Journal of Intelligent Information Systems
https://doi.org/10.1007/s10844-022-00746-0

TROMPA-MER: an open dataset for personalized music emotion recognition

Juan Sebastián Gómez-Cañón[1] · Nicolás Gutiérrez-Páez[2] · Lorenzo Porcaro[1] · Alastair Porter[1] · Estefanía Cano[3] · Perfecto Herrera-Boyer[1] · Aggelos Gkiokas[1] · Patricia Santos[2] · Davinia Hernández-Leo[2] · Casper Karreman[4] · Emilia Gómez[1,5]

Received: 3 August 2022 / Revised: 5 September 2022 / Accepted: 6 September 2022

Abstract
We present a platform and a dataset to help research on Music Emotion Recognition (MER). We developed the Music Enthusiasts platform aiming to improve the gathering and analysis of the so-called "ground truth" needed as input to MER systems. Firstly, our platform involves engaging participants using citizen science strategies and generate music emotion annotations – the platform presents didactic information and musical recommendations as incentivization, and collects data regarding demographics, mood, and language from each participant. Participants annotated each music excerpt with single free-text emotion words (in native language), distinct forced-choice emotion categories, preference, and familiarity. Additionally, participants stated the reasons for each annotation – including those distinctive of emotion perception and emotion induction. Secondly, our dataset was created for personalized MER and contains information from 181 participants, 4721 annotations, and 1161 music excerpts. To showcase the use of the dataset, we present a methodology for personalization of MER models based on active learning. The experiments show evidence that using the judgment of the crowd as prior knowledge for active learning allows for more effective personalization of MER systems for this particular dataset. Our dataset is publicly available and we invite researchers to use it for testing MER systems.

Keywords Music emotion recognition · Personalization · Active learning · Citizen science

✉ Juan Sebastián Gómez-Cañón
juansebastian.gomez@upf.edu

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona 08019, Spain

[2] Interactive and Distributed Technologies for Education Group, Universitat Pompeu Fabra, Barcelona 08019, Spain

[3] Songquito, Erlangen 91054, Germany

[4] Muziekweb, Rotterdam 3011 PV, Netherlands

[5] Joint Research Centre, European Commission, Seville 41092, Spain

Published online: 19 September 2022

1 Introduction

Music Emotion Recognition (MER) is a computational task that aims at automatically predicting emotions that are expressed by music (*emotion perception*) or those felt by the listener (*emotion induction*) (Yang & Chen, 2011). It has been historically built around signal processing to extract emotionally relevant features from music and associate these features to the listeners' emotion judgments. This task seeks to narrow the so-called semantic gap between high-level musical constructions/concepts and low-level, handcrafted representations of sound – centering it around the processing of musical content (e.g., acoustic features and lyrics). However, more recent studies have argued that user-aware music retrieval systems should include user-related factors (Barthet et al., 2013; Schedl et al., 2013; Zangerle et al., 2021; Yang et al., 2021): *user context* – fluctuating characteristics from the listener, and *user properties* – features from the listener which are more constant. In short, *user context* involves collecting data regarding listening mood, uses of music, or physiological signals, while *user properties* include demographics, musical experience, or preference. In this direction, we have previously argued that the typical response variability in annotation gathering may be exploited positively to account for listener diversity – we promote the creation of *context-sensitive* MER systems (which draw upon the users' context), and *personalized* MER systems (which rely on information from the users' properties and annotations) (Gómez-Cañón et al., 2021).

To this extent, the present study aims at delivering a MER dataset that includes anonymized user data (containing context and properties). To demonstrate the potential of the dataset, we present a personalization strategy based on the collective judgment of the crowd of participants. To collect the dataset, we created the Music Enthusiasts platform[1] in the context of the TROMPA EU research project (Towards Richer Online Music Public-domain Archives).[2]

1.1 Contributions

A typical MER workflow is summarized in five steps (see Fig. 1): (1) researchers determine a music selection and a particular emotion taxonomy for annotation, (2) listeners annotate the perceived or induced emotion in music, (3) emotionally relevant features are extracted and matched to subjective annotations, and (4) a machine learning model is trained and tested with the annotated data. Thus, the music selection becomes central to experimental design and performance of MER systems (Warrenburg, 2020) – motivating the contributions of this work:

- Inspired on work by Kim et al. (2008), Aljanaki et al. (2016), Eerola et al. (2021), and Honing (2021), we propose a citizen science approach by engaging music enthusiasts (see step 2, Fig. 1): we provide music training on the musical properties of emotion and incentivize listeners with personalized recommendations for discovery of non-Western music (in fact, music from the Global South).
- We use a categorical emotion taxonomy (see step 1, Fig. 1): free-text annotations, categorical core affects (Barrett, 2017), *perceived* basic emotions tags (Ekman, 1992),

[1] https://ilde.upf.edu/trompa/
[2] https://trompamusic.eu/

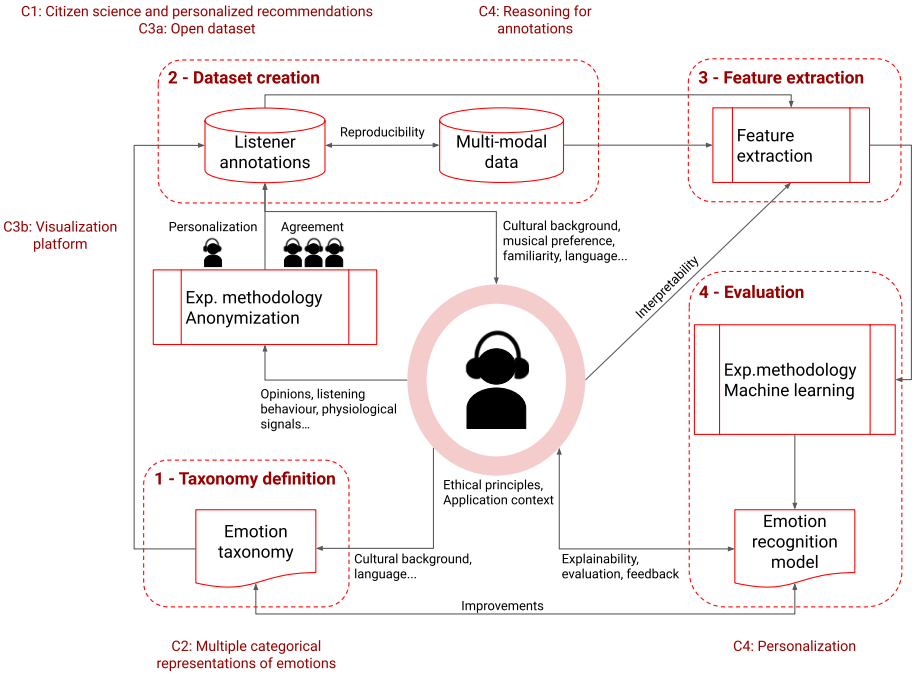

Fig. 1 Capturing personalized judgments in human-centered Music Emotion Recognition - adapted from Gómez-Cañón et al. (2021)

and musically-specific *induced* emotions (Zentner et al., 2008). We gather annotations of music in different styles and containing lyrics in different languages. Our aim is to account for different cultural backgrounds, languages, and response diversity.

- In order to favor reproducibility and open source research (see step 2, Fig. 1), our dataset is publicly available and we offer a list of links to the musical excerpts, extracted acoustic features, anonymized annotations and user profiles, and a complementary website to reproduce the findings of this study.[3] Additionally, we deliver a visualization platform to navigate and listen to the music from the dataset, and better understand the agreement and diversity of the annotations.[4]
- We enhance listeners' annotations with text data about the reasoning for each annotation (see step 2, Fig. 1). This allows to better understand if a listener chose a particular emotion due to properties of music (*perceived* emotions) or to psycho-physiological responses (*induced* emotions).
- We test and improve a personalization strategy that uses prior knowledge about the uncertainty of an excerpt with respect to the collective judgment of a crowd. Music excerpts on which we disagree upon define our personal opinions and should be taken into account for personalization (see step 4, Fig. 1). Our aim is to create an explicit

[3] https://github.com/juansgomez87/vis-mtg-mer
[4] https://trompa-mtg.upf.edu/vis-mtg-mer/

feedback loop between the listener and the models, that allows for progressive evaluation and improvement.

The rest of this paper is structured as follows. Section 2 reviews previous work and definitions for MER. In Section 3 we specify the music selection, annotation gathering approach, incentivization strategies, and personalization methodologies. Section 4 provides statistical analysis on the annotations and the validation of the personalization scheme, later discussed in Section 5.

