Emotion annotation of music: a citizen science approach

Nicolás Felipe Gutiérrez Páez[*1][0000−0003−3929−6980], Juan Sebastián
Gómez-Cañón[1][0000−0002−2544−6311], Lorenzo Porcaro[1][0000−0003−0218−5187],
Patricia Santos[1][0000−0002−7337−2388], Davinia
Hernandez-Leo[1][0000−0003−0548−7455], and Emilia Gómez[1][0000−0003−4983−3989]

Universitat Pompeu Fabra, Barcelona 08005, Spain
{nicolas.gutierrez,juansebastian.gomez,lorenzo.porcaro,patricia.santos,
davinia.hernandez-leo,emilia.gomez}@upf.edu

Abstract. The understanding of the emotions in music has motivated
research across diverse areas of knowledge for decades. In the field of
computer science, there is a particular interest in developing algorithms
to "predict" the emotions in music perceived by or induced to a listener.
However, the gathering of reliable "ground truth" data for modeling the
emotional content of music poses challenges, since tasks related with
annotations of emotions are time consuming, expensive and cognitively
demanding due to its inherent subjectivity and its cross-disciplinary na-
ture. Citizen science projects have proven to be a useful approach to solve
these types of problems where there is a need for recruiting collaborators
for massive scale tasks. We developed a platform for annotating emo-
tional content in musical pieces following a citizen science approach, to
benefit not only the researchers, who benefit from the generated dataset,
but also the volunteers, who are engaged to collaborate on the research
project, not only by providing annotations but also through their self
and community-awareness about the emotional perception of the music.
Likewise, gamification mechanisms motivate the participants to explore
and discover new music based on the emotional content. Preliminary
user evaluations showed that the platform design is in line with the mo-
tivations of the general public, and that the citizen science approach
offers an iterative refinement to enhance the quantity and quality of con-
tributions by involving volunteers in the design process. The usability
of the platform was acceptable, although some of the features require
improvements.

Keywords: Citizen science · Crowdsourcing · Collaborative annotation
· Music Emotion Recognition · Motivations.

1 Introduction

Citizen science has been defined as the participation of different audiences in sci-
entific research, specifically to gather, submit or analyze large quantities of data

---

[*] corresponding author

[8]. The increase of citizen science projects in recent years is attributed to the developments in information science, data informatics, graphical user interfaces and system-based web applications, as well as the refinement of the strategies for retaining volunteers and the methods for assuring data quality [19, 48]. Citizen science projects that are entirely mediated by information and communication technologies are often considered a form of crowdsourcing applied to science [47]. Yet, crowdsourcing is considered the most basic level of the participatory research approaches, and it is necessary to increase public involvement in the research process to enhance the benefits for both researchers and citizens [18, 21]. Crowdsourcing has been used in different research fields to collect, analyze and process data. Platforms such as Amazon Mechanical Turk[1] or Prolific[2] are widely used to recruit experiment participants and or collaborators for massive scale tasks such as image annotation, classification or content analysis by offering tangible rewards in the form of monetary compensation. Other platforms such as Zooniverse[3] focus on participants' intrinsic motivation (e.g., to have fun or to gain knowledge) to engage people in such tasks. In the field of music, there are several crowdsourcing initiatives to engage public participation.

One of the principal reasons why people engage and interact with music is emotion [16] - songs produce goosebumps and chills, evoke inner feelings, or provoke particular memories. This has motivated research across diverse areas of knowledge for decades: music cognition [26, 14, 27], philosophy [12], musicology [13], neuroscience [35], and computer science [49, 42, 4]. In the latter, there is a particular interest in developing algorithms that could "predict" the emotions in music perceived by or induced to a listener – known as the field of music emotion recognition (MER). Nevertheless, the gathering and study of reliable "ground truth" data for modeling the emotional content of music pose challenges, since tasks related with annotations of emotions in music are time consuming, tedious, and expensive [2], as well as cognitively demanding due to its inherent subjectivity and its cross-disciplinary nature [29]. Several studies have collected emotion annotations in music through human annotation following different approaches. On the one hand, traditional surveys provide a straightforward technique for collecting information about emotional content in music [16], but the required costs in terms of human resources, time and money reduces the chances to fully share the gathered data within the research community [29]. On the other hand, collaborative online games have been designed for the collection of music data, such as MajorMiner [33], Listen Game [43], TagATune [31], MoodSwings [28] or Emotify [3]. Even when this approach seems to engage volunteers, some studies have suggested that certain participants may perceive gamification as a trivialization of the research objects, or can demotivate participants with lower scores or low contribution rates [7, 41].

