UTILIZING LISTENER-PROVIDED TAGS FOR MUSIC EMOTION RECOGNITION: A DATA-DRIVEN APPROACH

J. Affolter, M. Rohrmeier
Ecole Polytechnique Fédérale de Lausanne, EPFL

ABSTRACT

This work introduces a data-driven approach for assigning emotions to music tracks. Consisting of two distinct phases, our framework enables the creation of synthetic emotion-labeled datasets that can serve both Music Emotion Recognition and Auto-Tagging tasks. The first phase presents a versatile method for collecting listener-generated verbal data, such as tags and playlist names, from multiple online sources on a large scale. We compiled a dataset of $5,892$ tracks, each associated with textual data from four distinct sources. The second phase leverages Natural Language Processing for representing music-evoked emotions, relying solely on the data acquired during the first phase. By semantically matching user-generated text to a well-known corpus of emotion-labelled English words, we are ultimately able to represent each music track as an 8-dimensional vector that captures the emotions perceived by listeners. Our method departs from conventional labeling techniques: instead of defining emotions as generic "mood tags" found on social platforms, we leverage a refined psychological model drawn from Plutchik's theory [1], which appears more intuitive than the extensively used Valence-Arousal model.

1. INTRODUCTION

Several studies on music listener behavior have identified an increasing interest in music discovery based on its emotional content [2]. It is therefore hardly surprising that the field of Music Emotion Recognition (MER), which explores how emotions can be identified in music [3], is a growing area of research.

MER research is dominated by the use of supervised machine learning methods, in which systems are trained on music excerpts previously labeled with emotion descriptors through crowdsourcing. A major hurdle in this field is the lack of large-scale emotion-annotated datasets [4]. The complexity of collecting suitable training data contributes significantly to this issue, as the process is time-consuming, labor-intensive and expensive. The subjective nature of musical emotions further complicates the data collection process [5].

Recognizing language as a powerful medium for conveying musical signification, we proceed on the premise that emotions can be inferred from textual data—specifically, from listener-generated tags and playlist names on music platforms. We thus introduce a novel method for assigning, to any given song, an emotion vector within an 8-dimensional space defined by Plutchik's model. This enables us to propose a new dataset comprising $5,892$ tracks, specifically tailored for Music Emotion Recognition (MER) tasks.

2. RELATED WORK

Yuan et al. [4] propose the Music Audio Representation Benchmark for universaL Evaluation (*MARBLE*) as a unified standard for assessing various Music Information Retrieval (MIR) tasks. They employ 12 publicly available datasets to evaluate 18 distinct tasks, including the *Emomusic* [5] and *MTG-MoodTheme* [6] datasets for MER evaluation. Table 1 provides an overview of commonly used datasets in MER research, along with their size, data collection method and emotion labeling approach.

Dataset	Size	Data collection	Emotion model	Ref.
Emomusic	744	C	AV	[5]
MTG-MT	17,982	DM	56 labels	[6]
AMC	600	C	5 clusters	[7]
EMMA	364	C	GEMS	[8]
CAL500	500	C	174 labels	[9]
MoodSwings	240	Game	AV	[10]
NTWICM	2,648	C	AV	[11]
Soundtracks	470	C	9 labels	[12]
DEAP	120	EEG	AVD	[13]
AMG1608	1,608	C	AV	[14]
Emotify	400	Game	GEMS	[15]
Moodo	200	C	AV	[16]
4Q-emotion	900	C	AV	[17]
PMEmo	794	EEG	AV	[18]

Table 1: Overview of existing MER datasets.
C: crowdsourcing, DM: data mining, AV(D): arousal/valence/dominance, EEG: electroencephalography, GEMS: Geneva Emotion Music Scale

Through the examination of these datasets, three areas for potential improvement have been identified.

Dataset size. Datasets annotated with labels according to a psychological emotion model (AV, AVD, GEMS) do not exceed $2,648$ tracks, with an average size of 801. Furthermore, most datasets fail to cover a wide range of

Attribution: J. Affolter, M. Rohrmeier, "Utilizing Listener-Provided Tags for Music Emotion Recognition: A Data-Driven Approach", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

musical genres—they are often limited to four or fewer, or do not provide clear genre definition, resulting in imbalanced datasets. This limited size and diversity complicate the training of accurate music emotion recognition models, raising concerns about issues such as group fairness and generalization capability.