2 Related work

MER has been subject to extensive criticism given the ambiguous and subjective nature of emotions in music (Sturm, 2013; Hong et al., 2017; Lange & Frieler, 2018; Schedl et al., 2018; Vempala & Russo, 2018; Gómez-Cañón et al., 2021; Grekow, 2021). Namely, different listeners are likely to provide diverse emotional judgments due to several factors: (1) intrinsic constructions of music (e.g., lyrics content and style), (2) socio-cultural conventions (e.g., functionality of music), (3) personal differences (e.g., listener's mood, preferences, personality, and musical experience), (4) high-level emotional evaluation mechanisms (e.g., language differences, aesthetic experience, familiarity, episodic memory, and identity confirmation), and (5) generalized confusion between the concepts of *induced* and *perceived* emotions in music. We stress the distinction between these concepts: *perceived* emotions are recognized by the listener through judgment/interpretation of musical properties (e.g., Western *happy* music is typically in major mode and has fast tempo); *induced* emotions are felt by the listener and involve psycho-physiological responses to music (e.g., *happy* music might induce *sadness* when triggering a nostalgic memory). Despite the criticism, great advances have been accomplished over the years to tackle the inherent subjectivity of the task - for in-depth reviews on MER see Laurier (2011); Kim et al. (2010); Yang & Chen (2012); Yang et al. (2018); Panda et al. (2020); Han et al. (2022).

2.1 Datasets for MER

Open datasets for MER have been developed mainly from benchmarking initiatives from the Audio Mood Classification task in the Music Information Retrieval Evaluation eXchange (MIREX)[5] and the Multimedia Evaluation Benchmark (MediaEval)[6]. The former released the DEAM dataset (Aljanaki et al., 2017): 1802 excerpts of royalty-free music rated by at least 10 MTurk[7] workers with dynamic annotations of arousal and valence.[8] The latter resulted in the Jamendo Mood and Theme subset (Bogdanov et al., 2019): 18486 audio tracks from Jamendo annotated with multi-label mood or theme tags gathered from users (e.g., *happy*, *positive*, or *epic*). We also highlight the AMG1608 dataset, unique for personalization purposes (Chen et al., 2014): 1608 excerpts from All Music Guide that

[5] https://www.music-ir.org/mirex/wiki
[6] https://multimediaeval.github.io/
[7] https://www.mturk.com/
[8] Core affects are arousal (the amount of energy/activation) and valence (pleaseantness/positiveness).

were annotated with static arousal and valence ratings by MTurk workers and students from the National Taiwan University. We refer the reader to Warrenburg Warrenburg (2020) for a thorough review of stimuli selection in music emotion studies and to Gómez-Cañón et al. (Gómez-Cañón et al., 2021) for a broad introduction on the typical methodologies for dataset creation in MER. We also offer the reader a summary website, with detailed information regarding existing datasets for MER (annotation strategy, music style, and emotion taxonomy).[9]

Importantly, it has been argued that the annotation procedure is time-consuming, tedious, expensive, and a heavy cognitive load for listeners (Yang & Chen, 2011) – interfering with the dataset creation step (see step 2, Fig. 1). Crowdsourcing strategies have been designed to attract participants (Law et al., 2007; Schedl et al., 2014) – collaborative gamification strategies lower the cost of labeling and could maximize consensus across listeners. For the case of MER, the MoodSwings game was created to annotate *continuous, perceived* emotion values of valence and arousal (Kim et al., 2008), while the Emotify game collected *discrete, induced* emotion labels (Aljanaki et al., 2016) - for theory on discrete and continuous models of music emotion see (Eerola & Vuoskoski, 2011). Thus, approaches for dataset creation that incentivize participation and reward listeners are highly desirable.

2.2 Personalized MER models

In the context of this paper, the use-case for our dataset is to produce personalized models – a central topic in the field of Affective Computing (Tkalčič et al., 2016). Moreover, personalization could promote emotion-based music recommendation applications with beneficial purposes in the context of healthcare and well-being (Grekow, 2021): improving concentration (Agres et al., 2021), promoting prosociality (Cespedes-Guevara & Dibben, 2021), aiding learning (Hu et al., 2021), or supporting tinnitus patients (Tarnowska, 2021). Given the subjective nature of emotions and the typical low agreement across annotations (Liebetrau & Schneider, 2013; Schedl et al., 2018), it is reasonable to train models exclusively with the annotations from a particular person. Particularly for MER, Yang et al. (2007) considered the role of individuality from two perspectives: (1) assembling group-based MER models that average annotations from groups of listeners with similar properties (i.e., demographics and musical experience), and (2) training personalized models which train on the annotations of a specific user. Evidence has shown that personalized models significantly outperform general models (i.e., those trained with an average rating across all listeners), although group-based models do not (Yang & Chen, 2011). Contrarily, Gómez-Cañón et al. (2020) found that group-based MER algorithms trained on the annotations of users that reported understanding the lyrics, consistently outperformed general models for a small dataset with a large amount of annotations per excerpt.

Regarding personalization for MER, Su and Fung (2012) proposed using active learning (Settles, 2012; Aggarwal et al., 2014; Yang, 2018) – the strategy relies on cleverly selecting unlabeled data instances so that the algorithms require less training. Chen et al. (2014; 2017) proposed model adaptation – they developed a general MER regression model (namely, Gaussian Mixture Models) and progressively tied the Gaussian components

[9] https://github.com/juansgomez87/datasets_emotion

to adapt the models based on the maximum a posteriori (MAP) linear regression. More recently, Gómez-Cañón et al. (2021) introduced *consensus entropy* for MER: disagreement of a committee of classifiers and/or listeners is used to progressively re-train models with personal annotations (Cohn et al., 1994).

Aiming at personalization, we evaluated the following research questions:

RQ1: Are there differences in emotion judgments of both perceived and induced emotions, according to musical and user properties?

RQ2: Can novel personalization classification schemes, i.e., consensus entropy (Gómez-Cañón et al., 2021), generalize to the TROMPA-MER dataset?

3 Dataset creation

3.1 Music selection

The TROMPA project included several international partners including the Stichting Centrale Discotheek Rotterdam (CDR) and their online library Muziekweb.[10] Muziekweb is the music library of the Netherlands, which offers an automatic selection of 30 second excerpts of each song and metadata accessible to everyone: 600,000 CDs, 300,000 LPs, and 30,000 music DVDs. Muziekweb presents a general taxonomy including genres, but is also indexed by different countries of the world. This allowed us to focus on non-Western styles of music from the Global South, since cross-cultural research of MER models is important to target user groups with different cultural and socioeconomic backgrounds (Gómez-Cañón et al., 2021). WEIRDness in music psychology studies has been openly discussed recently by the research community (Henrich et al., 2010; Jacoby et al., 2020) – researchers and participants are typically from Western, Educated, Industrialized, Rich, and Democratic backgrounds.

To gather initial data from the possible pool of songs, we used the Spotify Web API[11] to pre-select songs which were both available to Muziekweb and Spotify – we gathered information regarding danceability, key, mode, energy, and valence. Our aim was to select songs which, according to Spotify, had strong indicators of belonging to a particular quadrant. Since the calculation of energy and valence is unknown (i.e., Spotify's algorithms are essentially a black box), we standardized the values retrieved from the API (zero mean and unit variance), and balanced the selection of songs for each resulting quadrant.

- *Music in Spanish (100 excerpts) and Portuguese (100 excerpts).* We gather recent popular music containing lyrics in both languages. Over 91% of the songs were recorded after the year 2000.
- *Music from Africa (120 excerpts).* We rely on Muziekweb's music taxonomy to retrieve music with lyrics in the Griot tradition from West Africa, which usually contains Kora (a stringed instrument with 21 strings played by plucking with the fingers).
- *Music from Latin America (539 excerpts).* We gather traditional and popular music with lyrics from the following countries: Argentina, Bolivia, Brazil, Chile, Colombia,

[10] https://www.muziekweb.nl/

[11] https://developer.spotify.com/documentation/web-api/

Costa Rica, Cuba, Dominican Republic, Ecuador, Mexico, Panamá, Perú, Puerto Rico, Uruguay, and Venezuela.
- *Music from the Middle-East (219 excerpts).* We gather traditional and popular music with and without lyrics from the following countries: Syria, Irak, Yemen, Kurdistan, and Lebanon.
- *Choir music (83 excerpts).* We gather music with singing voice balanced in the following languages: Spanish, English, German, and other languages (Dutch, French, Danish, and Finnish).

Following recommendations from Gebru et al. (2018) and Prabhakaran et al. (2021) to document the properties a dataset, we offer an interactive "datasheet" for our dataset: a complementary website to listen to the music collection, emotion representations, and participants' annotations and agreement.[12] The aim is to highlight the issue of subjectivity to future MER researchers.

3.2 Citizen science and music recommendations

As a citizen science approach, the platform offers examples as concise guides to Western musical properties which are typically related particular emotions – allowing the listener to better understand the annotation task. We used two types of music recommendations to incentivize participation to our platform:

- *Informative recommendations*: we split the music into annotation campaigns of 20 songs, and participants received a recommendation belonging to their most rated category after completing the campaign (e.g., if a listener is biased toward annotating music with a given category, the platform presents recommendations from this category). We inform musical properties of tempo, mode (major or minor), and danceability. For example, a song with 120 BPM, major tonality, and 70% danceability is likely be classified as *happy* – reinforcing concepts of musical properties of emotion.
- *Personalized recommendations*: using the personalization framework described in Section 3.4, each user annotated 10 songs, personalized models were trained, participants chose the location of the desired recommendation (West Africa, Latin America, and Middle East), and received four recommendations according to their personalized model. Participants gave explicit feedback of their agreement with these recommendations – they rated if each recommendation category agrees to their opinions.