The aim of this work is to present the iterative process followed for the design of a platform to collect emotion annotations in music. This work follows a citizen

---

[1] https://www.mturk.com/

[2] https://www.prolific.co/

[3] https://www.zooniverse.org/

science approach and explores different mechanisms for engaging participation. The outline of this document is as follows. In section 2 we present the platform requirements definitions for both researchers and volunteers, and the importance of participants' motivations in the design process. Section 3 provides an overview of the designed platform and the theoretical background that support the design decisions. Section 4 describes the preliminary results obtained from different user evaluation testing scenarios and section 5 describes the conclusions and future directions.

2   Platform design requirements

2.1   "Ground truth" data for MER

In the context of MER, there are different representations of emotions, supported by a large body of music cognition research [26, 29, 15, 22]. These representations are categorized in two taxonomies: categorical/discrete [17] and dimensional/continuous [40]. In discrete/categorical models, a defined set of basic emotional descriptors (tags) is used to describe the emotions in music. Examples of such models are basic emotions (e.g., such as happy or sad) or clusters of emotional adjectives like those created for the Music Information Retrieval Evaluation eXchange (MIREX) audio mood classification task [23]. Dimensional models suggest that emotions can be scaled and measured by a continuum of descriptors or simpler multidimensional metrics, such as the circumplex model of affect (i.e., arousal refers to energy/activation, and valence relates to pleasantness/positiveness of an emotion) [40]. Both categories have been extensively studied and compared, and results suggest that even when there are clear methodological differences, there is a high correspondence between the discrete and dimensional models [15, 16, 14]. However, both approaches present drawbacks. In fact, the categorical/discrete approach is naturally ambiguous to describe rich human emotions, while the dimensional/continuous approach makes the mapping of a particular emotion to arousal-valence space vague and unreliable. Thus, it is important to provide different layers of annotation (from free-text to concise classes of emotions) in order to produce enriched, multi-level datasets. Additionally, recent research [24, 20] has shown impact of inter-rater agreement of the emotional content in music on the performance of MER systems, hence there is a need to better understand how agreement is achieved in different layers of annotation.

R1. *Annotations should follow different representations emotions to better understand the agreement among annotators (inter-rater agreement) and to obtain results that can be compared and/or integrated to other available datasets.*

Another important aspect to consider from the theoretical and methodological points of view is the distinction between perceived and induced emotions [6]. Perceived emotions refer to those recognized by the listener when interpreting musical properties (music can communicate qualities associated with emotions)

[27]. Induced or felt emotions involve psycho-physiological responses to music (emotional response to music of the listener) [30]. The challenge for annotators is the common confusion between both of these concepts, adding a layer of complexity to the annotation process.

R2. *Annotations should include information to identify whether the listener response is based on the judgement of perceived or induced emotions.*

In addition to the emotion annotation, it is of special interest to identify participants' factors that have some influence on the behavior/preference of the annotators in emotion annotation [42]. These factors may include spoken languages, birthplace, current mood, familiarity with music and their preference for the music excerpt.

R3. *Annotations should include individual characteristics of the listener, and the context in which the annotation was generated.*

2.2   Citizen science approach for collaborators' recruitment

As it was mentioned before, the problem of generating "ground truth" data for MER can be tackled from the perspective of citizen science. According to the four level categorization proposed by Haklay [21] for citizen science initiatives, this project falls in the category of distributed intelligence, since participants are asked to take basic training and then collect data or carry out data interpretation.

R4. *Training material and tutorials must be designed in order to gather quality annotations. Likewise, the designed platform should support participants' learning beyond the initial training.*

R5. *The designed platform should foster collaborations between researchers and communities, as well as to promote self-reflection and awareness regarding the research in MER.*

In citizen science research, the understanding of participants' motivations to contribute has a special relevance for projects' sustainability [37, 34, 39, 25]. Even when these studies follow different motivation theories, their results highlight the importance for project organisers to determine the motivations they may wish to appeal through the project tasks, and that it should be also noted that not all motivations will be catered for within each project [45, 46].

R6. *The designed platform should implement different incentive mechanisms to engage contributions from participants with different short and long-term motivations.*

One of the issues raised from the implementation of citizen science is regarding data sharing to allow others to build on completed work, and to foster debate and critical feedback [38]. In order for a citizen science project to produce outcomes that impact the research field, the information and data generated should be made publicly accessible in a timely manner and in a suitable form for multiple uses [9].

