this section presents the overall evaluation strategy employed in this study, followed by descriptions of the baseline methodologies developed for this study.

We performed various experiments to establish a baseline for benchmarking and provide a comprehensive evaluation of the proposed dataset. State-of-the-art approaches were employed for both quadrant classification and AV regression in each modality (audio-only, lyrics-only or bimodal), in two categories: i) feature engineering and classical machine learning; ii) and deep learning. In [52], an in-depth survey of approaches from the two categories is provided.

We note that the main purpose of these experiments is to assess the dataset’s feasibility rather than to focus on the novelty of the employed models or to propose new methods. Our goal with the selected baseline methods is to establish a foundation for future, more advanced models. We are aware that other methods in the literature may achieve higher classification and regression scores.

Algorithm 1 Dataset creation algorithm.

1. Gather songs and emotion data from AllMusic services.
 - 1.1. Retrieve the list of 289^a emotion tags, E , using the AllMusic API.
 - 1.2. For each emotion tag gathered, E_i , query the API for the top 10000 songs related to it, S .
2. Bridge the tags from AllMusic with Warriner's list.
 - 2.1. For each emotion tag, E_i , retrieve the associated AV (arousal, valence) values from Warriner's dictionary of English words. If the word is missing, remove it from the set of tags, E .
 - 2.2. Using the retrieved AV values, map each emotion tag, E_i , onto one of the four Russell's quadrants.
 - 2.3. Assign a quadrant to each song, S_i , based on the quadrant where the majority of the emotion tags, E_i , fall.
3. Perform data pre-processing and filtering to reduce the massive amount of gathered data to a more balanced but still sizable set, FS .
 - 3.1. Filter ambiguous songs (where a dominant emotional quadrant is not present).
 - 3.1.1. For all the songs in S_i , calculate the average arousal and valence values of all the emotion tags gathered, E_i .
 - 3.1.2. If the average value of valence or arousal is in the range $[-0.2, 0.2]$ (on a $[-1, 1]$ scale), remove the song from the dataset.
 - 3.2. Remove duplicated or very similar versions of the same songs by the same artists (e.g., different albums) by using approximate string matching against the combination of artist and title metadata.
 - 3.3. Eliminate songs without genre information. This ensures that the algorithms that maximize genre diversity can function correctly.
4. Generate a subset, GS , maximizing genre variability in each quadrant.
5. Obtain the manually validated audio dataset, ASV .
 - 5.1. Distribute all the songs in the set GS equally among all team members.
- 5.2. For each song, GS_i , validation and annotation are performed according to Russell's quadrants.
 - 5.2.1. Verify that the song is valid (e.g., does not contain clapping, noise, or silence) and that the emotion present in the song is not ambiguous.
6. Retrieve the lyrics dataset, LS , corresponding to the validated audio clips, AVS , from the following platforms: lyrics.com, ChartLyrics, MaxiLyrics, and MusixMatch, leading to the lyrics dataset (instrumental songs will be discarded from the lyrics dataset).
7. Perform lyrics cleaning, e.g., spell checking, removal of non-English lyrics, removal of metadata, etc.
8. Obtain the manually validated lyrics dataset, LSV .
 - 8.1. Distribute all the songs in the LS set equally among all team members.
 - 8.2. For each song, LS_i , perform validation and annotation of the song according to Russell's quadrants.
 - 8.2.1. Verify that the lyrics file is well structured, belongs to the correct audio clip, and that the emotion in the file is not ambiguous.
9. Define the bimodal dataset, Bm , by keeping only the songs where audio and lyrics annotations match.
 - 9.1. For each song, ASV_i and LSV_i , if the annotated audio and lyrics quadrants match, the song is added to the bimodal dataset; otherwise, the song will be discarded (but present in the audio subset with a given quadrant and in the lyrics subset with a different quadrant).
10. Create the final complete and balanced audio, lyrics, and bimodal datasets.
 - 10.1. The above ASV , LSV , and Bm datasets form the complete sets, containing annotations for discrete quadrants and continuous AV values (on a $[-1, 1]$ scale).
 - 10.2 From the datasets in 10.1, obtain balanced datasets, ASV_b , LSV_b , and Bm_b , respectively, by discarding samples from the more represented quadrants, respecting genre balancing.

^aWhen we began creating this dataset, AllMusic offered 289 emotion labels; the current total is now 305, as previously mentioned.

A. Evaluation Strategy

As previously mentioned, we approach static MER as both classification and regression problems.

For these two problems, two evaluation methods were used to evaluate the performance of each methodology: stratified 10-fold cross-validation with 10 repetitions, and the two previously described TVT splits (70-15-15 and 40-30-30). Optimal hyperparameters were found through Bayesian search optimization. The obtained optimal parameters for all models are provided as supplementary materials.

Statistical significance tests were performed to compare the classification results from the proposed models and modalities on the relevant experiments. Differences are statistically significant for $p < 0.05$.

B. Audio Classical ML

Our approach in [7] served as the basis for the conducted classical ML experiments. All songs are standardized, and features related to the eight standard musical dimensions (melody, harmony, rhythm, dynamics, expressivity, texture, and form) are extracted, as detailed in [7].

The ReliefF algorithm was then employed for feature ranking and selection. For classification, SVMs¹² were used, and their optimal hyperparameters were obtained through a Bayesian search. As for regression, Support Vector Regression (SVR) was employed. The kernel radial basis function (RBF) was selected, as earlier experimentation led to better results.