Data collection. Most datasets rely on human annotations from crowdsourcing/online games, or from EEG experiments, which, while reliable, are both expensive and time-consuming. Moreover, these datasets encounter challenges in participant diversity. Typically, the assignment of an emotion label to a track requires consensus between few annotators. In the case of datasets featuring mood tags, it is common for tracks to have, on average, no more than two tags associated with them, potentially leading to misleading data.

Emotion model. Emotion labeling generally falls into two categories. (1) Mood-based emotion tags. For instance, in *MTG-MoodTheme*, the 56 emotion labels correspond to tags directly retrieved from the Jamendo music platform. This can result in a large number of emotion labels, making it difficult for end-users to understand and use the system. This approach may also not align with an established emotion model. (2) Discrete- or continuous-based annotations derived from a predefined emotion model. While the VA model, with its two-dimensional structure, has been criticized to be restrictive and open to overly subjective interpretation [19], the Geneva Emotion Music Scale (GEMS) is specifically crafted for the music domain, and proposes a more detailed taxonomy.

3. APPROACH

This section introduces the key design decisions underlying our methodology for inducing music-evoked emotion descriptors from a collection of tracks. Our approach aims to enhance the study of emotions in music by introducing a novel representation of emotions based on a psychological model that has been hitherto unacknowledged in the field of MER.

3.1 Plutchik's Emotion Model

We recognize the importance of grounding our research framework in a well-established emotion model. In search of a more intuitive alternative than the Valence/Arousal (VA) framework, we opted for Plutchik's model, which, to our knowledge, has not yet been utilized in the field of music and, we believe, strikes a good balance between complexity and usability. Plutchik's emotion model is founded on eight primary emotions (joy, fear, anger, sadness, disgust, surprise, anticipation, trust) that we believe are accessible and instinctive for listeners. As a recognized model in psychology, it has been employed across various domains beyond music, enabling us to leverage existing resources, such as the *NRC Lexicon* [20], a crowdsourced list of $14,182$ English words and their binary associations with Plutchik's primary emotions. Highly aligned with our research goal, its single-word structure bears a strong resemblance to our textual data, which includes tags and playlist names. Its origin in actual annotations by human subjects, rather than derivative interpretations, is also crucial to the accuracy of our emotion mappings. Furthermore, the model's categorical approach can be expanded by combining emotions as depicted in Figure 1, thus enabling the representation of more complex emotions.

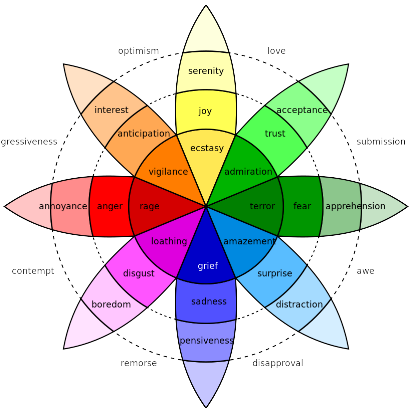

Figure 1: Plutchik's wheel of emotions [1].

3.2 Emotion Vector Representation

Instead of viewing emotions as discrete labels, as traditional classification MER systems do, we propose to represent a music track as an 8-dimensional emotion vector that captures the emotions perceived by listeners. We define our emotion domain mathematically as a vector space V, represented by a basis B={joy, fear, anger, sadness, disgust, surprise, anticipation, trust}. Each emotion in B corresponds to a standard basis vector in V, with $e_{joy} = [1, 0, ..., 0]$ through $e_{trust} = [0, ..., 0, 1]$. The emotion vector v of a track is defined as

$$v = \sum_{i \in B} \lambda_i e_i, \qquad (1)$$

where $\lambda_i \in [0, 1]$ represents the intensity of emotion i.

By representing emotion intensities within an 8-dimensional vector, our framework aims to effectively capture the complex spectra of music-evoked emotions and discern the subtle emotional nuances of music tracks.