To obtain greater response variability, we translated our platform into English, Spanish, Italian, Mandarin, and Dutch. We refer the reader to Gutiérrez Páez et al. (2021) for a full description of other incentivization strategies that were implemented (e.g., using gamification to compare participants' engagement).

3.3 Annotations and emotion taxonomies

Previous to annotation, all participants accept the data management consent form and we collect the annotators' current mood (with a free-text in native language and a visual

[12] https://trompa-mtg.upf.edu/vis-mtg-mer/

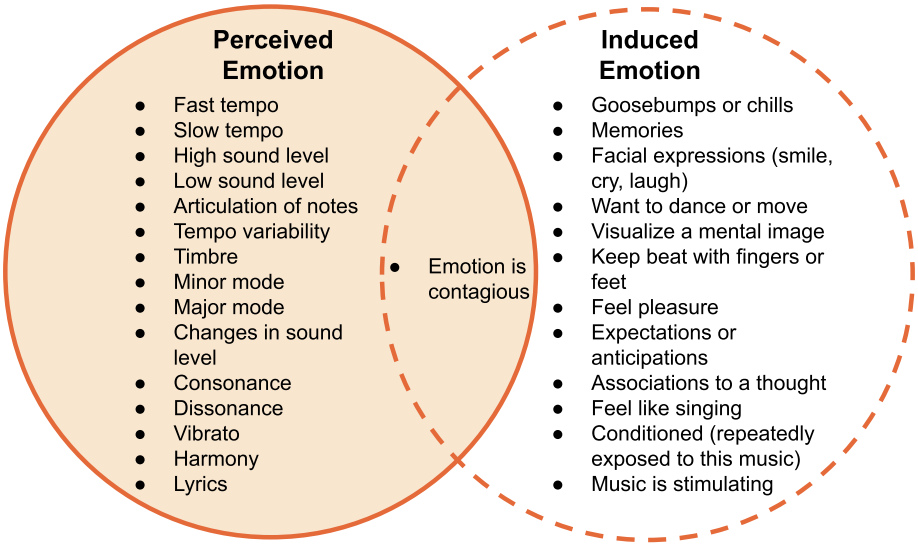

Fig. 2 Collection of reasons presented to the participants

category from Pick-a-mood (Vastenburg et al., 2011)).[13] We refer the reader to Meyer (1961), Budd (1992), Yang and Chen (2011), Eerola & Vuoskoski (2011), Eerola (2018), Céspedes-Guevara & Eerola (2018), Juslin (2019), Warrenburg (2020), and Dufour & Tzanetakis (2021) for thorough studies of emotion representation in music. For our particular platform, we offer variable granularity for the categorical/discrete representation of emotions (Ekman, 1992): each annotation describes a particular class/label (e.g., happy, sad). Although this approach is naturally ambiguous by using language to describe emotions and is limited compared to the richness of human emotion (Yang & Chen, 2011), it has been argued that musical emotions are likely to be prototypical (Juslin, 2019). Inspired by Cowen (Cowen et al., 2019) and Warrenburg (Warrenburg, 2020; 2020), we attempt to provide coarse and fine emotion granularity to our participants through different categories: single free-text emotion word in the listener's native language, forced-choice categorical arousal and valence, and eleven emotion adjectives typically used in music emotion studies. We use four basic emotions (*anger*, *surprise*, *fear*, and *sadness*) (Ekman, 1992), and seven music-specific emotions from the Geneva Emotion Music Scale – GEMS (*joy*, *power*, *tension*, *bitterness*, *peace*, *tenderness*, *transcendence*) (Zentner et al., 2008). Since Russell's circumplex model of emotion is dimensional (Russell, 1980), we can further categorize music into four quadrants resulting from the separation of arousal and valence axes (see Panda et al. (2018): Q_1 corresponds to positive arousal and valence, Q_2 to positive arousal and negative valence, Q_3 to negative arousal and valence, and Q_4 to negative arousal and positive valence. In order to decrease the amount of time for each annotation and following our previous research (Gómez-Cañón et al., 2020), we present subsets of the emotion adjectives depending on the choice of quadrant: Q_1 (*joy*, *power*, *surprise*), Q_2

[13] Moods reflect affective states that have lower intensity than emotions, do not have a clear "object", have a low intensity, and are more lasting than emotions – they reflect an overall status of the person (e.g., gloomy).

(*anger*, *fear*, *tension*), Q_3 (*bitterness*, *sadness*), and Q_4 (*peace*, *tenderness*, *transcendence*). It is questionable that these emotion words are effectively mapped to the given quadrants (e.g., the adjective *surprise* could have either positive or negative valence). However, the annotation of each musical excerpt is highly demanding and this procedure reduced annotation time significantly (see (Gómez-Cañón et al., 2020)). Furthermore, we collect the listener's preference and familiarity for each music excerpt. Finally, listeners state the reasons behind each annotation. From a citizen science perspective, the aim is to educate each participant and as a result obtain higher quality annotations. We offer a pool of reasons for both emotion perception (musical properties like consonance, harmony, sound level) and for induced emotions (psycho-physiological reactions such as chills, thought associations, memories). The latter are simplifications of the BRECVEMA model that includes eight mechanisms of emotion induction (Juslin, 2013): Brain Stem Reflex, Rhythmic Entrainment, Evaluative Conditioning, Contagion, Visual Imagery, Episodic Memory, Musical Expectancy, and Aesthetic Judgment. We offer reasons such as the music "*makes me feel pleasure*", "*makes me want to dance or move*", or "*makes me feel like singing*" (see Fig. 2). Importantly, the set of reasons for induced emotions can have positive or negative valence (e.g., goosebumps, memories, mental images, thought associations). Thus, we can analyze data reliability and better understand how emotional judgments are converted into annotations. It must be noted that research on physiological reactions to music should involve analyzing biosignals such as heart rate variability, electrodermal activity, or electro-encephalography (Poli et al., 2021; Saganowski et al., 2022) – however, our approach is centered on understanding the underlying reasoning in order to understand agreement and improve personalization. Thus, one annotation contains the following information, for example: Q_1 (positive arousal and valence), *joy* (force-choice), *merry* (free-text), positive familiarity, negative preference, and three reasons for choosing arousal (e.g., *fast tempo*), valence (e.g., *major mode*), and emotion (e.g., *thought associations*).

3.4 Personalization methodology

Our personalization methodology is based on the assumption that prior knowledge about the collective judgment of the crowd of annotators results in indicative instances of classification boundaries across individual listeners (Gómez-Cañón et al., 2021). For this reason, we gathered multiple annotations from the same excerpt and evaluated their agreement as input to the algorithm. Music excerpts on which we disagree upon define our personal opinions and should be taken into account to create personalized classification boundaries. We used active learning to progressively fine-tune an ensemble of pre-trained classifiers according to each participant's annotations of four classes (i.e., quadrants in arousal-valence space). Active learning is based on consensus entropy (Cohn et al., 1994; Settles, 2009): (1) an ensemble of classifiers predict the quadrant probability of unlabeled data, (2) probabilities are averaged across classifiers and Shannon's entropy is calculated across quadrants, (3) instances with highest entropy are queried to the participant, and (4) classifiers are re-trained with the provided annotations. For example, an ensemble of four classifiers analyze a song and reach high uncertainty: each classifier predicts one quadrant with 100% probability yielding an average probability of $p_{avg} = \{Q_1 : 0.25, Q_2 : 0.25, Q_3 : 0.25, Q_4 : 0.25\}$ and high inter-class entropy of 1.386. In Gómez-Cañón et al. (2021), we enhanced this methodology by introducing inter-rater agreement as input. Instead of using the output probabilities from the classifiers, we calculate entropy on the normalized annotation histogram. For example, given six annotations for song i we calculate a relative frequency $f_i =$