¹²Other classical ML approaches were evaluated (e.g., K-Nearest Neighbours and Random Forest), but SVMs achieved the best results.

This kernel requires tuning cost (C), set to [1e-6, 1500], and γ , set to [1e-6, 100]. A logarithmic uniform step size was defined in both, allowing a higher number of smaller values to be tested.

Most of the implementations (SVMs and the Bayesian search algorithm used for model optimization) are provided by the scikit-learn Python library¹³. ReliefF is an exception, being provided by the *attrEval* function of the CORElearn R package¹⁴, for which no equivalent was found in Python.

In both evaluation methods, different numbers of features were tested to find the optimal feature set. This was performed independently for each dataset containing audio, as mentioned in Section IV-C.

C. Audio Deep Learning

Our audio DL baseline uses as inputs audio embeddings from the wav2vec2 model [53]. These audio embeddings are processed to reduce them to a 1024-sized vector (depicted at the top of Figure 2), followed by two dense layers (at the right side of the same figure).

The wav2vec2 model family comprises pre-trained models specifically designed for speech recognition. The large version, trained on 960 hours of 16 kHz speech data, is used in this study. The model processes raw audio through a multi-layer convolutional feature encoder and seven temporal convolutional blocks, outputting audio features. A Transformer-like context network employs dynamic convolutions for relative positional embeddings on these features, enabling the extraction of information primarily related to acoustics and timbre.

The model produces a vector consisting of 1,024 values for each timestep across the 25 hidden layers of its context network. Since we focus on static emotion, we averaged the timestep information, resulting in a concise set of 1,024 values for each of the 25 layers (i.e., a 1024x25 matrix). This matrix is then fed to a convolutional 1D layer, followed by a flatten layer, to reduce the audio embedding to a 1024-sized vector.

The sample rate is lower compared to classical experiments to reduce the complexity of the model. It has also been stated that such reduction does not impact the model's performance [54], as confirmed experimentally.

A Bayesian optimization strategy akin to the one used for the classical approach is applied using the Optuna library¹⁵. The hyperparameters tuned using this strategy include the optimizer (Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (Adam), and Adamax), learning rate, and batch size. Search spaces are defined as follows: the learning rate is searched in the range [1e-5, 1e-2], and batch size in [32, 256].

We implemented an early stopping strategy for the classifier methodologies to prevent the models from overfitting on the training data. If the model's validation accuracy does not improve noticeably after 15 epochs, the optimization is stopped, and the next set of candidate hyperparameters is tested.

¹³<https://scikit-learn.org/>.

¹⁴<https://www.rdocumentation.org/packages/CORElearn/>.

¹⁵<https://optuna.org/>.

D. Lyrics Classical ML

The basis for the following machine learning experiments, which includes data pre-processing, feature selection, and the creation of classification and models, is described in [13].

Regarding feature extraction, we use the features proposed in [13], which are divided into content-based (e.g., bags-of-words), stylistic (e.g., number of occurrences of nouns, adjectives, adverbs, slang words, etc.), song-structure (e.g., number of repetitions of the chorus and song title, etc.), and semantic features (e.g., features extracted from frameworks such as Synesketch¹⁶, ConceptNet¹⁷, LIWC¹⁸ and General Inquirer¹⁹, as well as features based on word dictionaries (gazetteers) related to each of Russell's emotion quadrants.

As in audio, SVMs are used to create classification models, which are parameterized with an RBF kernel and tuned using Bayesian parameter search. We also employ the ReliefF algorithm for feature selection and ranking.

The optimization strategy presented in Section V-B (repeated 10-fold cross-validation and TVT) was also applied to the lyrics counterpart.

E. Lyrics Deep Learning

Similar to audio, our lyrics DL baseline uses as inputs word embeddings from the Robustly Optimized BERT Pre-Training Approach (RoBERTa) pre-trained model [55]. These word embeddings serve as inputs to a neural network with two dense layers (as depicted at the bottom of Figure 2).

RoBERTa employs self-attention and pre-training mechanisms on large text corpora to capture long-range emotional contexts. After experimentation, we found that encoding the full lyrics performed better than encoding individual verses. However, a caveat of obtaining RoBERTa's embeddings from the available HuggingFace implementation²⁰ limits the input to 512 characters, meaning that some lyrics had to be truncated.

As with audio, a Bayesian optimization strategy was employed using the Optuna library, with the same search spaces.

F. Bimodal Classical ML

We perform feature-level fusion to combine audio and lyrics in classical machine learning. The combined audio and lyrics features are fed to the ReliefF feature selection algorithm altogether, with the rest of the pipeline remaining unchanged.

One caveat that required special attention was the thousands of content-based features extracted from lyrics, which initially led to worse results when combined with the audio features. This was possibly due to the feature selection algorithm's inability to handle such high dimensionality. As such, in the developed bimodal classical machine learning approach, content-based lyrics features were discarded.

¹⁶https://github.com/parthenocissus/synesketch_v2.1/.

¹⁷<https://conceptnet.io/>.

¹⁸<https://www.liwc.app/>.

¹⁹<https://inquirer.sites.fas.harvard.edu/>.

²⁰Available at <https://huggingface.co/sentence-transformers/all-roberta-large-v1>.