3.3 Textual Data Encoding

To effectively encode textual data, we selected the Sentence-BERT (SBERT) model [21], an NLP neural network known for generating semantically meaningful embeddings at the sentence level. Specifically fine-tuned for semantic similarity tasks, SBERT enhances the original BERT architecture by integrating siamese and triplet network structures. NLP techniques such as Semantic Search are significantly enhanced with this encoding model, as it enables the retrieval of the closest elements in the embedding space based on semantic similarity.

By using SBERT, we are able to create single embeddings that accurately encode the semantic content of each textual element while preserving context. This choice is particularly suitable for our data, which includes short phrases—such as single- and multi-word tags, playlist names, and English words from the *NRC Lexicon*—that need to be compared in terms of semantic similarity.

4. IMPLEMENTATION: A TWO-STAGE FRAMEWORK FOR EMOTION ATTRIBUTION

Building upon the challenges and insights discussed in Section 2, we introduce a two-stage framework for extracting music-evoked emotions from a collection of tracks. Drawing inspiration from *MTG-MoodTheme*, the first phase focuses on collecting verbal tags through data mining across platforms such as *Last.fm* and *Spotify*, while the second phase leverages NLP techniques to computationally associate emotion vectors with music tracks by relying solely on the tags acquired during phase one.

4.1 Large-Scale Listener-Generated Data Collection

4.1.1 Track Selection Process.

We started with a baseline dataset of $20,000$ music tracks, spanning 20 distinct genres. We selected the top $1,000$ tracks with the highest popularity index on *Spotify* for each genre, in order to increase the likelihood of finding them in multiple sources when retrieving tags.

4.1.2 Data Mining.

We extracted listener-generated tags from three popular rating websites—Last.fm, AllMusic, Rate Your Music [1] —and retrieved playlist names from the dataset provided for the Spotify Million Playlist Dataset Challenge, which includes $1,000,000$ playlists created by Spotify listeners between 2010 and 2017 [22].

4.1.3 Data Pre-Processing.

While the tags from *Rate Your Music* and *All Music* were already normalized by the platform, those from *Spotify* and *Last.FM* required extensive cleaning. The objective was twofold: first, to eliminate irrelevant data, such as playlist names along the lines of '*Favorite hits*', and second, to remove tags that could introduce bias when assigning emotions. Indeed, some tags—such as album names, artist names, or musical genres—are intended as mere filters for finding music. Others, like 'roadtrip tunes', are too neutral and may suggest contexts unrelated to emotions, while tags such as 'love it' reflect personal opinions and could bias our results by conflating perceived with induced emotions [23].

We first translated multilingual text into English, expanded abbreviations, replaced slang words and emoticons with their standard equivalents, and corrected misspelled words. We then implemented four iterative filtering processes to eliminate listener-generated tags that cannot be considered emotion descriptors.

Metadata Filtering. Since artists, song titles, album names, and musical genres were retrieved as metadata for all tracks in our dataset, we first eliminated any textual inputs containing terms from these categories. The set of musical genres was expanded to include a broader range beyond the 20 genres under study.

Named Entities Filtering. We then used the pre-trained BERT model fine-tuned for Named Entity Recognition (NER) [2] to identify named entities within predefined categories, such as person names, song titles, and locations. We filtered out sequences containing at least one token classified as a named entity of a target category with a confidence score above 0.9.

Neutral Tag Filtering. Sentiment analysis was subsequently performed using the pre-trained RoBERTa model fine-tuned for this task [3] . We removed tags with a neutral sentiment proportion greater than 70% (where 100% was distributed among positive, neutral, and negative sentiments for each input sequence). This threshold was deliberately chosen to avoid losing potentially useful tags like 'energetic'. In subsequent stages of this framework, tags that are too neutral and not intended for emotion description will nonetheless be matched with words from the *NRC Lexicon* that do not have associated emotions, thereby not impacting the final emotions associated with music tracks.

Listener Judgment Filtering. Finally, we eliminated tags closely tied to listener preferences and judgments. Briefly put, we established predefined categories specifically designed to capture tags for exclusion, based on their semantic content. For example, we defined a category titled '*This track is great*' and tags like '*Love it!*' would semantically align with this category and be filtered out. To do so, we computed sentence-level embeddings for both the tags and the categories (augmented by the *NRC Lexicon*) using the SBERT model to capture their semantic content. We then matched each tag to its closest category using cosine similarity on their embeddings, removing tags that fell into any unwanted category.