R7. *The designed platform should follow open science principles to make analysis and results of collected data public. To do this, data within the platform needs to conform as much as possible to existing data and process standards, and data privacy must be ensured for participants personal information.*

3 Platform description

The Music Enthusiasts (ME) web platform was designed to fulfill the identified requirements. The platform allows participants to annotate the perceived emotions in different music pieces, following circumplex model of affect and providing basic and GEMS [50] emotion tags for each quadrant. Participants can compare their annotations with the community results, and receive musical recommendations based on the provided annotations. They also have training material to understand the goals of the research and how to contribute with their annotations according to the differences between evoked and perceived emotions. Likewise, the prototype implements gamification mechanisms (scoring systems, rankings and rewards) to engage participants to contribute. Besides, researchers can create different annotation tasks (campaigns) within the platform. Each campaign consists of a determined number of songs to be annotated, and participants can choose in which campaign they want to participate. Furthermore, the platform was translated into English and Spanish. This section describes the relevant aspects of the ME platform.

3.1 Definition of the annotation task

In order to gather annotations and implement **R1**, a hybrid multi-level representation of emotions was selected. As shown in Figure 1, the core of the annotation is composed of three main sections. The first two sections collect ratings of the perceived arousal (low or high) and valence (negative or positive). To reduce the complexity and as an attempt to improve the inter-rater agreement, the annotation task only gathers binary information about arousal and valence. In this layer of annotation, we produce distinct classes from the annotations depending on the combination of arousal and valence dimensions. Thus, every annotation is mapped into one of the four quadrants of the dimensional taxonomy. The second part is a discrete tag of emotion corresponding to the specified quadrant. Eleven discrete emotion tags from the basic and the GEMS model are used, namely joy, surprise, power (quadrant 1: positive arousal and valence), anger, fear, tension (quadrant 2: positive arousal and negative valence), sadness, bitterness (quadrant 3: negative arousal and valence), peace, tenderness and transcendence (quadrant 1: negative arousal and positive valence).

To implement **R2** each annotation also includes free-text fields for explaining the reasons behind each choice. The implementation of **R3** is divided into three parts. As shown in part (C) of Figure 1, a first set of individual characteristics is collected during the annotation task. The platform stores a binary indication of the annotator's familiarity with and preference for the excerpt, as well as the
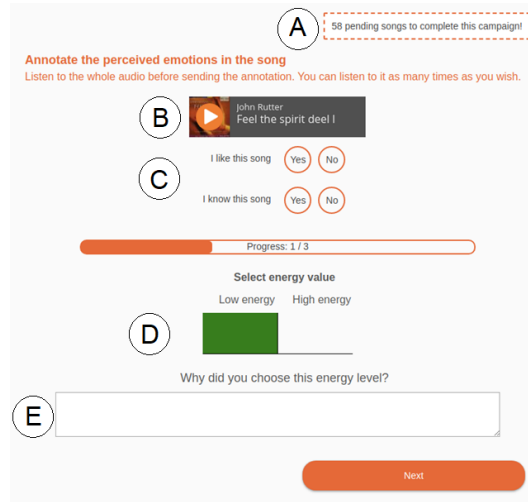
Fig. 1: Annotation interface. (A) Indicator with the number of pending songs within the campaign. (B) Music player. Participants must click on play to start listening. (C) Participants indicate if they are familiar with the song and if they like/dislike it. (D) Core of the annotation. It has three steps: arousal, valence and a discrete emotion tag (Only arousal is shown in this image). (E) Participants can add their reasons for each of their choices.

timestamp of the annotation. The second part is collected previous to an annotation. Participants report their current mood through a free-text field and a force-choice using the Pick-A-Mood tool [44]. The participant's mood report is only asked after six hours of the last report (Figure 2a). The third part of the individual characteristics is collected when users register in the platform (Figure 2b). This information is anonymized and participants have full control over the shared information (i.e., following the implementation of **R7**). They can modify or delete their information from the platform at any time they wish. Internally, the ME platform links each annotation with a unique identifier associated to the user. This is the only field that external researchers have to link contextual information with the emotion annotation.

3.2 Motivations and incentive mechanisms

This research follows the motivational model proposed and evaluated by Nov et al. [34], in which participants' motivations to contribute in a citizen science project are categorized as collective motives (importance attributed to the collective goals of the project), reward motives (potential benefits to be gained from participation, such as reputation or social interaction), norm-oriented motives (expected reactions of important others such as family and friends) and intrinsic motives.

In order to implement **R4**, **R5** and **R6**, the ME platform includes intrinsic and extrinsic incentives for users to participate in the community (Figure 3).