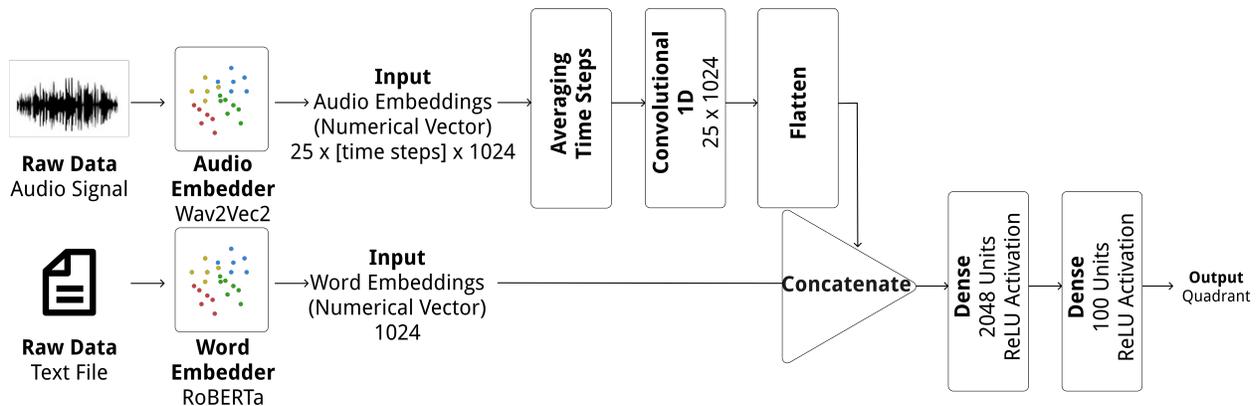


Fig. 2: Architecture for bimodal DL experimentation. Audio and word embeddings are obtained for each sample. The former’s output is further reduced before being concatenated with the latter. The concatenated vector is then classified with two dense layers.

G. Bimodal Deep Learning

As shown in Figure 2, we employ an early fusion strategy with the previously discussed DL methodologies, akin to the ML approach described in the previous subsection. Audio embeddings are reduced as described in Section V-C and concatenated with the corresponding word embeddings for the same sample. The prediction is obtained using a simple two-layer dense network.

The search spaces were mostly kept, except for the batch size, whose range was modified to [16, 128]. As observed in the lyrics DL experiments, the optimal batch size is relatively small compared to audio, so it made sense to adjust the corresponding range.

VI. RESULTS AND DISCUSSION

This section presents and discusses the results obtained for the audio, lyrics, and bimodal datasets after training models using the described baseline methodologies and evaluation strategies. For classification, F1-scores are presented; regarding regression, RMSE and R^2 scores are provided for arousal and valence, along with F1-scores obtained for classification based on regression (i.e., mapping the predicted AV values to the corresponding quadrant and measuring the resulting F1-score). As mentioned, we performed statistical significance tests in all comparisons, with a threshold set to $p < 0.05$. Additionally, we utilize 4QAED and LED as baseline datasets for comparison purposes with the new datasets.

In the following subsections, several tables are presented to summarize the attained results. In those tables, CV stands for 10x10-fold Cross Validation, TVT for Train-Validate-Test (using 70-15-15 and 40-30-30 splits), CML for Classical ML, HF for Handcrafted Features, nAE for Audio Embeddings, WE for Word Embeddings, and DNN for Dense Neural Network. Regarding TVT, we only present F1-scores for compactness since we obtained similar recall and precision values. In those tables, several results are highlighted with bold font.

Besides the summary result tables mentioned, we also show the confusion matrices for the main profile, i.e., MERGE Bimodal Complete with TVT 70-15-15. Moreover, since the

results obtained with the deep learning approach are typically higher than those obtained with the classical approach, we will present confusion matrices using the former.

A. MERGE Audio

Table III shows the overall results for the audio modality. Starting with a comparison between the baseline dataset (4QAED) and MERGE, similar results were obtained for the classical approach: 71.71% for 4QAED²¹ and around 71% for all MERGE datasets (in the cross-validation experiment). This suggests that the baseline and MERGE datasets have comparable complexity.

Comparing the classical ML approach (relying on handcrafted audio features) and the DL approach, the former attains a maximum F1-score of 71.03% (using the MERGE AC dataset with cross validation) and the latter tops at 70.84% (using the MERGE BC dataset with the 70-15-15 TVT split). Although these top scores are comparable, we observe that the classical approach outperforms the DL approach in most cases. Results are only similar in the MERGE Bimodal Complete Dataset. This suggests that the employed audio embeddings have room for improvement compared to state-of-the-art acoustic features, such as through fine-tuning.

Regarding the influence of the size and imbalance of the new datasets, these factors showed little impact, as the results obtained for the four datasets (MERGE AC and AB, MERGE BC and BB) are similar. This is observed across all modalities covered in this study (audio, lyrics, and bimodal), as shown in the following subsections.

As for the standard deviation of the F1-scores for 10x10-fold CV, we can observe that they are low (from 1.97% to 4.81%), which denotes low sensitivity to the defined folds. This is also observed in the lyrics and bimodal methodologies.

²¹ Although our implementation is identical to the original, we were unable to achieve the original 76.4% score. This is a consequence of updates to the underlying feature extraction frameworks, leading to different values for some extracted features. For the sake of fair comparison between 4QAED and the novel datasets, we decided to report the results obtained under the same conditions. We will address this issue in future work.