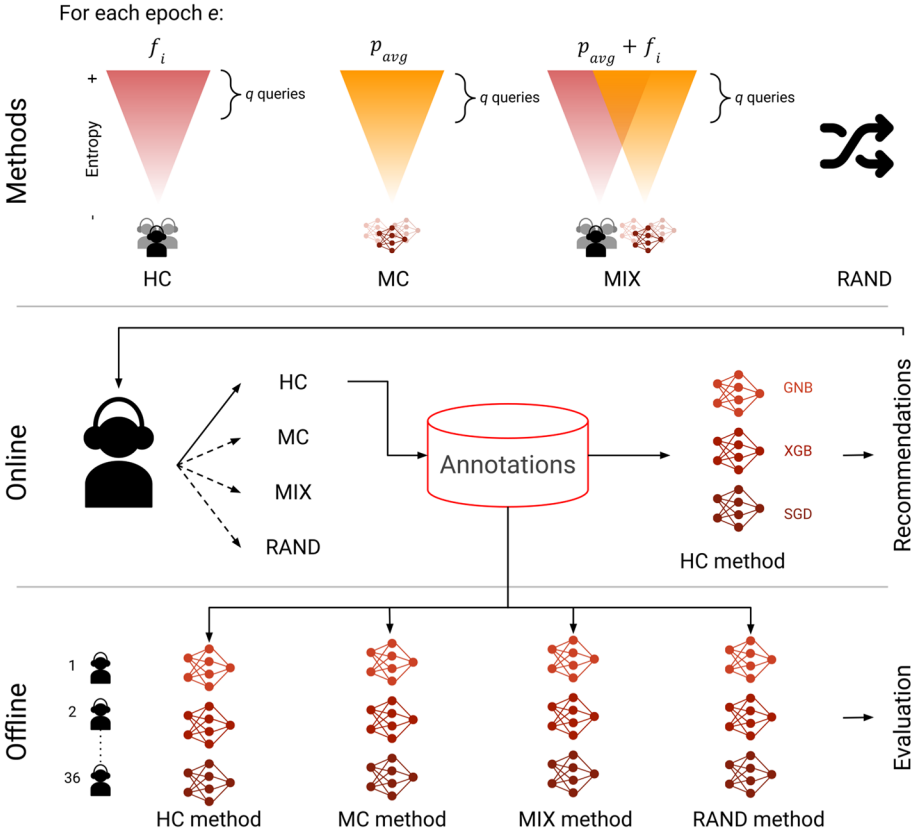

Fig. 3 Schematic of each proposed method. We show online personalized recommendations and offline evaluation

$\{Q_1 : 1/6, Q_2 : 2/6, Q_3 : 3/6, Q_4 : 0/6\}$ and its corresponding entropy of 1.011. We proposed four methods to choose informative query instances: (1) analyzing the agreement achieved by an ensemble of pre-trained models (*machine consensus - MC*), (2) analyzing the agreement from an ensemble of annotators (*human consensus - HC*), (3) taking into account both ensembles (*hybrid consensus - MIX*), and (4) randomly selecting instances as a baseline (*RAND*). In short, q instances with highest entropy (i.e., highest uncertainty) with respect to the method are queried to each participant, annotations are used to re-train each algorithm, and the process is repeated throughout e epochs (see top schematic in Fig. 3).

In this work, we enhance the methodology from Gómez-Cañón et al. (2021) to improve class balancing for each epoch: (1) before calculating entropy, we split the probabilities p_{avg} or f_i into four matrices corresponding to the instances with high probability of belonging to each quadrant, (2) we calculate entropy independently for each matrix, and (3) we select instances with highest entropy from each matrix. Thus, we alleviate the issue of imbalanced classes for each epoch, since the instances selected for query are more likely to belong to a particular quadrant. In the case that the probabilities do not favor a particular quadrant (i.e., models/annotators are biased towards particular classes), we simply select

the instances with highest entropy from the initial matrix. We extracted 260 emotionally relevant acoustic features (mean and standard deviation of 65 low-level music descriptors and their first order derivatives) from segments of 1 second (Aljanaki et al., 2017), with 50% overlap, and standardize across features – using the IS13 ComParE feature set (Weninger et al., 2013) and the OpenSMILE toolbox (Eyben et al., 2013). For each participant, we created an ensemble of 15 machine learning algorithms, pre-trained using the DEAM dataset (Aljanaki et al., 2017) – 5 Gaussian Naive Bayes classifiers (GNB), 5 Logistic Regression classifiers optimized with Stochastic Gradient Descent (SGD), and 5 Extreme Gradient Boosting classifiers (XGB). Each ensemble of 5 classifiers is trained from a different cross-validation split from the DEAM dataset to assure different prediction outputs. We use these algorithms since they offer probabilistic outputs and are less computationally expensive than novel methodologies like neural networks. In total, 181 annotators completed 4721 annotations of 691 music excerpts – which we use as the prior information for the HC and MIX methods. For the online platform, each participant was randomly assigned to a sampling method and $q = 10$ the sampling method is kept throughout their participation: each participant annotated 10 excerpts and then received recommendations as the models were re-trained. However, we had no control over the amount of epochs e that our participants were willing to complete. Hence, we used their annotations to test all methods offline (*MC, HC, MIX, RAND*) with different combinations of q and e, and analyzed which models were effectively personalized with respect to each participant (see online and offline schematics in Fig. 3).

4 Analysis and discussion

After implementing the personalization strategy on the platform, we result with a total of 181 participants (44 new participants) that completed 4721 annotations (1312 new annotations) of 691 music excerpts (470 still require annotations). Given past research on the impact of language on music emotion annotation (Gómez-Cañón et al., 2020), we gathered information on the birth place, native language and second languages. 85% of our participants were born in Europe, 8% in America (6% in South America and 2% in North America), and 6% in Asia. The main native languages from our participants were Spanish (32.7%) and Català (34.9%), followed up by Romanian, Dutch, English, Croatian, Ukranian, Greek, Italian, French, Chinese, Russian, Galician, Turkish, Portuguese, Korean, German, Serbian, Japanese, Swedish, and Arabic ($\leq 3\%$). The main second languages were English (45.4%), Spanish (21.4%), Català (11.4%), French (8.7%), followed up by German, Italian, Basque, Dutch, Russian, Chinese, Polish, Hungarian, Bosnian, Vietnamese, Galician, and Arabic ($\leq 2\%$).

4.1 Inter-rater agreement

Following (Schedl et al., 2018; Gómez-Cañón et al., 2020), we assessed the agreement of participants using reliability statistics with respect to self-reported characteristics of annotations (i.e., preference and familiarity) and music properties retrieved from Spotify (i.e., mode, tempo, danceability, acousticness, and popularity) (Krippendorff, 2004). Among other agreement indexes considered, the benefits of using Krippendorff's α are: offering several types of metric (in this case, nominal since annotations are categorical), handling of missing data and any number of observers, and not requiring a minimum of sample

Table 1 Krippendorff's α for annotations of quadrants, arousal, valence, and emotion

Properties	Configuration	Annotations	%	Quadrant	Arousal	Valence	Emotion
–	*All*	*4721*	*100%*	*0.313*	*0.449*	*0.310*	*0.182*
Annotation Properties	Positive Preference	1843	39%	0.305	0.455	0.229	**0.189**
	Negative Preference	2878	61%	**0.322**	**0.456**	**0.329**	0.175
	Positive Familiarity	1418	30%	**0.343**	**0.479**	**0.360**	0.172
	Negative Familiarity	3303	70%	0.299	0.434	0.287	**0.184**
Musical Properties	Major tonality	2835	60%	**0.318**	0.421	**0.325**	**0.191**
	Minor tonality	1886	40%	0.296	**0.482**	0.273	0.160
	Tempo $\geq$ 100 bpm	2412	51%	**0.341**	**0.485**	**0.329**	**0.208**
	Tempo $<$ 100 bpm	2309	49%	0.273	0.397	0.286	0.148
	Danceability $\geq$ 0.35	2397	51%	**0.316**	**0.440**	**0.315**	**0.203**
	Danceability $<$ 0.35	2324	49%	0.263	0.388	0.279	0.131
	Acousticness $\geq$ 0.98	2280	48%	0.250	0.388	0.256	0.113
	Acousticness $<$ 0.98	2441	52%	**0.337**	**0.448**	**0.360**	**0.227**
	Popularity $\geq$ 0.1	1976	42%	0.274	0.412	0.259	0.177
	Popularity $<$ 0.1	2745	58%	**0.330**	**0.468**	**0.330**	**0.181**

The annotation dataset is also filtered by preference, familiarity, mode, tempo, danceability, acousticness, and popularity. Bold indicates higher agreement for positive or negative filters

size. When there is no disagreement among annotators, there is perfect reliability ($\alpha = 1$). Conversely, if agreement and disagreement are due to chance, there is a lack of reliability of the data ($\alpha = 0$). However, α could be smaller than zero if agreement is below what is expected by chance or the sample size is too small. According to Krippendorff (2004), data with $0.4 \leq \alpha \leq 0.67$ shows fair agreement and $\alpha \geq 0.8$ is considered to have good agreement.