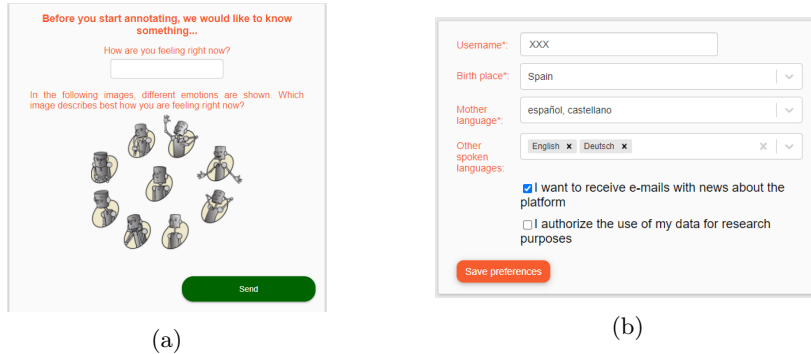(a)   (b)

Fig. 2: Individual characteristics gathering. (a) Mood selection screen. (b) Participants are the owners of their own personal data, and they can modify, erase or add information at any time. This process only affects the available contextual information associated to the user, but not the collected emotion annotations.

The platform is designed to be promoted as a platform to get recommendations based on emotional content (external reward). After completing five annotations, participants receive a musical recommendation that is stored in their profiles, creating customized playlists that can be accessed at any time. Likewise, the promotion and the information and text within the platform is designed to highlight the importance of every contribution to the research in Music Emotion Recognition (collective motives). Participants are also engaged to contribute to gain knowledge about how certain musical features relate with the emotional content in music, and to understand the wisdom of crowds and how other community members perceive music (social interaction and performance improvement). A feedback dashboard is presented to the participants after completing an annotation, so they are able to compare their perceptions with the community, to foster self-reflection and social learning. Gamification mechanisms such as a scoring system based on output-agreement, user progress statistics (number of annotations provided, number of campaigns completed) and ranking systems (fun, social interaction and public recognition) are also included.

3.3   Contributor Environment

The ME platform has been developed in the context of TROMPA (Towards Richer Online Music Public-domain Archives) project. To implement R6, the data infrastructure of the ME platform is centered around the TROMPA Contributor Environment (CE), which is a web-based platform that interlinks items hosted within different music repositories. The CE is implemented as a graph database that does not incorporate the content itself, but it refers to content hosted in web-accessible public repositories using URIs [32]. Thus, the annotations generated through the ME are stored following the Web Annotation Data Model[4] as generic core representation, which enables annotations to be shared
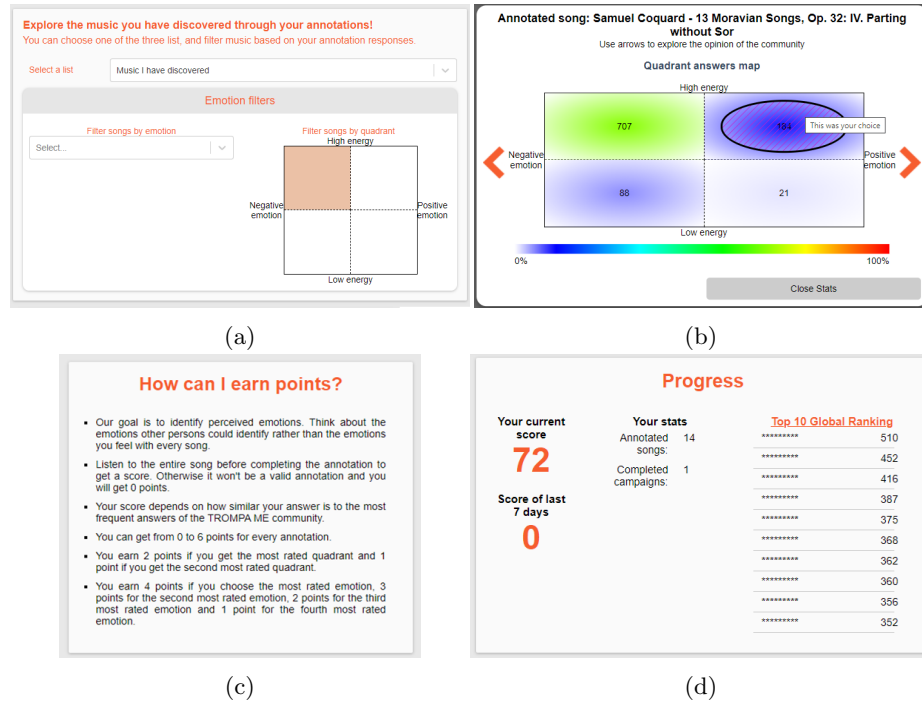
---

[4] https://www.w3.org/TR/annotation-model

Fig. 3: Incentive mechanisms. (a) My Musical profile section. Participants can explore the annotated music as well as the recommended music at any time, and apply filters based on the community results or on the MER algorithms trained with the available data. (b) Participants receive feedback of their performance compared with the community responses (output-agreement mechanism) (c) Participants can access the scoring system information at any time. Based on their responses and how close they are to the previously obtained annotations, participants receive 0 to 6 points. (d) Progress dashboard. It includes the user position within the community, number of annotations/campaigns completed and a global ranking based on the scoring system.

and reused across different hardware and software platforms. Additionally, all the annotations can be linked to the anonymized information of the participant who generated the annotation. The user information is stored either in the ME's database or in the users' decentralized data stores (Pods). Thus, users keep control of their information and can unlink or remove their information at any time, without affecting the emotion annotation.