TABLE III: **Audio** Best Results

Dataset	Methodology	Approach	Cross Val	TVT 70-15-15			TVT 40-30-30	
			F1-score	F1-score	R ² (A/V)	RMSE (A/V)	F1-score	
4QAED	HF + SVM (CML)	C	71.71% ± 4.50	-	-	-	-	
Audio Complete	MERGE	HF + SVM (CML)	C	71.01% ± 2.31	70.12%	-	-	66.38%
		R	-	70.18%	0.479 / 0.364	0.281 / 0.375	-	
	AE + DNN (DL)	C	67.88% ± 3.04	68.90%	-	-	64.69%	
		R	-	66.71%	0.475 / 0.314	0.241 / 0.390	-	
Audio Balanced	MERGE	HF + SVM (CML)	C	70.91% ± 2.32	68.52%	-	-	68.20%
		R	-	67.47%	0.527 / 0.306	0.193 / 0.385	-	
	AE + DNN (DL)	C	67.12% ± 2.26	66.75%	-	-	67.79%	
		R	-	68.03%	0.538 / 0.359	0.171 / 0.362	-	
Bimodal Complete	MERGE	HF + SVM (CML)	C	71.03% ± 2.62	67.61%	-	-	69.22%
		R	-	63.65%	0.498 / 0.284	0.192 / 0.403	-	
	AE + DNN (DL)	C	70.57% ± 3.18	70.84%	-	-	67.97%	
		R	-	70.60%	0.529 / 0.379	0.165 / 0.353	-	
Bimodal Balanced	MERGE	HF + SVM (CML)	C	71.02% ± 2.79	66.13%	-	-	67.70%
		R	-	68.76%	0.522 / 0.360	0.182 / 0.351	-	
	AE + DNN (DL)	C	65.38% ± 4.81	66.81%	-	-	68.84%	
		R	-	68.04%	0.531 / 0.255	0.168 / 0.404	-	

Notes: HF = Handcrafted Features, AE = Audio Embeddings.

TABLE IV: **Audio** Confusion Matrix Results using the Main Profile (MERGE BC + TVT 70-15-15) with the DL approach)

		Predicted			
		Q1	Q2	Q3	Q4
Actual	Q1	76.3%	5.0%	5.0%	13.8%
	Q2	7.0%	93.0%	0.0%	0.0%
	Q3	6.9%	0.0%	55.5%	37.6%
	Q4	9.8%	2.0%	33.3%	54.9%

When compared with cross-validation (CV), TVT attains, in general, slightly lower but comparable results (for example, a top result of 71.03% in CV against 70.84% in TVT 70-15-15). This indicates the robustness of the proposed TVT splits and their feasibility for benchmarking, resulting in more straightforward and faster model training compared to cross-validation (CV). Comparing the two proposed splits, 70-15-15 outperforms 40-30-30 (a top F1-score of 70.84% in the former against 68.84% in the latter). This might result from the more extensive training set in the 70-15-15 split.

Comparing the F1-scores using direct classification approaches with the ones from regression-based classification, we observe that the results attained by the regression-based strategy are around 2-3% lower. This is to be expected considering the semi-automatic approach employed to obtain AV values for samples. However, this is not always the case, as results are slightly higher for the DL approach on the MERGE BC set, while the drop is only 1% in the BB set.

Regarding R² and RMSE values, as expected, valence performs significantly worse than arousal. Moreover, it fluctuates more across datasets, yielding R² intervals between 0.255 and 0.379, compared to 0.475 and 0.538 for arousal. RMSE are also consistently higher for valence in comparison to arousal. Given the fact that the reference AV values were obtained

automatically, we believe the obtained scores are acceptable.

Finally, the confusion matrix for the main profile is presented in Table IV. As can be observed, the model can accurately predict Q2, followed by Q1. However, despite our efforts to reduce ambiguity in the datasets, some confusion remains between Q3 and Q4, which leads to a lower score in these quadrants. This aligns with other studies in the literature that show the difficulty in distinguishing valence in low-arousal quadrants using only acoustical information [7].

B. MERGE Lyrics

Table V shows the overall results for the lyrics modality. Starting with a comparison between the baseline dataset (LED) and MERGE, contrary to audio, the results in the new datasets using the classical approach slightly underperform the ones attained in the baseline dataset (from 72.94% in LED²² to a maximum of 69.31% on MERGE BC). Despite the increased size, this suggests that the complexity of the novel lyrics datasets increased compared to LED.

Comparing the classical ML approach (relying on handcrafted text features) and the DL approach, the former attains a maximum F1-score of 71.31% and the latter tops at 75.05% (both using the MERGE BC dataset with the 70-15-15 TVT split). We observe that the DL approach outperforms the classical one in most cases. This suggests that the employed word embeddings can capture the emotional content of the lyrics more accurately than the handcrafted features. This is unsurprising, as these embeddings were trained on large amounts of text data. Additionally, fine-tuning the word embeddings may further improve the obtained F1-scores.

²²It is worth noting that the 73.6% F1-score reported in [13] was obtained on the LED2 771-lyrics subset; here, we performed 10x10-fold CV on the combined LED1+LED2 datasets (180 + 771 = 942 lyrics), hence, the slight differences.