4.2 Emotion Vector Attribution

The second phase of our approach relies on the *NRC Lexicon* to computationally associate emotion vectors with music tracks by relying solely on the acquired tags. We decided to represent words from the Lexicon as vectors w within the Plutchik emotion space, where $w = \sum_{i \in B} c_i e_i$ and c_i is a binary indicator denoting the absence or presence of the corresponding emotion.

Given that tags are assigned by individual listeners on music platforms, we can treat them as independent entities. This assumption enables us to first assign emotions to each unique tag in the dataset, and then derive the emotion vector of a track by combining the emotions of its associated tags.

[1] https://www.last.fm/, https://www.allmusic.com/, https://rateyourmusic.com/

[2] https://huggingface.co/dslim/bert-base-NER

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

4.2.1 Assigning Emotions to Individual Tags

To infer the emotions connoted by a given tag, we first input both the tag and words from the *NRC Lexicon* into the **SBERT model**. This model generates embeddings for each, representing both the tag and words in the same semantic space.

We then perform **semantic search**, which involves retrieving the top-k entries $\{y_i\}_{i=1,\dots,k} \in Y$ of a corpus (NRC Lexicon) that are closest to a query x (the tag) by maximizing the cosine similarity on their embeddings, effectively identifying the words that are semantically similar to the tag: $y_i \in \arg\max_{y \in Y} \frac{x \cdot y}{\|x\|\|y\|}$.

Finally, a **weighted majority vote** is performed. This method involves directly selecting emotion vectors when a match with a high similarity score is found. If no such match exists, emotions with the highest consensus among a broader set are chosen.

Algorithm 1 Weighted Majority Vote

```
Input: Hyperparameters $\alpha_1, \alpha_2, \alpha_3, \beta$; tag embedding x; embeddings, similarity scores, and emotion vectors from the top-k matches $\{(x_i, s_i, w_i) \mid s_u \geq s_v$ when $u < v\}_{i=1,\dots,k}$;
Output: Emotion vector of the tag $v \in R^8$
Initialize the set of chosen matches : $m \leftarrow \emptyset$.
if $s_1 \geq \alpha_1$ then
    $m \leftarrow \{w_1\}$
else
    $m \leftarrow \{w_i \mid s_i \geq \alpha_2\}$
    if $m = \emptyset$ then
        $m \leftarrow \{w_i \mid s_i \geq \alpha_3\}$
$\mu \leftarrow \sum_{w_i \in m} s_i w_i \in R^8$
$v \leftarrow (v_i)_i$ where $v_i = 1$ if $\mu_i > \beta$, 0 otherwise
Return $v$
```

We conducted hyperparameter tuning using Grid Search to optimize the parameters α_1, α_2, α_3, β and k with the evaluation method outlined in Section 5.2. By selecting hyperparameters that maximize the F1-score between the original and inferred vectors, we ensured optimal accuracy in identifying the correct emotion vectors from words in the lexicon, considering both false positives and false negatives. The optimal values obtained were $0.95, 0.9, 0.5, 0.5$, and 7, respectively.

4.2.2 Deriving the Emotion Vector for Each Track

Now that each tag is assigned an emotion vector with binary values indicating the presence or absence of each primary emotion, we can derive the emotion vector for each track. However, two issues must be addressed first. Tag occurrences should be normalized to ensure comparability across different sources; and intersubjective variability in music perception should be accounted for, since it can lead to differing tags among listeners and misleading inferred emotion vectors.

Tag Occurrences Normalization. We divide each occurrence by the maximum occurrence encountered within the source, resulting in normalized occurrences within the [0,1] range. For tags from *Rate Your Music*, where occurrences were not provided, we set their count to the average occurrence at the track level.

Tag Selection for Inter-rater Agreement. For each track, we select tags that exhibit good inter-rater agreement, estimated using the Intra-class Correlation Coefficient (ICC) with one-way random effects for absolute agreement [24]—a widely used metric for assessing inter-rater reliability when the same set of raters evaluates all subjects. In our approach, each emotion is treated as an individual "subject" and each tag as a "rater". The emotion vectors of each tag, weighted by their normalized occurrences, serve as ratings for the respective emotion.