We report our inter-rater agreement results in Table 1. Although annotation properties (i.e., preference and familiarity) depend on the users' annotations, we attempted to split the dataset into a balanced number of annotations with respect to musical properties. Hence, the bounds to split the annotations with respect to properties of tempo, danceability, acousticness, and popularity were selected accordingly. In general, our results are consistent with findings from (Lange & Frieler, 2018; Gómez-Cañón et al., 2020): only arousal annotations have a fair agreement according to Krippendorff (2004), which justifies the creation of personalized models for MER – although we only use four quadrants to describe emotion, subjectivity plays a major role in the creation of a possible "ground truth" to machine learning algorithms for MER. We also remark the following findings, with respect to other comparisons on the table: (1) participants that reported disliking and being familiar with the music show higher agreement for arousal-valence annotations, while the opposite happens for emotion words – we observe that negative preference and familiarity could result in basing judgments on musical properties (i.e., arousal-valence), but preference and negative familiarity relate more to *induced* emotions (i.e., emotion words – see Section 4.2); (2) although higher agreement for arousal is reached over songs with minor tonality, higher agreement for quadrants, valence, and emotion words is reached for major tonality – it is likely that the annotated excerpts are not necessarily in the same mode as the one calculated by Spotify's algorithms (which produce song-level predictions); (3) excerpts with tempo over 100 BPM, high danceability, low acousticness and not popular show higher

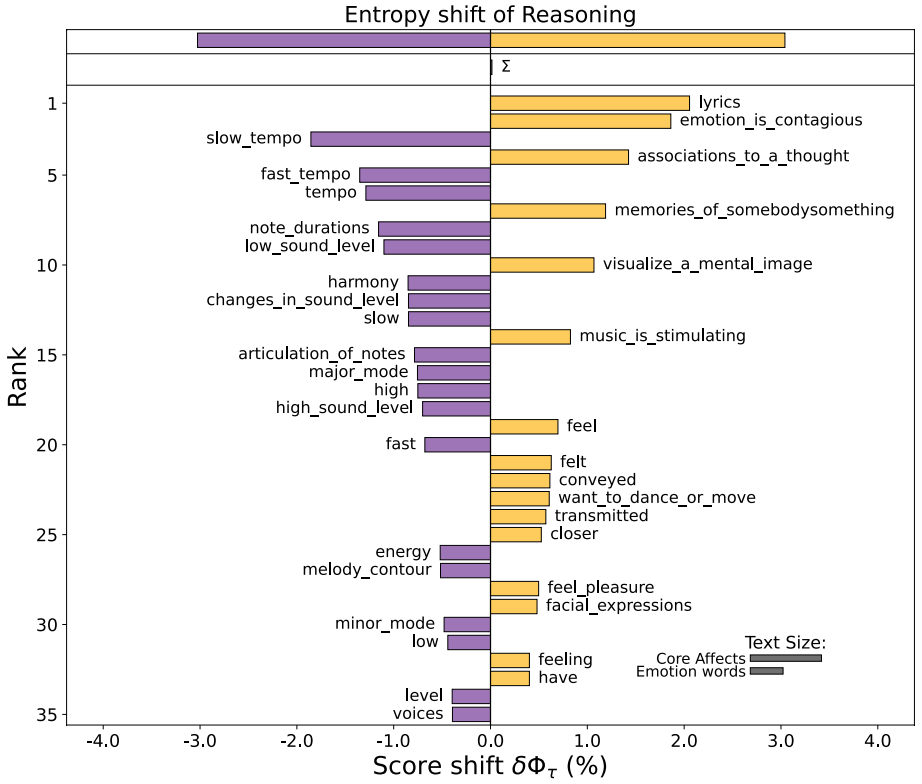

Fig. 4 Entropy shift for reasoning of arousal/valence (in purple) vs. emotion words (in yellow). Σ denotes the overall contribution of the compared corpora

agreement – these are song-level features from Spotify and it is likely that their algorithms are mainly trained on Western music, which do not completely capture the variability of music from Africa, Latin America, and the Middle East.

4.2 Reasoning behind emotion judgments

The collection of reasons behind each annotation allowed to perform a deeper analysis to understand the emotion judgments (for both *perceived* and *induced* emotions). As mentioned in Section 3, we gathered four sources of text data: (1) free-text emotion annotations, (2) reasoning for arousal annotation, (3) reasoning for valence annotation, and (4) reasoning for emotion words. Importantly, the reasoning from (2), (3), and (4) can be free-text or chosen from a pool of explanations that relate to *perceived* emotions (musical properties) and *induced* emotions (psycho-physiological responses). To perform this analysis we employ *generalized word shift graphs* (Gallagher et al., 2021), a methodology that provides word-level explanations on how and why two texts differ. This method uses two text corpora (e.g., reasoning for positive/negative valence) and the relative word frequencies for each corpus to perform pairwise comparisons between texts and obtain a list of the most characteristic words for each text. For our case, we used Shannon's entropy to calculate each word's contribution to a given text: a word will appear higher in the ranking when it is

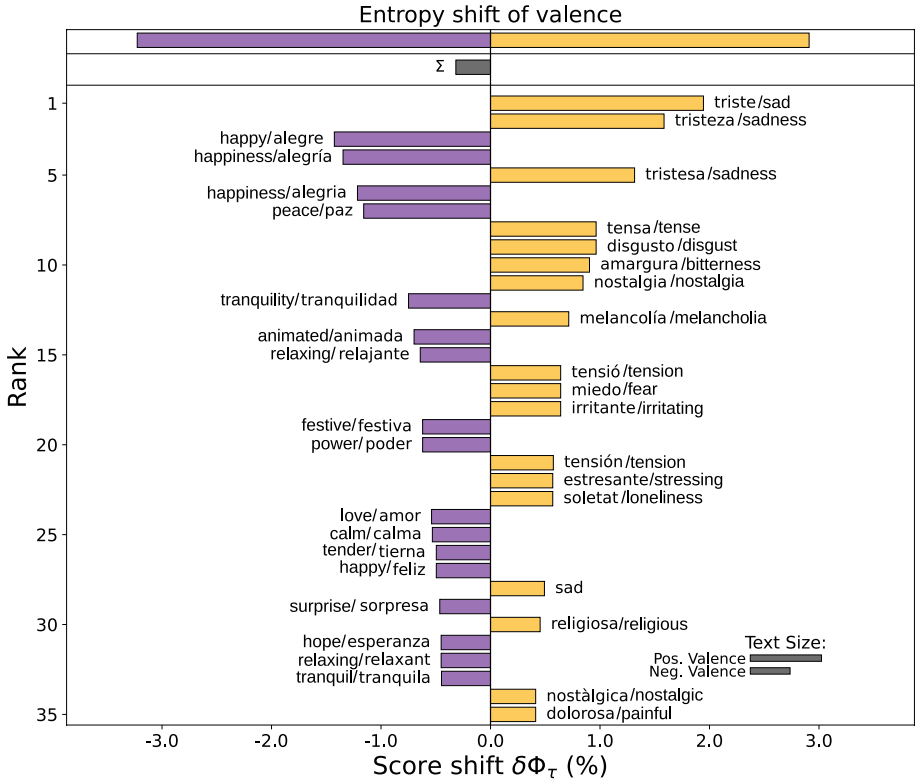

Fig. 5 Entropy shift for free-text annotations of positive valence (in purple) vs. negative valence (in yellow). Σ denotes the overall contribution of the compared corpora

more *suprising* (i.e., when it is more diverse from the other corpus). Hence, the difference of contribution $\delta\Phi_t$ for each word τ can be calculated as:

$$\delta\Phi_\tau = p_\tau^{(2)} log_2(1/p_\tau^{(2)}) - p_\tau^{(1)} log_2(1/p_\tau^{(1)}) \quad (1)$$

where $p_\tau^{(1,2)}$ is the relative frequency of each word in texts T_1 and T_2.

In order to analyze these texts, we performed two comparisons: (1) how explanations for core affects (T_1) differ from explanations for emotion words (T_2) – we assume that core affects (i.e., arousal and valence) are more related to *perceived* emotions, and (2) how free-text words contribute to differences between annotations of positive (T_1) and negative valence (T_2) – as seen in Table 1, valence is highly subjective and shows less inter-rater agreement as a core affect. Regarding comparison 1, we obtained 16951 reasons to explain arousal/valence and 8135 to explain emotion words (see Fig. 4). Interestingly, we find that the explanations for core affects (i.e., arousal and valence) relate more to musical properties (consistent with work by Barrett (2017) and Panda et al. (2018)). Conversely, the annotations of emotion words (e.g., *joy*) are typically explained with reasons for induced emotions. We remark that this is particularly important finding when creating a labeled dataset – careful thought must be given to describe emotions with words in music datasets.

Regarding comparison 2, we analyzed 1043 free-text annotations for positive and negative valence (see Fig. 5). We found that most of our annotators use Spanish and Català

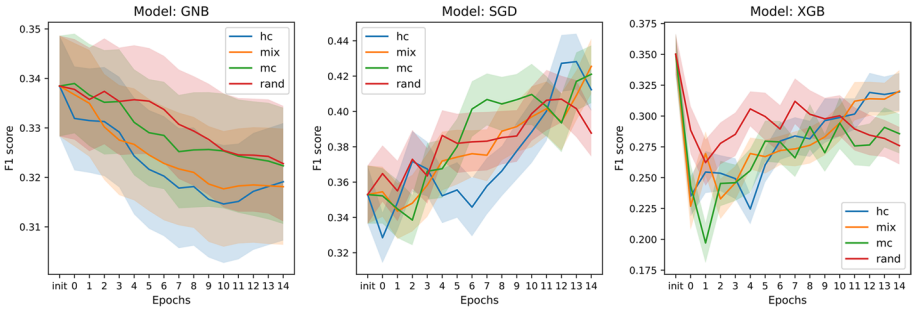

Fig. 6 Average results of weight-averaged F1-scores for each type of model, across 36 users and 5 classifiers (shaded area corresponds to $CI = 95\%, n = 180$). HC stands for Human Consensus, MC for machine consensus, MIX for hybrid consensus and RAND for random selection. Initial refers to the F1-score previous to introducing active learning and $\{q : 4, e : 15\}$

languages (see matched translations). In short, it is interesting to see that several of the words can refer to both *perceived* and *induced* emotions, and the word *surprise* is more related towards a positive valence. However, more work is needed in order to map multilingual description of emotion words used to describe music.