TABLE V: Lyrics Best Results

Dataset	Methodology	Approach	Cross Val	TVT 70-15-15			TVT 40-30-30	
			F1-score	F1-score	R ² (A/V)	RMSE (A/V)	F1-score	
LED	HF + SVM (CML)	C	72.94% ± 4.42	-	-	-	-	
Lyrics Complete	MERGE	HF + SVM (CML)	C	67.46% ± 2.87	70.98%	-	-	64.95%
		R	-	59.18%	0.333 / 0.401	0.263 / 0.370	-	
	WE + DNN (DL)	C	68.82% ± 5.17	71.40%	-	-	74.45%	
		R	-	67.42%	0.366 / 0.569	0.205 / 0.298	-	
Lyrics Balanced	MERGE	HF + SVM (CML)	C	67.48% ± 2.49	69.25%	-	-	65.99%
		R	-	67.47%	0.358 / 0.387	0.193 / 0.385	-	
	WE + DNN (DL)	C	68.01% ± 3.79	74.15%	-	-	74.92%	
		R	-	65.33%	0.311 / 0.492	0.213 / 0.332	-	
Bimodal Complete	MERGE	HF + SVM (CML)	C	69.31% ± 3.27	71.31%	-	-	69.57%
		R	-	59.37%	0.350 / 0.389	0.255 / 0.391	-	
	WE + DNN (DL)	C	70.90% ± 2.86	75.05%	-	-	74.99%	
		R	-	69.42%	0.365 / 0.556	0.191 / 0.305	-	
Bimodal Balanced	MERGE	HF + SVM (CML)	C	66.96% ± 3.35	69.50%	-	-	66.74%
		R	-	57.60%	0.351 / 0.377	0.260 / 0.369	-	
	WE + DNN (DL)	C	69.34% ± 4.59	72.83%	-	-	72.84%	
		R	-	70.69%	0.350 / 0.552	0.196 / 0.299	-	

Notes: HF = Handcrafted Features, WE = Word Embeddings.

TABLE VI: Lyrics Confusion Matrix Results using the Main Profile (MERGE BC + TVT 70-15-15) with the DL approach)

		Predicted			
		Q1	Q2	Q3	Q4
Actual	Q1	76.5%	2.9%	5.9%	14.7%
	Q2	2.9%	87.3%	5.9%	3.9%
	Q3	2.8%	6.9%	76.4%	13.9%
	Q4	25.6%	2.2%	13.3%	58.9%

Additionally, we observe that higher scores were obtained in the lyrics-only experiment compared to the audio-only experiment (a maximum of 75.05% for lyrics, against 70.84% for audio, using the main profile). This may suggest a higher maturity level of the word embeddings in comparison to audio embeddings, which was expected.

As before, the new datasets' size and imbalance had little impact. Once again, the results attained for the four datasets (lyrics complete and balanced, bimodal complete and balanced) are similar.

Regarding the standard deviation of the F1-scores for 10x10-fold CV, we can again observe a reasonably low sensitivity to the data folds (from 2.49% to 4.59%).

When compared with CV, TVT consistently attains higher scores (for example, a top result of 75.05% in TVT against 71.31% in CV), indicating its robustness. Once again, the 70-15-15 split outperforms the 40-30-30 split, although in a less notorious way (a top F1-score of 75.05% in the former against 74.99% in the latter).

Continuing the trend observed in the audio modality, regression-based classification methodologies underperformed compared to direct classification approaches. The difference is most noticeable in the classical approaches, which range from 10% to 13% lower F1-score. Embeddings-based methodolo-

gies appear more robust, showing at most a 9% lower F1-score and 2% at best.

In the R² metric, as expected for lyrics, valence outperformed arousal prediction. Interestingly, the classical approach yields very similar results for arousal and valence in terms of R². However, the fact that the RMS error was also higher seems contradictory. This may indicate that the present model does not accurately predict samples outside the mean of the data. As for variability, the results do not vary as much as in audio for either axis.

Finally, the confusion matrix for the main profile is presented in Table VI. As can be observed, the model can accurately predict Q2, followed by Q1 and Q4. Compared to audio, the F1-scores obtained for Q1 are comparable, while the ones for Q2 are lower (though still high, i.e., 93.0% against 87.3%). Regarding Q3, we observe a significant improvement (from 55.5% in audio to 76.4% in lyrics), although some confusion between Q3 and Q4 still remains. As for Q4, results are only slightly higher compared to audio (58.9% against 54.9%). However, while in audio the main confusion came from Q4 and Q3, in lyrics the main observed confusion stemmed from Q4 and Q1). This was expected, given the theoretical knowledge that lyrics are better at capturing valence, while audio is superior for arousal [13].

C. MERGE Bimodal

Regarding the experiments using the bimodal datasets, Table VII summarizes the results achieved.

When comparing the bimodal, audio-only, and lyrics-only approaches, results in the novel datasets for the classical approach show that the bimodal strategy significantly outperforms the best methods from the isolated modalities, as expected: the bimodal model attained a maximum F1-score of 78.58%, against 71.03% for audio and 69.31% for lyrics, all on MERGE BC. The same happens for DL approaches, where

TABLE VII: **Bimodal Best Results**

Dataset	Methodology	Approach	Cross Val	TVT 70-15-15		TVT 40-30-30	
			F1-score	F1-score	R ² (A/V)	RMSE (A/V)	F1-score
Bimodal Complete	HF + SVM (CML)	C	78.58% ± 2.47	77.98%	-	-	75.90%
		R	-	62.04%	0.320 / 0.433	0.256 / 0.377	-
	AE + WE + DNN (DL)	C	81.74% ± 2.69	80.53%	-	-	78.69%
		R	-	74.58%	0.505 / 0.528	0.169 / 0.305	-
Bimodal Balanced	HF + SVM (CML)	C	77.34% ± 2.41	77.28%	-	-	75.18%
		R	-	71.15%	0.545 / 0.412	0.189 / 0.320	-
	AE + WE + DNN (DL)	C	79.42% ± 2.72	75.75%	-	-	77.97%
		R	-	75.64%	0.484 / 0.508	0.176 / 0.310	-

Notes: HF = Handcrafted Features, AE = Audio Embeddings, WE = Word Embeddings.