To attain the acceptable threshold of 0.75 for inter-rater agreement (values between 0.75 and 0.90 indicate good reliability, according to [24]), we perform backward selection to iteratively eliminate conflicting tags. Starting with the initial set of tags for a given track, we remove the tag whose exclusion results in the highest ICC score. This process continues until the threshold is attained or only two tags remain.

Track Emotion Vector. We derive the emotion vector of a track by calculating the weighted average of the emotion vectors v_i from the p tags that demonstrated good inter-rater agreement. The weights α_i are set to the tags' normalized occurrences, thus giving more importance to emotions from prevalent tags.

$$v = \frac{1}{\sum_{i=1}^{p} \alpha_i} \sum_{i=1}^{p} \alpha_i w_i = \sum_{j \in B} \lambda_j e_j, \qquad (2)$$

5. EVALUATION

5.1 Tag Extraction Method

To assess the reliability of our tag extraction method, we compared our tags to *human-generated* annotations from two crowdsourced MER datasets: *AMC Mirex* [7] and *Cal500* [9]. These datasets were selected for being the only ones to include emotion tags and share common tracks with our collection.

First, we calculate the percentage of common tags at the track level between each dataset and ours. Next, to assess the alignment between emotion tags, we derive emotion vectors of tracks from the two crowdsourced datasets using our method for emotion vector attribution (see Section 4.2), and then compare them with ours using semantic similarity.

	AMC		**Cal500**	
	mean	med.	mean	med.
Percentage of common tags for each track	**56.0**	**100.0**	3.24	0.0
Similarity score between emotion vectors	0.68	0.78	**0.75**	**0.82**

Table 2: Comparison of tags and emotion vectors

Comparing the resulting tags either directly or via the emotions they convey, our findings demonstrate that our method's results align well with human-generated annotations. In the *AMC* dataset we observed a strong direct match with our tags, with an average tag overlap of 56% at the track level ($[41.69, 70.81]$ 95% CI) and a median reaching 100%. Considering the structure of the *AMC* dataset, whose tracks are usually assigned only one tag, this means that for 56% of the tracks the *AMC* tag is contained in the set of tags we collected from music platforms. For *Cal500*, despite a low tag overlap of 3.24%, we observed significant alignment, with a mean similarity score of 0.75 between the derived emotion vectors. Note that the lower similarity score observed in *AMC* (0.68) may be attributed to its limited number of tags—with an average of one per track—compared to ours (~8 tags per track) and that of *Cal500* (~15 tags per track).

5.2 Tag Emotion Assignment Method

To assess the reliability of our method for assigning emotions to individual tags (see Section 4.2.1), we applied the same technique to the words in the NRC Lexicon. By treating each word as a 'query' and using the NRC Lexicon, excluding the query word itself, as the 'corpus', we derive emotion vectors for each word and compare them with the original vectors provided by the NRC Lexicon.

We achieved an average accuracy of 84% in identifying emotions represented by a given tag, with balanced scores across emotions. Joy was the most accurately identified emotion (93%), while fear had the lowest score (76%). Our F1-score was 77%—an expected result, as our method prioritizes emotions that align across all matched words, leading to a higher number of false negatives and thus lower recall. Nonetheless, the method's ability to generally identify emotions demonstrates its overall effectiveness.

6. DISCUSSION

In this section, we discuss the strengths and weaknesses of our approach for generating emotion-labeled datasets.

6.1 A Dual-Use Model and a Reproducible Framework

The proposed framework, methodically divided into two distinct phases, facilitates the creation of synthetic datasets suitable for both Music Emotion Recognition (with emotion vectors) and auto-tagging tasks (with tags retrieved from music platforms). Its flexibility renders it applicable to any existing dataset that includes tags on music tracks, therefore allowing researchers to create their own emotion-labeled dataset [4].

Additionally, our work can be easily extended to incorporate future resources similar to the NRC Lexicon, albeit based on other emotion models (GEMS etc.), as such resources become available. Since this approach only requires a mapping from English words to emotion labels, collecting these resources is significantly easier than obtaining direct emotion annotations on music excerpts.