4.3 Personalized models

36 participants annotated more than 80 excerpts each (i.e., between 80 and 173 annotations) and we used their annotations to evaluate personalization independently from the online platform. Hence, we have a total of 3289 annotations and split each participant's dataset with non-overlapping excerpts into 85% for training (at least 68 excerpts) and 15% for testing (at least 12 excerpts). We tested different combinations of number of query instances q and epochs e to use most of the training data for all participants and balanced amount of queries with respect to the four quadrants: $\{q : 4, e : 12\}$, $\{q : 4, e : 15\}$, and $\{q : 8, e : 6\}$. We also test non-balanced queries: $\{q : 2, e : 33\}$, $\{q : 33, e : 2\}$, $\{q : 5, e : 10\}$, $\{q : 10, e : 5\}$, $\{q : 6, e : 11\}$, and $\{q : 11, e : 6\}$.

We obtained a total of 2160 trained classifiers – 36 participants × 3 algorithms (GNB, SGD, XGB) × 5 models per ensemble (cross-validation iteration) × 4 consensus entropy methods (HC, MC, MIX, RAND). We evaluated the classification scores and reported weight-averaged F1-scores, since most personal datasets turn out to be class-imbalanced (i.e., it is likely that annotators are possibly biased towards a particular class).

Our findings are consistent with Gómez-Cañón et al. (2021) and summarized in Fig. 6 for $\{q : 4, e : 15\}$: (1) while the overall classification performance is low, we observe that SGD and XGB models consistently improve as new annotations are presented to the algorithms – personalized MER models appear to benefit of using active and ensemble learning. (2) We use pairwise, one-sided t-tests ($d.f. = 179$, statistical significance $p < 0.0125$ with Bonferroni correction) to evaluate differences after training among consensus entropy methods (HC, MC, MIX, and RAND) – for the SGD algorithm, all methods are significantly better than the random baseline (HC: $p = 0.003$, MIX: $p = 0.001$, and MC: $p = 0.001$). For the XGB algorithm, our proposed methods significantly outperform the random baseline (HC: $p < 0.001$, MIX: $p < 0.001$), and the typical MC approach (HC: $p = 0.001$, MIX: $p = 0.002$). (3) Two-way ANOVAs were made independently for each algorithm to compare the effect of the consensus entropy method and cross-validation iterations

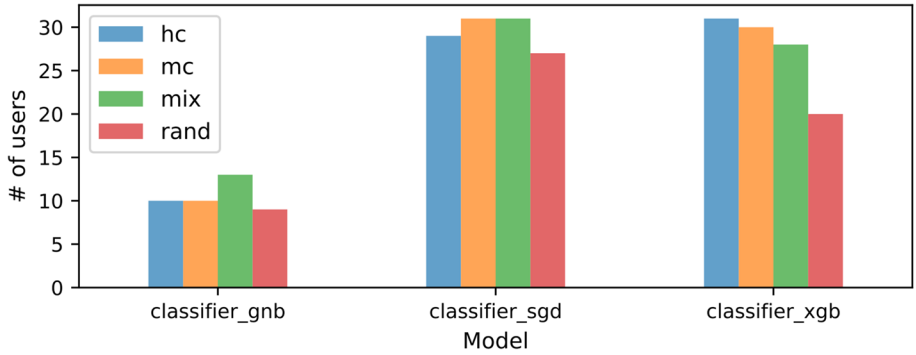

Fig. 7 Number of users with effective personalization per algorithm from a total of 35 participants

on the F1-score after training. These tests revealed significant interactions between the methods and iterations for the GNB and SGD algorithms ($p < 0.001$ and $p = 0.010$ respectively) and a significant effect of the consensus entropy method for the XGB algorithm ($p = 0.048$). (4) Tukey HSD tests for multiple comparisons found that the mean value of F1-score for the XGB algorithm was significantly different between HC and RAND/MC ($p < 0.001$ and $p = 0.010$ respectively) and between MIX and RAND/MC ($p < 0.001$ and $p = 0.005$ respectively). (5) RAND is a strong baseline – significant differences appear across consensus entropy methods, but improvements only appear after most of the data has been used to train the models and deeper analysis of each personalized ensemble showed varying behaviors (i.e., effective personalization is not reached for every participant). (6) GNB models appear not to personalize, consistent with findings from Gómez-Cañón et al. (2021) – naive bayesian models are expected to have limited generalization to new data. (7) Empirical tests with q and M show improvements with a low amount of queries and higher amount of epochs (e.g., $\{q : 5, e : 10\}$, $\{q : 2, e : 33\}$, $\{q : 6, e : 11\}$) – it is likely that more iterations are beneficial for the models to progressively adapt to the new annotations, however best performance is achieved by attempting to balance the queries, using the HC method, and XGB.

Since personalization is not reached for all participants, we evaluate them independently and fitted linear regressors on the average performance metrics of each ensemble of classifiers for each algorithm and assume effective personalization when the slope of the regressor is positive (see Fig. 7). We fitted a linear regressor on the five F1-scores for the each participant's ensemble and each epoch $e = \{0..14\}$ – we assume that an ensemble is effectively personalized when the slope is positive and performance increases with increasing number of epochs. Hence, we obtained 432 ensembles – 36 participants × 3 algorithms (GNB, SGD, XGB) × 4 consensus entropy methods (HC, MC, MIX, RAND). We summarize our findings as follows: (1) SGD models show a similar amount of effectively personalized models across sampling methods. (2) XGB models show that all proposed consensus entropy methods appear to produce more personalized models than RAND (particularly for the HC method). This indicates that, although the tendencies described in Fig. 6 show that RAND is a strong baseline, other methods produce more cases with effective personalization. (3) As expected, GNB models produce a few personalized models regardless the consensus entropy method – it is likely this algorithm is inappropriate given the non-gaussian distribution of data, reducing the performance of the ensemble. (4) MIX method produces

Table 2 Krippendorff's α for annotations of quadrants, arousal, valence, and emotion for typical (AVG) and outlier users (HIGH and LOW)

User type	# of users	Quadrant	Arousal	Valence	Emotion
HIGH	5	0.345	0.567	0.299	0.189
AVG	26	0.324	0.455	0.312	0.176
LOW	5	0.159	0.165	0.314	0.127

a high amount of personalized GNB and SGD models – MIX possibly exploits complementary advantages from HC and MC.

We also analyzed the information regarding the users that exhibited effective personalization. Each of the 36 participants had 12 ensembles: 3 algorithms (GNB, SGD, XGB) $\times$ 4 consensus entropy methods (HC, MC, MIX, RAND). On average for all participants, seven ensembles showed effective personalization ($\mu = 7.47 \pm \sigma = 2.26$). Thus we selected outlier participants: 5 participants had 10-12 effective ensembles (HIGH personalization) and 5 participants had 2-4 effective ensembles (LOW personalization). We analyze inter-rater agreement for these groups and we present our findings in Table 2: users for which personalization was effective (HIGH) exhibit high agreement in annotations and viceversa – this group of participants appear to exhibit the highest agreement in annotations even when comparing to the results from Table 1.

5 Conclusions

The Music Enthusiasts platform is an ongoing citizen science project which explores learning about music and emotion. This paper presents the TROMPA-MER dataset, an openly available dataset with diverse categorical annotations to account for the subjectivity of the task and foster response diversity. Our dataset offers advantages over existing MER datasets which can be exploited by future researchers: (1) additional information regarding each annotator (spoken languages, current mood, and birth place), (2) diverse styles of non-Western music belonging to the Global South, (3) high amount of annotations for certain excerpts (up to 143 annotations), (4) gathering reasons for annotation improves understanding of our participants' judgments, and (5) free-text annotations in native language enhance response diversity. Comparatively to other MER datasets, ours shows the following limitations: (1) using a categorical taxonomy is restricted due to language ambiguity (see Soundtracks dataset by Eerola & Vuoskoski (2011)), and (2) not taking into account temporal variability of annotations (see Aljanaki et al., (2017)). Our aim was to enrich our dataset with high quality annotations from multiple listeners, in order to place the listener "in-the-loop" of the MER framework by means of: (1) promoting the study of non-Western music and allowing our participants to discover music from different parts of the world, (2) engaging our participants with musical training and gamification strategies, (3) allowing for explicit feedback in order to continuously improve the performance of the algorithms.