TABLE VIII: **Bimodal Confusion Matrix Results using the Main Profile (MERGE BC + TVT 70-15-15) with the DL approach)**

		Predicted			
		Q1	Q2	Q3	Q4
Actual	Q1	79.3%	3.7%	3.7%	13.4%
	Q2	3.1%	95.9%	0.0%	1.0%
	Q3	1.4%	1.4%	80.0%	17.2%
	Q4	13.3%	1.2%	21.7%	63.9%

the bimodal methodology reached 81.74% (the top overall result achieved in all the experiments conducted in this study), against 75.05% for lyrics-only and 70.84% for audio-only.

When comparing TVT splits, as before, the 70-15-15 split outperforms the 40-30-30 split. An F1-score of 80.53% was achieved with 70-15-15 against 78.69% in 40-30-30. When compared with CV, TVT attains again comparable results.

The obtained results for regression-based classification are overall lower, as expected. Again, deep learning approaches appear more robust than the classical counterparts, with the largest difference from the direct classification approach being 6% F1-score on MERGE BC set when compared to the 16% difference of the classical methodology on the MERGE BC.

The best results for R² and RMSE are observed in this modality. The highest attained R² score was 0.545 for arousal (on the MERGE BB dataset) and 0.528 for valence (on MERGE BC), respectively. The smallest error observed for both axes in terms of RMSE is also observed on MERGE BC, achieving 0.169 for arousal and 0.305 for valence.

Finally, Table VIII contains the confusion matrix for the main profile. As can be observed, the scores increased for all quadrants compared to the audio-only and lyrics-only solutions. Compared to audio, a particularly noticeable improvement was obtained for Q3 (from 55.5% to 80.0%). For Q4, the increase was not so dramatic, although still significant (from 54.9% to 63.9%). Compared to lyrics, the improvements in these quadrants were not so notable, but were still very relevant (3.6% and 5% improvements in Q3 and Q4, respectively). This reinforces the conclusion that most of the improvement in the classification of the lower arousal quadrants is due to the lyrics.

The previous results confirm the potential of bimodal approaches to reduce the confusion between low-arousal quadrants. Yet, the attained results show that there is plenty of room

for improvement and that the separation between Q4 and Q3, and Q4 and Q1, is far from being solved [7].

Due to the current dataset sizes, the DL-based methods used in this work have not yet fully utilized the potential of deep learning. Although current DL methodologies in the literature open many exciting research paths, the need for extensive data remains an issue to be addressed. In this respect, a preliminary study [30] shows the promise of hybrid approaches. The combination of handcrafted features with deep neural networks outperformed traditional feature engineering and machine learning methods. Therefore, despite its (still) limited size, the MERGE dataset is a step toward unlocking the potential of deep learning solutions for MER.

VII. CONCLUSION

This article proposed a new bimodal audio-lyrics dataset. For each modality, both a complete and balanced variation are available. Two TVT splits were created and released alongside these datasets to enable fast experimentation and guarantee uniformity for all research works that employ them.

To validate the proposed datasets and data splits, we performed experiments using classical approaches (based on handcrafted features and standard ML algorithms) and DL methodologies (based on audio and word embeddings).

Based on the obtained results, we conclude that the proposed datasets (along with the related semi-automatic creation protocol) and TVT data splits are viable for MER benchmarking. The methods employed provide a solid baseline for comparison with future works using the MERGE dataset.

Among the proposed subsets, MERGE Bimodal complete typically led to the highest F1-scores. Also, the proposed data splits, especially the 70-15-15 strategy, are well-suited for optimizing and quickly validating MER systems. For this reason, we propose the MERGE BC dataset and the 70-15-15 TVT split as the *main profile* for future benchmarking.

Additionally, the proposed datasets are designed for various research purposes. In addition to emotion quadrant and AV annotations, the datasets also include metadata like genre, artist, album, year, and complete emotion tags. These features could benefit a wide range of MIR research and advanced MER tasks, including multi-label emotion classification.

This study addresses a critical need in the research area, specifically the bimodal dataset, which is its main contribution.

To the best of our knowledge, this is the largest publicly available and quality-controlled MER bimodal dataset.

Nevertheless, its size remains a limitation to fully exploiting the potential of deep learning solutions.

Another limitation is that the annotations do not account for temporal variability in emotion. Being a dataset designed for static MER, it is not suited for MEVD studies.

In addition, English is the only language supported in the dataset. In the future, we plan to extend the MERGE dataset to support multiple languages, which could be automatically translated into English using the current breakthroughs in this field, particularly Transformers.

ACKNOWLEDGMENTS

This work is funded by FCT - Foundation for Science and Technology, I.P., within the scope of the projects: MERGE - DOI: 10.54499/PTDC/CCI-COM/3171/2021 financed with national funds (PIDDAC) via the Portuguese State Budget; and project CISUC - UID/CEC/00326/2020 with funds from the European Social Fund, through the Regional Operational Program Centro 2020. Renato Panda was supported by Ci2 - FCT UIDP/05567/2020.