3.4 A framework to assess outcomes and usefulness of the incentives

The impact of a citizen science project can be measured in terms of research outcomes (quantity of results, quality of results), participation outcomes (including sustained participation in time) and learning outcomes (knowledge exchange between citizens and researchers) [1]. We propose to collect data from different sources: questionnaires about usability and participants' perceptions, data logs to analyze participants' behavior within the community (number of annotations, number of registered users, average time spent in annotation tasks, analysis of navigation paths within the platform, etc.), and the results of annotations' analysis in the context of MER. These sources of information will allow us to obtain real time feedback about the outcomes of the project, and to create a collaborative environment between researchers and citizens to refine the ME platform.

4 User evaluation and results

4.1 Evaluation study

The platform has been tested in a controlled online study. The objective of this study was to evaluate the functionality and usability of the ME platform, determine the effectiveness of the incentives created within the platform and evaluate compliance with the design requirements. We created three campaigns of choral music within the platform, to be completed by the participants. Participants worked individually to complete the assigned tasks. The duration of the study was approximately two hours distributed over several days, and the content of the material was made available in Spanish and English. The study was divided into two phases. During the first phase, participants completed seven tasks (explore the *Home* section, explore the *About us* section, register in the platform, edit their user profile, annotate all the songs of a campaign, explore the annotation feedback, explore the recommended music). Participants self-assessed their ability to complete the tasks (*Did you complete the requested task?*), answered single-choice questions to validate their understanding of different concepts and mechanisms within the platform (the dimensional taxonomy, the scoring system usage, and the reward system usage), and answered Likert-scale questions to self-report their motivations to participate (table 1). During the second phase, participants were requested to complete a long annotation campaign (60 songs) during several days (spending in total around 1 hour). For evaluating the usability of the system we applied the System Usability Scale (SUS) test, which has been widely used and proved to be a reliable instrument [10, 11].

Table 1: Likert-scale questions to self-report participants' motivations. Adapted from Nov et al. [34].

Code	Question
KNO	I want to learn about musical properties and the emotions associated with them.
FUN	I want to have fun while I discover new music based on its emotional content.
REP	I like to compete with other users.
SOI	I like to be able to compare my results with other users.
COM	I want to help in this project for the automatic recognition of emotions in music

A total of 21 participants (11 male, 10 female) with an average age of 26 years old (rank=12) participated in the evaluation study. All the participants self-reported they were able to complete the requested tasks, and we verified completion using the platform log. Regarding the understanding of concepts and mechanisms, 19 participants (90.5%) understood the reward system, but only 14 (66.7%) fully understood the descriptors used to classify the music (e.g., arousal, valence, evoked emotions, perceived emotions) and 11 participants (52.4%) understood the scoring system and the feedback. Regarding the participants' motivations (Figure 4), have fun while discovering music based on its emotional content (FUN) and learn about musical properties and the emotions associated with them (KNO) were the best rated motivators, followed by altruist motives (COM). On the other hand, the scoring system and the community feedback (REP, SOI) had the lowest rating. Regarding the usability, the average SUS score was 72.9/100 (SD=8.8, rank=35). Usability and Learnability scores had similar values to the general SUS score (71.6 and 74.2 respectively).
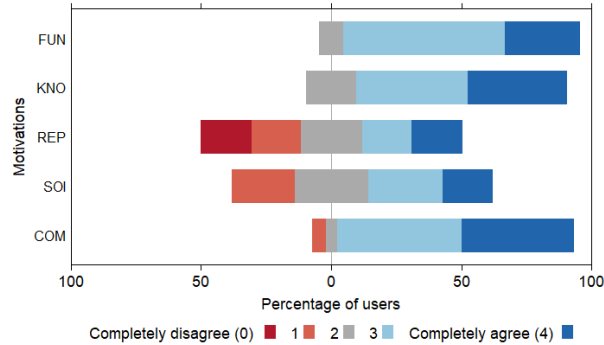

Fig. 4: Self-reported motivations to contribute in tasks for emotion annotation of music.