Our answers to the research questions posed in Section 2 would be as follows. Regarding *RQ1* - Are there differences in emotion judgments of both perceived and induced emotions, according to musical and user properties? In short, we found an overall lack of agreement of categorical annotations of emotion in music – the illusion of universality in emotion has led to the typical practice of averaging annotations to create "ground truths". Moreover, we found higher inter-rater agreement in cases that participants reported to be familiar with the music – we remark that employing cultural experts with knowledge

from music styles is important for data reliability purposes. In our intention to enrich our dataset, we asked participants to explicitly state the reasons behind each annotation. We found that, not only the annotation task is highly subjective, but participants tended to confuse reasoning behind the annotation of *perceived* and *induced* emotions. We also found that the reasons which were chosen by participants to describe annotations of arousal and valence are more related to purely musical properties and *perceived* emotions (e.g., tempo, note duration, sound level), while the reasoning for emotion words relate more to psychophysiological responses to music and *induced emotions* (e.g., associations to a thought, wanting to dance or move). We remark that these findings pose a fundamental issue to the generation of emotion-based classification strategies relying on tags for streaming services – broad arousal/valence annotations should be used to describe *perceived* emotion and precise free-text annotations in native language could better capture *induced* emotions. Thus, we propose that the issue of subjectivity can be addressed by creating personalized models – personalized MER uses the annotations from a specific user to train a user-tailored model.

With respect to *RQ2* - Can novel personalization classification schemes generalize to the TROMPA-MER dataset? To the best of our knowledge, the use of the collective judgment as a personalization strategy has never been explored in MER so far. We are able to replicate and improve the findings from (Gómez-Cañón et al., 2021) on a different dataset and we argue that a human-centered approach that selects informative instances for active learning might result beneficial for personalization. Particularly, the proposed HC and MIX methods appear to significantly outperform the random baseline for SGD and XGB models. This finding leads us to promote the creation of subjectivity-aware machine learning methods which could have a high impact in novel applications of immersion in virtual reality (Warp et al., 2022) and emotion-based music recommendation (Grekow, 2021; Tarnowska, 2021)– several other tasks display low inter-rater agreement too: music auto-tagging (Bigand & Aucouturier, 2013), music similarity and diversity (Flexer et al., 2021; Porcaro et al., 2022), automatic chord estimation (Koops et al., 2019), and beat tracking (Holzapfel et al., 2012). In short, MER has been openly criticized due to the subjectivity issue (Gómez-Cañónet al., 2021) – however we advocate for "*embracing subjectivity and potentially leveraging the opportunities it offers for better learning*" (Rizos & Schuller, 2020).

Acknowledgements This work was funded by the European Commission under the TROMPA project (H2020 770376) and the Project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033 funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI). We thank our music enthusiasts for allowing us to develop and improve our platform, and to Cynthia Liem (TU Delft) for the valuable feedback when designing the platform.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by JSGC and NGP. The first draft of the manuscript was written by JSGC and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The TROMPA-MER dataset is publicly available at the visualization platform (https://trompa-mtg.upf.edu/vis-mtg-mer/). Data including extracted features and models are available from the corresponding author.

Code availability The active learning methodology is openly available at a Github repository (https://github.com/juansgomez87/vis-mtg-mer).

Declarations

Human and animal ethics Not applicable.

Ethics approval and consent to participate Ethical approval from UPF ethical committee CIREP (attached). The Information sheet (attached) was seen and approved by each participant during registration to the platform.

Consent for publication See above.

Conflict of interests Not applicable.

References

Aggarwal, C. C., et al. (2014). Active learning: a survey. In *Data classification: algorithms and applications, pp. 571–605. CRC Press, New York*.

Agres, K. R., et al. (2021). Music, computing, and health: a roadmap for the current and future roles of music technology for health care and well-being. *Music & Science*, *4*, 1–32. https://doi.org/10.1177/2059204321997709.

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2016). Studying emotion induced by music through a crowdsourcing game. *Information Processing and Management*, *52*(1), 115–128. https://doi.org/10.1016/j.ipm.2015.03.004.

Aljanaki, A., Yang, Y. -H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PloS One*:1–22. https://doi.org/10.1371/journal.pone.0173392.

Barrett, L. F. (2017). *How emotions are made: the secret life of the brain*. Houghton Mifflin Harcourt.

Barthet, M., Fazekas, G., & Sandler, M. (2013). Music emotion recognition: from content- to context-based models. In *From sounds to music and emotions, pp. 228–252. Springer, Heidelberg*.

Bigand, E., & Aucouturier, J.-J. (2013). Seven problems that keep mir from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems*, *41*(3), 483–497. https://doi.org/10.1007/s10844-013-0251-x.

Bogdanov, D., et al. (2019). The mtg-jamendo dataset for automatic music tagging. In *Machine learning for music discovery workshop, international conference on machine learning (ICML 2019)*, pp. 1–3.

Budd, M. (1992). *Music and the emotion*. Routledge.

Cespedes-Guevara, J., & Dibben, N. (2021). Promoting prosociality in colombia: is music more effective than other cultural interventions? *Musicae Scientiae*, *25*(3), 332–357. https://doi.org/10.1177/1029864921101350 5.

Céspedes-Guevara, J., & Eerola, T. (2018). Music communicates affects, not basic emotions - a constructionist account of attribution of emotional meanings to music. *Frontiers in Psychology*, *9*(Feb), 1–19. https://doi.org/10.3389/fpsyg.2018.00215.

Chen, Y. -A., et al. (2014). Linear regression-based adaptation of music emotion recognition models for personalization. In *Proceedings of the IEEE international conference on acoustic, speech and signal processing* (ICASSP), pp. 2149–2153.

Chen, Y. -A., et al. (2017). Component tying for mixture model adaptation in personalization of music emotion recognition. *IEEE/ACM Transactions on Audio Speech, and Language Processing*, *25*(7), 1409–1420. https://doi.org/10.1109/TASLP.2017.2693565.

Cohn, D., et al. (1994). Improving generalization with active learning. *Machine Learning*, *15*, 201–221. https://doi.org/10.1007/BF00993277.

Cowen, A. S., et al. (2019). What music makes us feel: at least 13 dimensions organize subjective experiences associated with music across different cultures. Proceedings of the National Academy of Sciences:1–11. https://doi.org/10.1073/pnas.1910704117.

Dufour, I., & Tzanetakis, G. (2021). Using circular models to improve music emotion recognition. *IEEE Transactions on Affective Computing*, *12*(3), 666–681. https://doi.org/10.1109/TAFFC.2018.2885744.

Eerola, T. (2018). Music and emotion. In R. Bader S. Koelsch (Eds.) *Handbook of systematic musicology, pp. 539–556. Springer, Switzerland*. https://doi.org/10.1007/978-3-662-55004-5.

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology Music*, *39*(1), 18–49. https://doi.org/10.1177/0305735610362821.

Eerola, T., et al. (2021). Online data collection in auditory perception and cognition research: recruitment, testing, data quality and ethical considerations. *Auditory Perception & Cognition*:1–30. https://doi.org/10.1080/25742442.2021.2007718.

Ekman, P. (1992). Are there basic emotions. *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295x.99.3.550.

Eyben, F., et al. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on multimedia*, New York, pp. 835–838.

Flexer, A., Lallai, T., & Rašl, K. (2021). On evaluation of inter- and intra-rater agreement in music recommendation. *Transactions of the International Society for Music Information Retrieval*, *4*(1), 182–194. https://doi.org/10.5334/tismir.107.

Gallagher, R. J., et al. (2021). Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts, EPJ Data Science, vol. 10(1). https://doi.org/10.1140/epjds/s13688-021-00260-3.

Gebru, T., et al. (2018). *Datasheets for datasets*. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Gómez-Cañón, J. S., et al. (2020). Joyful for you and tender for us: the influence of individual characteristics and language on emotion labeling and classification. In *Proceedings of the 21st international society for music information retrieval conference, Montréal, Canada* (Online), pp. 853–860.

Gómez-Cañón, J. S., et al. (2021). Music emotion recognition: toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Processing Magazine, vol. 38. https://doi.org/10.1109/MSP.2021.3106232.

Gómez-Cañón, J. S., et al. (2021). Let's agree to disagree: consensus entropy active learning for personalized music emotion recognition. In *Proceedings of the 22nd international society for music information retrieval conference (ISMIR), pp. 237–245*.

Grekow, J. (2021). Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*, *57*(3), 531–546. https://doi.org/10.1007/s10844-021-00658-5.

Grekow, J., Ras Z. W., Wieczorkowska, A., & Tsumoto, S. (Eds.) (2021). *Music recommendation based on emotion tracking of musical performances*, pp. 167–186. Cham: Springer. https://doi.org/10.1007/978-3-030-66450-3_11.