REFERENCES

[1] D. Huron, "Perceptual and cognitive applications in music information retrieval," *Perception*, vol. 10, no. 1, pp. 83–92, 2000.

[2] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition*, 1st ed. USA: CRC Press, Inc., 2011.

[3] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, vol. 5, no. 1_suppl, pp. 123–147, 2001. [Online]. Available: <https://doi.org/10.1177/10298649020050S105>

[4] M. Zhang, Y. Zhu, W. Zhang, Y. Zhu, and T. Feng, "Modularized composite attention network for continuous music emotion recognition," *Multimedia Tools Appl.*, vol. 82, no. 5, p. 7319–7341, Aug. 2022. [Online]. Available: <https://doi.org/10.1007/s11042-022-13577-6>

[5] P. Paukner, M. Ripoll, D. Sabir, D. O. Erdoğan, L. Sacchetto, and K. Diepold, "Classifying music-induced emotion using multi-modal ensembles of eeg and audio feature models," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[6] E. J. Oh, H. Kim, and K. M. Lee, "Which audio features can predict the dynamic musical emotions of both composers and listeners?" in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*. ISMIR, 2024, pp. 352–359.

[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] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-based approach towards music emotion recognition from lyrics," in *Proceedings of Advances in Information Retrieval - 43rd European Conference on IR Research*, vol. 2, 2021, pp. 167–175.

[9] R. Orjeseck, R. Jarina, and M. Chmulik, "End-to-end music emotion variation detection using iteratively reconstructed deep features," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5017–5031, Feb. 2022.

[10] Q. Hu, M. A. A. Murad, and Q. Li, "Advancing music emotion recognition: Large-scale dataset construction and evaluator impact analysis," *Multimedia Systems*, vol. 31, no. 2, p. 123, Feb. 2025.

[11] P. Donnelly and S. Gaur, "Mood dynamic playlist: Interpolating a musical path between emotions using a knn algorithm," in *Human Interaction & Emerging Technologies (IHJET-AI 2022): Artificial Intelligence & Future Applications*, ser. AHFE (2022) International Conference. AHFE Open Access, T. Ahram and R. Taiar, Eds., vol. 23. AHFE International, USA, 2022. [Online]. Available: <http://doi.org/10.54941/ahfe100894>

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

[13] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, "Emotionally-Relevant Features for Classification and Regression of Music Lyrics," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 240–254, Apr. 2018.

[14] P. N. Juslin and P. L. and, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of New Music Research*, vol. 33, no. 3, pp. 217–238, 2004.

[15] M. Besson, F. Faita, I. Peretz, A.-M. Bonnel, and J. Requin, "Singing in the brain: Independence of lyrics and tunes," *Psychological Science*, vol. 9, no. 6, pp. 494–498, 1998.

[16] X. Hu, "Improving Music Mood Classification Using Lyrics, Audio and Social Tags," Ph.D. dissertation, University of Illinois, Urbana, Illinois, 2010.

[17] 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>

[18] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 Mirex Audio Mood Classification Task: Lessons Learned," in *Proceedings of the 9th International Society for Music Information Retrieval Conference*, Drexel University, Philadelphia, Pennsylvania, USA, 2008, pp. 462–467.

[19] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell, "On the importance of both dimensional and discrete models of emotion," *Behavioral sciences (Basel, Switzerland)*, vol. 7, no. 4, p. 66, 2017.

[20] P. E. and, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[21] R. PLUTCHIK, "Chapter 1 - a general psychoevolutionary theory of emotion," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. Academic Press, 1980, pp. 3–33. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780125587013500077>

[22] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *The American Journal of Psychology*, vol. 48, no. 2, p. 246, Apr. 1936.

[23] G. A.V., M. T., P. D., and U. E., "Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions," *Information Fusion*, vol. 105, p. 102218, 2024.

[24] X. Hu, "Improving music mood classification using lyrics, audio and social tags," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[25] E. Schubert, "Update of the hevner adjective checklist," *Perceptual and Motor Skills*, vol. 96, no. 3_suppl, pp. 1117–1122, 2003, pMID: 12929763. [Online]. Available: <https://doi.org/10.2466/pms.2003.96.3c.1117>

[26] X. Hu and J. Downie, "Exploring mood metadata: Relationships with genre, artist and usage metadata." in *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*, 01 2007, pp. 67–72.

[27] 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, Feb. 2008.

[28] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*. New York, NY, US: Oxford University Press, 2001.

[29] X. Wu, J. Wang, J. Yu, T. Zhang, and K. Zhang, "Popular hooks: A multimodal dataset of musical hooks for music understanding and generation," in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2024, pp. 1–6.

[30] 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, p. 2201, 2024.

[31] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music Emotion Recognition: A State of the Art Review," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Porto, Portugal, 2010, pp. 225–266.

[32] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami, Florida, USA, 2011, pp. 591–596.

[33] E. Çano and M. Morisio, "MoodyLyrics: A Sentiment Annotated Lyrics Dataset," in *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*. Hong Kong Hong Kong: ACM, Mar. 2017, pp. 118–124.

[34] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLOS ONE*, vol. 12, no. 3, p. e0173392, Mar. 2017.