These results suggested that the information within the platform was not sufficient as training material, and the explanation of the scoring system was not clear enough. To tackle these issues some refinements were implemented within the platform: the different annotation fields were merged into one single screen

to make participants aware of their choices at any time during the annotation task (Figure 5a). Additionally, a new field was added to the annotation to collect a free-text tag of the most salient emotion perceived within the excerpt in the native language of the participant, adding a new level of emotion annotation to the fixed categories for each musical excerpt and consistent with research from neuroscience [5]. Predefined options were added as suggestions for the free-text fields for explaining the reasons behind each choice. Moreover, a help button was added in the annotation section, so the participants could access a detailed explanation of how to complete the annotation task and how the scoring and reward systems works (Figure 5b). Furthermore, the platform was translated into two additional languages (Italian and Dutch) to reach a higher audience.

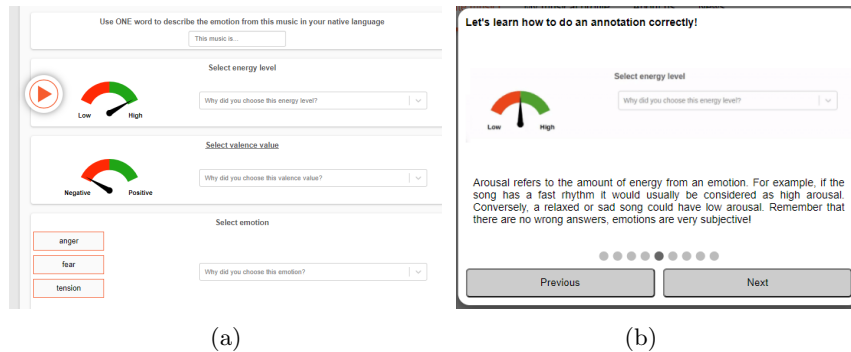(a)   (b)

Fig. 5: Platform refinements.(a) Annotation fields were merged into one single screen. (b) New help menu added during the annotation tasks. This menu is mandatory for the first annotation of new users or users with no recent annotations, and can be accessed by any user in the annotation section.

4.2   Evaluation in real conditions

After the refinements, the ME platform was released[5] and three different engagement campaigns were executed: a first contest with two external rewards for the top contributors (29th of June to 5th of July 2020), a second contest with three additional external rewards for the top contributors and for volunteers who completed an usability survey (14th to 20th of October 2020) and a long promotion campaign without external rewards (January 2021). Regarding the usability, the average SUS score was 72.9/100 (SD=13.8, rank=47.5). The results of the t-test showed that there is no significant differences between the results from the version used in the evaluation study and the results in real conditions ($t = 0.003$, $p\text{-}value > 0.05$). Additionally, participants were asked about their perception regarding the music they discovered through the platform (Figure 6), and results suggest that participants discovered completely new music

---

[5] https://enthusiasts.trompamusic.eu/

when annotating (DM1) since most of it was music that they do not tend to listen (REC1). Still, the recommendations require refinement to make them more appealing for participants (DM2, REC2).
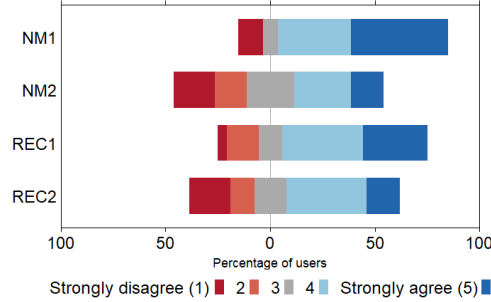

Fig. 6: Perception of users about the musical recommendation and the discovered music. **NM1:** *I discovered new music using this application.* **NM2:** *I found new music that I like during the annotation campaigns.* **REC1:** *I thought that the music recommendations received were distant from what I usually listen to.* **REC2:** *I would like to receive music recommendations more in line with my taste*

The log data recorded from the platform usage was extracted (from 1st of March 2020, until 1st of April 2021) to analyze the users behavior and to evaluate the effectiveness of the implemented incentives. At the time of the data extraction, there were 189 registered users and there were 3556 annotations. As it can be seen, both campaigns with external rewards had a significant impact in the number of registered users (Figure 7a) and the generated annotations (Figure 7b). Additionally, the users' navigation data (Figure 7c) suggest that users do not use the help after the first mandatory display. Likewise, they do not tend to access their user settings. Furthermore, the musical reward (music recommendation) and the rankings have similar accesses, suggesting that participants main motivation is to have fun. The reliability of the collected data using Krippendorff's coefficient α was measured to understand the importance of inter-rater agreement on the collected annotations. In summary, we obtained: (1) Arousal=0.505, (2) Valence=0.364, and (3) Emotion=0.192. Despite the low agreement obtained in the annotations (which is consistent with previous studies [4, 24, 20]), this motivates to use the fine-grained annotations (i.e., free-text) produced in this study to develop more personalized and context-sensitive MER applications [49].