6.2 Large-Scale Data Collection with Emphasis on Data Quality

Our method for extracting listener-generated textual data from music platforms overcomes the usual limitations of data collection—including time, cost and feasability constraints—through crowdsourced experiments and therefore enables data collection on a larger scale. Our final collection contains 5,892 emotion-labelled tracks, more than twice the size of the hitherto largest emotion model-based dataset of 2,648 tracks (*NTWICM* [11]).

Notably, specific attention was paid to retain only relevant tags, removing those that lack overt emotional signification, represent value judgments, or describe musical genres. Consequently, our dataset underwent significant refinement, with only a small percentage of total and unique tags retained (4.8% and 1.1%, respectively), enhancing its quality while maintaining its diversity (see Table 3). We actually ended up with 1,013 unique emotion tags, significantly more than the 157 emotion labels in the MER dataset with the largest number of labels to our knowledge (*Cal500*). Tag distribution across the dataset and within each source is presented in Figure 2, where the size of each tag reflects its frequency, taking into account its occurrence.

	Before	**After**
Tags across all tracks	1,007,847	48,737 (4.8%)
Unique tags	90,699	1,013 (1.1%)
Unique tracks	12,515	5,892 (47.1%)

Table 3: Data filtering overview, before pre-processing, after pre-processing

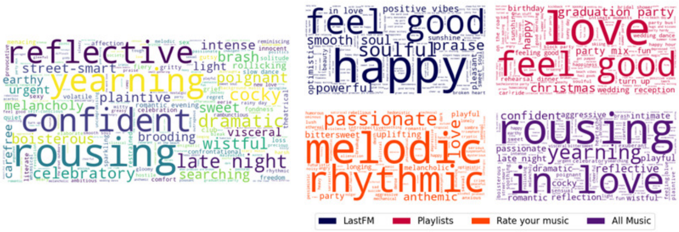

Figure 2: Tag Frequency Across Dataset and Sources

6.3 Reliance on Music Platforms

Our proposed framework relies on the availability of data on music platforms and may face challenges due to the under-representation of certain musical genres. Indeed, the final dataset exhibits a significant imbalance across genres, with world music and classical music represented only by 29 and 24 tagged tracks. We nonetheless pursued the creation of a balanced dataset by selecting the 280 most popular tracks for the top 15 most-represented genres, yielding 4,200 tracks in total [5].

The tag frequencies (see Figure 2) reveal a prevalence of certain tags, with positive-emotions particularly prominent. However, it is uncertain whether this reflects listener preferences, a wider music industry trend, or a tendency of listeners to engage with music platforms when in a positive mood. The latter possibility could potentially introduce bias into our results.

[4] Python code to derive emotion vectors from a set of tags is provided: `https://github.com/joanne-affolter/PlayMood`

[5] The dataset and its balanced version are made public : `https://github.com/joanne-affolter/PlayMood`

6.4 Emotion Association: A Focus on Explainability

Some critics might argue that our methodology for generating emotion-labeled datasets could introduce bias when training Music Emotion Recognition (MER) systems, as it relies on synthetic data. However, the use of a direct mapping from tags to emotions using a crowd-sourced Lexicon aims to ensure model explainability and interpretability. We deliberately chose not to use machine learning models to predict word emotions, opting instead for a resource curated by humans. However, we acknowledge that the strong reliance on the NRC Lexicon renders our work subject to the latter's limitations, including socio-cultural biases [25], the possibility of incorrect, nonsensical, or pejorative entries due to human error—inevitable with large-scale annotations—and potential ambiguities due to a lack of context in the lexicon [26].

6.5 Towards More Generalizable Findings

In addition to significantly reducing the need for human-generated annotations, our synthetic dataset in fact leverages the size and diversity of social music platforms. For instance, it features a larger average number of tags per track (mean: 7.52, min: 1, max: 171) compared to crowd-sourced datasets, which typically rely on a few tags (on average 1.62 for *MTG-MT* and 1.01 for *AMC Mirex*). This variety enables a broader range of interpretations and a more nuanced evaluation of listener feedback, although it may also present challenges in identifying emotions. Furthermore, agreement among listeners on music platforms tends to be relatively high, as indicated by the frequency of tag occurrences at the track level (mean: 6.01, min: 1, max: 200). In contrast to crowdsourced studies that generally require agreement between a few annotators, our method has an intrinsic potential for more robust and generalizable findings thanks to the higher number of listeners involved in the tagging process.