- [35] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*, 2nd ed. Sage, Thousand Oaks, CA, 2004.
- [36] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Technical report C-1, the center for research in psychophysiology . . . , Tech. Rep., 1999.
- [37] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, Dec. 2013.
- [38] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 135–142. [Online]. Available: <https://doi.org/10.1145/3206025.3206037>
- [39] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of Algorithms Using Games: The Case of Music Tagging," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, 2009, pp. 387–392.
- [40] K. Choi, G. Fazekas, and M. Sandler, "Automatic Tagging Using Deep Convolutional Neural Networks," in *Proceedings of the 17th Int. Society for Music Information Retrieval Conf. (ISMIR)*, New York City, United States, 2016, pp. 805–811.
- [41] R. Panda, R. Malheiro, and R. P. Paiva, "Musical Texture and Expressivity Features for Music Emotion Recognition," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 383–391.
- [42] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," 2021. [Online]. Available: <https://arxiv.org/abs/2108.01374>
- [43] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. [Online]. Available: <http://hdl.handle.net/10230/42015>
- [44] C. Akiki and M. Burghardt, "Muse: The musical sentiment dataset," *Journal of Open Humanities Data*, Jul 2021.
- [45] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal Music Mood Classification Using Audio and Lyrics," in *2008 Seventh International Conference on Machine Learning and Applications*, Dec. 2008, pp. 688–693.
- [46] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Mousallam, "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.
- [47] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues *et al.*, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2020, pp. 399–404.
- [48] H. Kim, K. Choi, M. Modrzejewski, and C. C. S. Liem, "The biased journey of MSD_AUDIO.ZIP," 2023.
- [49] D. Bogdanov, X. Lizarraga-Seijas, P. Alonso-Jiménez, and X. Serra, "Musav: A dataset of relative arousal-valence annotations for validation of audio models," in *International Society for Music Information Retrieval Conference (ISMIR 2022)*, Bengaluru, India, 2022. [Online]. Available: <http://hdl.handle.net/10230/54181>
- [50] E. Çano and M. Morisio, "Music mood dataset creation based on last.fm tags," pp. 15–26, 2017. [Online]. Available: <http://porto.polito.it/2669975/>
- [51] R. Panda, H. Redinho, C. Gonçalves, R. Malheiro, and R. P. Paiva, "How does the spotify api compare to the music emotion recognition state-of-the-art?" in *Proceedings of the 18th sound and music computing conference (smc 2021)*. Axea sas/SMC Network, 2021, pp. 238–245.
- [52] Y. Wang, X. Zhao, C. Deng, Y. Xiao, H. Ruan, P. Jin, and X. Cai, "A survey on music emotion recognition using learning models," *Multimedia Systems*, vol. 31, no. 4, p. 278, Jun 2025. [Online]. Available: <https://doi.org/10.1007/s00530-025-01871-w>
- [53] A. Baeviski, 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.
- [54] K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Multi-Modal Song Mood Detection with Deep Learning," *Sensors*, vol. 22, no. 3, p. 1065, 2022.
- [55] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019.



Pedro Lima Louro is a PhD Research Student at the University of Coimbra, where he also concluded his Masters degree, specializing in Intelligent Systems. He is a member of the Cognitive and Media Systems (CMS) research group at the Centre for Informatics and Systems of the University of Coimbra (CISUC). His main research interests include Music Information Retrieval (MIR), Music Emotion Recognition (MER), and Deep Learning.



Hugo Redinho is an MSc from the University of Coimbra, specializing in Intelligent Systems, where he also concluded his Bachelor's degree in Informatics Engineering. He is a member of the Music Information Retrieval (MIR) research team at the Center for Informatics and Systems of the University of Coimbra (CISUC). His main research interests are related to Music Emotion Recognition (MER) and MIR.



Tiago Filipe Rodrigues Ribeiro is a PhD Research Student at CISUC. His research focuses on applying machine learning to analyze and generate emotions in music, with a particular emphasis on lyrics-based methods. He holds a Master's degree in Data Science and a Bachelor's degree in Electrical and Computer Engineering from the Polytechnic of Leiria, Portugal. He is particularly interested in the intersection of technology and humanities.



Ricardo Santos is a PhD Research Student at the Centre for Informatics and Systems of the University of Coimbra (CISUC). His research interests include Music Emotion Recognition, Deep Learning, and Large Language Models. Santos received his Master's in Computer Engineering from IPT/USP, Brazil.



Ricardo Malheiro is a PhD from the University of Coimbra, where he also concluded his Master and Bachelor (Licenciatura - 5 years) degrees, respectively in Informatics Engineering and Mathematics. He is a Professor at the Polytechnic Institute of Leiria - School of Technology and Management, and a member of the CMS research group at CISUC. His main research interests are Natural Language Processing, Detection of Emotions in Music Lyrics and Text, and Text/Data Mining.



Renato Panda is an Assistant Researcher at Ci2, Polytechnic Institute of Tomar, Portugal. His main research interests are Music Emotion Recognition (MER) and Music Information Retrieval (MIR), as well as Applied Machine Learning and Software Engineering. He earned his PhD in Informatics Engineering from the University of Coimbra in 2019. Since then, he has been a member of the Cognitive and Media Systems group at CISUC, where he remains actively involved.



Rui Pedro Paiva is a Professor at the Department of Informatics Engineering of the University of Coimbra, where he concluded his Doctoral, Master and Bachelor degrees in 2007, 1999 and 1996, respectively. He is also a member of the CMS group at CISUC. His main research interests are in the areas of MIR and Health Informatics. The common research hat is the study of feature engineering, machine learning, and signal processing to analyze musical and bio signals.