Gutiérrez Páez, N.F., et al. (2021). Emotion annotation of music: a citizen science approach. In D. Hernández-Leo, R. Hishiyama, G. Zurita, B. Weyers, A. Nolte, & H. Ogata (Eds.) *Collaboration technologies and social computing, pp. 51–66. Springer, Cham*. https://doi.org/10.1007/978-3-030-85071-5_4.

Han, D., et al. (2022). A survey of music emotion recognition. *Frontiers of Computer Science*, *16*(6), 166335. https://doi.org/10.1007/s11704-021-0569-4.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X.

Holzapfel, A., et al. (2012). Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio Speech, and Language Processing*, *20*(9), 2539–2548. https://doi.org/10.1109/TASL.2012.2205244.

Hong, Y., Chau, C.-J., & Horner, A. (2017). An analysis of low-arousal piano music ratings to uncover what makes calm and sad music so difficult to distinguish in music emotion recognition. *Journal of the Audio Engineering Society*, *65* (4), 304–320. https://doi.org/10.17743/jaes.2017.0001.

Honing, H. (2021). Lured into listening: engaging games as an alternative to reward-based crowdsourcing in music research. *Zeitschrift für Psychologie*, *229*, 1–6. https://doi.org/10.1027/2151-2604/a000474.

Hu, X., Chen, J., & Wang, Y. (2021). University students' use of music for learning and well-being: a qualitative study and design implications. *Information Processing and Management*, *58*(1), 1–14. https://doi.org/10.1016/j.ipm.2020.102409.

Jacoby, N., et al. (2020). Cross-cultural work in music cognition: challenges, insights, and recommendations. *Music Perception*, *37*(3), 185–195. https://doi.org/10.1525/mp.2020.37.3.185.

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*, *10*(3), 235–266. https://doi.org/10.1016/j.plrev.2013.05.008.

Juslin, P. N. (2019). *Musical emotions explained*. Oxford University Press.
Kim, Y. E., Schmidt, E., & Emelle, L. (2008). Moodswings: a collaborative game for music mood label collection. In *Proceedings of the 9th international society for music information retrieval (ISMIR), pp. 231–236*.
Kim, Y. E., et al. (2010). Music emotion recognition: a state of the art review. In *Proceedings of the 11th international society for music information retrieval conference, pp. 255–266*.
Koops, H. V., et al. (2019). Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, *48*(3), 232–252. https://doi.org/10.1080/09298215.2019.1613436.
Krippendorff, K. H. (2004). *Content analysis: An introduction to its methodology*, 2nd edn. SAGE Publications.
Lange, E. B., & Frieler, K. (2018). Challenges and opportunities of predicting musical emotions with perceptual and automatized features. *Music Perception: An Interdisciplinary Journal*, *36*(2), 217–242. https://doi.org/10.1525/mp.2018.36.2.217.
Laurier, C. (2011). *Automatic classification of musical mood by content-based analysis*. Dissertation, Universitat Pompeu Fabra.
Law, E.L.M., Von Ahn, L., Dannenberg, R.B., & Crawford, M. (2007). Tagatune: a game for music and sound annotation. In *Proceedings of the 8th international society for music information retrieval (ISMIR), pp. 361–364*.
Liebetrau, J., & Schneider, S. (2013). Music and emotions: a comparison of measurement methods. In *134th convention of the audio engineering society,* Rome, Italy, pp. 1–7.
Meyer, L. B. (1961). *Emotion and meaning*. University of Chicago Press, Chicago.
Panda, R., Malheiro, R. M., & Paiva, R. P. (2020). Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*:1–20. https://doi.org/10.1109/TAFFC.2020.3032373.
Panda, R., Rui, R. M., & Paiva, P. (2018). Musical texture and expressivity features for music emotion recognition. In *Proceedings of the 19th international society for music information retrieval conference,* Paris, France, pp. 383–391.
Poli, A., et al. (2021). A preliminary study on the correlation between subjective sound quality perception and physiological parameters. In *150th convention of the audio engineering society, online,* pp. 1–7.
Porcaro, L., Gómez, E., & Castillo, C. (2022). Perceptions of diversity in electronic music: the impact of listener, artist, and track characteristics. In *Proceedings of the 25th ACM conference on computer-supported cooperative work and social computing* (CSCW), Taipei, Taiwan, pp. 1–26.
Prabhakaran, V., et al. (2021). On releasing annotator-level labels and information in datasets. In *Proceedings of the joint 15th linguistic annotation workshop (LAW) and 3rd designing meaning representations (DMR) workshop, Punta Cana, Dominican Republic,* pp. 133–138.
Rizos, G., & Schuller, B. W. (2020). Average jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty. In M.-J. Lesot et al. (Eds.) *Information processing and management of uncertainty in knowledge-based systems,* pp. 42–55. Springer, Switzerland. https://doi.org/10.1007/978-3-030-50146-4_4.
Russell, J. A. (1980). A circumplex model of affect. *Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714.
Saganowski, S., et al. (2022). Emotion recognition for everyday life using physiological signals from wearables: a systematic literature review. IEEE Transactions on Affective Computing:1–21. https://doi.org/10.1109/TAFFC.2022.3176135.
Schedl, M., Flexer, A., & Urbano, J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, *41*, 523–539. https://doi.org/10.1007/s10844-013-0247-6.
Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: recent developments and applications. *Foundations and Trends in Information Retrieval*, *8*(2-3), 127–261. https://doi.org/10.1561/1500000042.
Schedl, M., et al. (2018). On the interrelation between listener characteristics and the perception of emotions in classical orchestra music. *IEEE Transactions on Affective Computing*, *9*(4), 507–525. https://doi.org/10.1109/TAFFC.2017.2663421.
Settles, B. (2009). *Active learning literature survey*. University of Wisconsin–Madison: Computer Sciences Technical Report 1648.
Settles, B. (2012). *Active learning*. Morgan and Claypool Publishers.
Sturm, B.L. (2013). Evaluating music emotion recognition: lessons from music genre recognition?. In *Proceedings of the IEEE international conference on multimedia and expo workshops,* San Jose, USA, pp. 1–6.
Su, D., & Fung, P. (2012). Personalized music emotion classification via active learning. In *Proceedings of the second international ACM workshop on music information retrieval with user-centered and multimodal strategies,* New York, pp. 57–62.

Tarnowska, K.A., Ras Z. W., Wieczorkowska, A., & Tsumoto S. (Eds.) (2021). *Emotion-based music recommender system for tinnitus patients (EMOTIN),* pp. 197–221. Cham: Springer. https://doi.org/10.1007/978-3-030-66450-3_13.

Tkalčič, M., et al. (Eds.) (2016). *Emotions and personality in personalized services*. Switzerland: Springer. https://doi.org/10.1007/978-3-319-31413-6.

Vastenburg, M., et al. (2011). Pmri: development of a pictorial mood reporting instrument. In *CHI '11 extended abstracts on human factors in computing systems.* Chi ea '11, pp. 2155–2160.

Vempala, N. N., & Russo, F. A. (2018). Modeling music emotion judgments using machine learning methods. *Frontiers in Psychology, vol. 8*. https://doi.org/10.3389/fpsyg.2017.02239.

Warp, R., et al. (2022). Moved by sound: how head-tracked spatial audio affects autonomic emotional state and immersion-driven auditory orienting response in VR Environments. In *152nd convention of the audio engineering society,* Rome, Italy, pp. 1–7.

Warrenburg, L. A. (2020). Choosing the right tune: a review of music stimuli used in emotion research. *Music Perception*, *37*(3), 240–258. https://doi.org/10.1525/mp.2020.37.3.240.

Warrenburg, L. A. (2020). Comparing musical and psychological emotion theories. *Psychomusicology: Music Mind, and Brain*, *30*(1), 1–19. https://doi.org/10.1037/pmu0000247.

Weninger, F., et al. (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in Psychology*, *4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00292.

Yang, Y. (2018). *Towards practical active learning for classification*. Dissertation, TU Delft University.

Yang, Y. -H., & Chen, H. H. (2011). *Music emotion recognition*. CRC Press.

Yang, Y.- H., & Chen, H. H. (2012). Machine recognition of music emotion: a review. ACM Transactions on Intelligent Systems and Technology, vol. 3. https://doi.org/10.1145/2168752.2168754.

Yang, Y. -H., et al. (2007). Music emotion recognition: the role of individuality. In *Proceedings of the international workshop on human-centered multimedia,* pp. 13–22.

Yang, X., Dong, Y., & Li, J. (2018). Review of data features-based music emotion recognition methods. *Multimedia Systems*, *24*(4), 365–389. https://doi.org/10.1007/s00530-017-0559-4.

Yang, S., Reed, C. N., Chew, E., & Barthet, M. (2021). Examining emotion perception agreement in live music performance. IEEE transactions on affective computing, pp. 1–17. https://doi.org/10.1109/TAFFC.2021.3093787.

Zangerle, E., et al. (2021). Leveraging affective hashtags for ranking music recommendations. *IEEE Transactions on Affective Computing*, *12*(1), 78–91. https://doi.org/10.1109/TAFFC.2018.2846596.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, *8*(4), 494–521.