6.6 Emotion Modeling: Paving the Way for Future Research

By grounding our method on a set of eight primary emotions, we offer an intuitive alternative to the VA framework. Meanwhile, by using a continuous vector representation in the Plutchik emotion space, our framework is also able to capture subtle emotional nuances of music tracks, as illustrated by the emotional profiles in Figure 3 [6] .

Notably, we found that as the number of tags increased, the emotional spectra of a track became more complex, involving a wider variety of emotions with varying intensities. One may wonder whether to consider feedback from all listeners, resulting in more intricate emotion representations, or retain the best-aligned tags alone, thereby increasing consistency at the risk of missing individual nuances. In this work, we opted for mutually consistent emotion vectors with an emphasis on inter-rater agreement. By filtering out tags that represent contrasting emotions, we negotiated, on the one hand, the intersubjective variability of music perception and, on the other hand, its socially communicative potential by selecting a significant number of tags with high agreement, effectively producing complex emotion representations validated by the majority. We thus achieved an average ICC score of 0.76 for the emotion ratings associated with the selected tags for each track, indicating ***good*** *reliability* according to [24], compared to the initial score of 0.52, which suggested ***moderate*** *reliability*. It is noteworthy that, despite the filtering process, the average number of tags per track decreased only slightly from 7.52 to 6.55, demonstrating that our dataset still reflects the diversity of its participant pool.

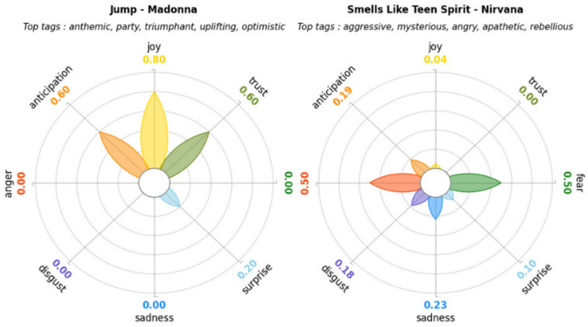

Figure 3: Visualization of tracks' emotions.

7. CONCLUSION

Our investigation introduces a novel approach to Music Emotion Recognition (MER) by benefiting from large-scale, listener-generated tagging alongside an original application of Plutchik's emotion model. With this work we aimed not only to address the scarcity of annotated datasets in the MER domain, but also to challenge traditional paradigms of music emotion by adopting an intensely empirical, psychological model-based framework. Through meticulous data collection and cleaning, we generated a dataset that surpasses existing collections in size and diversity, while maintaining a high degree of alignment with human-generated annotations. We believe that the integration of Natural Language Processing techniques in the semantic analysis of music tags is a methodological innovation that may effectively transpose the problem from audio to the textual domain. Furthermore, our approach's dual utility in MER and Auto-Tagging tasks demonstrates its versatility and potential for wide-ranging applications in Music Information Retrieval. By narrowing the gap between psychological emotion theories and computational music analysis, we pave the way for future research endeavors aimed at enriching our understanding of listeners' emotional engagement with music.

[6] A notebook for visualizing the emotional profiles across all tracks in our collection is available: `https://github.com/joanne-affolter/PlayMood`

8. ETHICS STATEMENT

Ethical Considerations in Data Handling It is important to note that our dataset does not contain any listener-specific information; our research involved the analysis of publicly available data only. By design, our approach prevents any direct links to individual listeners within the dataset, mitigating concerns around privacy and data security.

Addressing Societal and Cultural Considerations The diversity of music across cultures presents a challenge for Music Information Retrieval technologies, which should strive to prevent cultural homogenization in interpretive systems. Despite efforts to include a wide range of genres and styles, our dataset may not fully capture the breadth of global musical diversity. Additionally, the platforms from which data was sourced may primarily serve specific demographics, potentially biasing our dataset towards the musical preferences and emotional expressions of a particular segment of the global population. Future research should prioritize the collection of tags from a more culturally diverse set of sources, work towards the further mitigation of such biases, and enhance the inclusivity of MIR technologies.

9. ACKNOWLEDGMENTS

I would like to sincerely thank I. Rammos for his invaluable supervision throughout the entire project and his significant contributions to the review and correction of the paper. His expertise and support greatly enhanced the quality of this work.