5   Conclusions

The developed platform supports the collaborative gathering of "ground truth" data for MER algorithms, while engaging participants in the tasks by providing learning in the research field, self and community-awareness about the emotional perception of the music, and implementing gamification mechanisms (scoring systems, rankings and rewards). The results of the preliminary tests indicate that
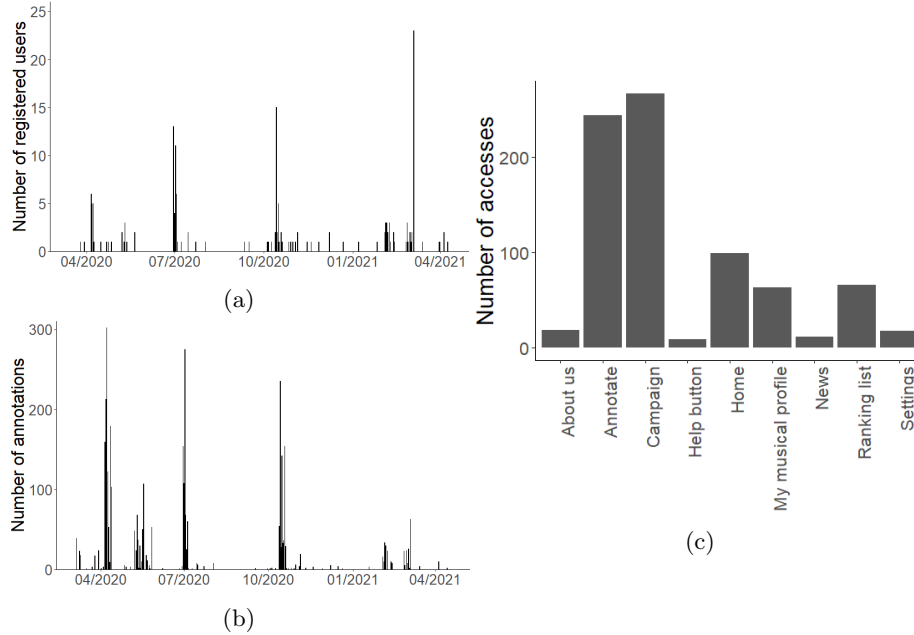
Fig. 7: Partial results of log data analysis. (a) Number of registered users over time. (b) Number of generated annotations over time. (c) Platform modules' accesses count.

contributors have intrinsic motivations to participate, i.e., they enjoy discovering new music and want to learn about the emotional content in music while contributing to the project. During the first months of deployment, tangible rewards worked as a strong motivator to engage new users in the platform. Nevertheless, according to the participants motivations, other incentives such as the recommendation system or the gamification mechanisms should be refined in order to provide long term motivation, and to provide a better landing page to make the project goals and the expected outcomes clearer (for both, volunteers and researchers). These required refinements are in line with the usability results, since the ME platform is in the range of "acceptable" applications (72.9 points over 100), and with the analysis made to the platform log data. Concerning the inter-rater agreement, the preliminary analysis shows low agreement, compared with other similar projects, but it is noteworthy that neither the individual characteristics nor the free-text field to explain participants' choices have been yet analyzed. Further analysis will include assessment of participants understanding of emotional representations and the difference between perceived and evoked. Likewise, the platform log data should include users' time line, similar to the ones proposed by Ponciano and Brasileiro [36] to define participants' engagement profiles and refine the incentive mechanisms.

References

1. Abu Amsha, O., Schneider, D.K., Fernandez-Marquez, J.L., Da Costa, J., Fuchs, B., Kloetzer, L.: Data analytics in citizen cyberscience: Evaluating participant learning and engagement with analytics. Human Computation **3**(1), 69–97 (Dec 2016). https://doi.org/10.15346/hc.v3i1.5

2. von Ahn, L.: Games with a purpose. Computer **39**(6), 92–94 (2006). https://doi.org/10.1109/MC.2006.196

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

4. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PLoS One pp. 1–22 (2017)

5. Barrett, L.F.: How Emotions are Made: The Secret Life of the Brain. Houghton Mifflin Harcourt (2017)

6. Barthet, M., Fazekas, G., Sandler, M.: Music emotion recognition: From content-to context-based models. In: From Sounds to Music and Emotions. pp. 228–252. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)

7. Baruch, A., May, A., Yu, D.: The motivations, enablers and barriers for voluntary participation in an online crowdsourcing platform. Computers in Human Behavior **64**, 923–931 (2016). https://doi.org/10.1016/j.chb.2016.07.039

8. Bonney, R., Phillips, T.B., Ballard, H.L., Enck, J.W.: Can citizen science enhance public understanding of science? Public Understanding of Science **25**(1), 2–16 (2016). https://doi.org/10.1177/0963662515607406, pMID: 26445860