10. REFERENCES

[1] R. Plutchik, *A general psychoevolutionary theory of emotion*, R. Plutchik and H. Kellerman, Eds. New York: Academic Press, 1980, vol. 1.

[2] M. Barthet, G. Fazekas, and M. Sandler, "Music emotion recognition: From content- to context-based models," in *Proceedings of the International Society for Music Information Retrieval Conference*, Virtual, 2021, pp. 1–7.

[3] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, vol. 16, no. 6, p. 166335, 2022.

[4] R. Yuan, Y. Ma, Y. Li, G. Zhang, X. Chen, H. Yin, L. Zhuo, Y. Liu, J. Huang, Z. Tian, B. Deng, N. Wang, C. Lin, E. Benetos, A. Ragni, N. Gyenge, R. Dannenberg, W. Chen, G. Xia, W. Xue, S. Liu, S. Wang, R. Liu, Y. Guo, and J. Fu, "Marble: Music audio representation benchmark for universal evaluation," in *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023, camera-ready version. arXiv:2306.10548 [cs.SD]. DOI: https://doi.org/10.48550/arXiv.2306.10548.

[5] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 2013, pp. 1–6.

[6] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *International Conference on Machine Learning (ICML)*, 2019.

[7] 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 International Society for Music Information Retrieval Conference*, 2008, pp. 462–467. [Online]. Available: https://ismir2008.ismir.net/papers/ISMIR2008_263.pdf

[8] H. Strauss, J. Vigl, P. Jacobsen *et al.*, "The emotion-to-music mapping atlas (emma): A systematically organized online database of emotionally evocative music excerpts," *Behavioral Research*, vol. 56, pp. 3560–3577, 2024. [Online]. Available: https://doi.org/10.3758/s13428-024-02336-0

[9] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. (Accessed 2024) Cal500. [Online]. Available: http://calab1.ucsd.edu/~datasets/cal500/

[10] Y. Kim, E. Schmidt, and L. Emelle, "Moodswings," Accessed 2024, offline.

[11] B. Schuller, J. Dorfner, and R. Gerhard, "Now that's what i call music," Accessed 2024. [Online]. Available: http://openaudio.eu/NTWICM-Mood-Annotation.arff

[12] T. Eerola and J. K. Vuoskoski, "Soundtracks," Accessed 2024. [Online]. Available: https://osf.io/p6vkg/

[13] S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap," Accessed 2024. [Online]. Available: http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

[14] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H.-H. Chen, "Amg1608," Accessed 2024. [Online]. Available: https://amg1608.blogspot.com/

[15] A. Aljanaki, F. Wiering, and R. Veltkamp, "Emotify," Accessed 2024. [Online]. Available: http://www2.projects.science.uu.nl/memotion/emotifydata/

[16] M. Pesek, G. Strle, A. Kavčič, and M. Marolt, "Moodo," Accessed 2024. [Online]. Available: http://moodo.musiclab.si

[17] R. Panda, R. Malheiro, and R. P. Paiva, "4q emotion dataset," Accessed 2024. [Online]. Available: http://mir.dei.uc.pt/downloads.html

[18] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "Pmemo," Accessed 2024. [Online]. Available: https://github.com/HuiZhangDB/PMEmo

[19] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models, and stimuli," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 307–340, 2013.

[20] S. M. Mohammad and P. Turney, "NRC Word-Emotion Association Lexicon (aka EmoLex)," Non-Commercial Use Only — Research or Educational, Released 2011. [Online]. Available: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[21] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1908.10084

[22] Spotify Million Playlist Dataset Challenge. Accessed 2024. [Online]. Available: https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge

[23] G. Kreutz, U. Ott, D. Teichmann, P. Osawa, and D. Vaitl, "Using music to induce emotions: Influences of musical preference and absorption," *Psychology of music*, vol. 36, no. 1, pp. 101–126, 2008.

[24] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[25] S. Zad, J. Jimenez, and M. Finlayson, "Hell hath no fury? correcting bias in the nrc emotion lexicon," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 2021.

[26] S. M. Mohammad, "Practical and ethical considerations in the effective use of emotion and sentiment lexicons," 2020. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2011.03492