9. Brenton, P., von Gavel, S., Vogel, E., Lecoq, M.E.: Technology infrastructure for citizen science, pp. 63–80. UCL Press (2018), http://www.jstor.org/stable/j.ctv550cf2.12

10. Brooke, J.: Sus - a quick and dirty usability scale. In: Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester, B. (eds.) Usability evaluation in industry. pp. 189–194 (1996)

11. Brooke, J.: Sus: a retrospective. Journal of usability studies **8**(2), 29–40 (2013)

12. Budd, M.: Music and the Emotion. Routledge (1992)

13. Cook, N.: Beyond the Score - Music as performance. Oxford University Press (2013)

14. Eerola, T.: Music and emotion. In: Bader, R., Koelsch, S. (eds.) Handbook of Systematic Musicology, chap. Music and Emotion, pp. 539–556. Springer (2018)

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

16. Eerola, T., Vuoskoski, J.K.: A review of music and emotion studies: Approaches, emotion models, and stimuli. Music Perception: An Interdisciplinary Journal **30**(3), 307–340 (2013)

17. Ekman, P.: Are there basic emotions. Psychological Review **99**(3), 550–553 (1992)

18. English, P., Richardson, M., Garzón-Galvis, C.: From crowdsourcing to extreme citizen science: Participatory research for environmental health. Annual Review of Public Health **39**(1), 335–350 (2018). https://doi.org/10.1146/annurev-publhealth-040617-013702, pMID: 29608871

19. Europe, S.: Science europe briefing paper on citizen science (2018), https://www.scienceeurope.org/media/gjze3dv4/se_briefingpaper_citizenscience.pdf

20. Gómez-Cañón, J.S., Cano, E., Herrera, P., Gómez, E.: 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. pp. 853–860. Montréal, Canada (2020)

21. Haklay, M.: Citizen science and volunteered geographic information: Overview and typology of participation. In: Sui, D., Elwood, S., Goodchild, M. (eds.) Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice, pp. 105–122. Springer Netherlands, Dordrecht (2013). https://doi.org/10.1007/978-94-007-4587-2_7

22. Hallam, S., Cross, I., Thaut, M.: The Oxford Handbook of Music Psychology. Oxford University Press (2016)

23. Hu, X., Downie, S., Laurier, C., Bay, M., Ehmann, A.: The 2007 mirex audio mood classification task: Lessons learned. In: Proc. 9th Int. Conf. Music Inf. Retrieval. pp. 462–467 (2008)

24. Hu, X., Yang, Y.H.: Cross-dataset and cross-cultural music mood prediction: A case on Western and Chinese Pop songs. IEEE Transactions on Affective Computing **8**(2), 228–240 (2017)

25. Jennett, C., Cox, A.L.: Digital citizen science and the motivations of volunteers. In: The Wiley Handbook of Human Computer Interaction, chap. 39, pp. 831–841. John Wiley & Sons, Ltd (2018). https://doi.org/10.1002/9781118976005.ch39

26. Juslin, P.N.: Handbook of Music and Emotion: Theory, Research, Applications. Oxford University Press, Oxford (2010)

27. Juslin, P.N.: Musical Emotions Explained. Oxford University Press, Oxford (2019)

28. Kim, Y.E., Schmidt, E.M., Emelle, L.: Moodswings: A collaborative game for music mood label collection. In: Ismir. vol. 8, pp. 231–236 (2008)

29. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J., Speck, J.A., Turnbull, D.: Music emotion recognition: A state of the art review. In: Proc. ismir. vol. 86, pp. 937–952 (2010)

30. Krumhansl, C.L.: An exploratory study of musical emotions and psychophysiology. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale **51**(4), 336 (1997)

31. Law, E.L., Von Ahn, L., Dannenberg, R.B., Crawford, M.: Tagatune: A game for music and sound annotation. In: ISMIR. vol. 3, p. 2 (2007), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.106.7184

32. M. Weigl, D., Goebl, W., Crawford, T., Gkiokas, A., F. Gutierrez, N., Porter, A., Santos, P., Karreman, C., Vroomen, I., C. S. Liem, C., Sarasúa, A., van Tilburg, M.: Interweaving and enriching digital music collections for scholarship, performance, and enjoyment. In: 6th International Conference on Digital Libraries for Musicology. p. 84–88. DLfM '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3358664.3358666

33. Mandel, M.I., Ellis, D.P.: A web-based game for collecting music metadata. Journal of New Music Research **37**(2), 151–165 (2008). https://doi.org/10.1080/09298210802479300

34. Nov, O., Arazy, O., Anderson, D.: Scientists@home: What drives the quantity and quality of online citizen science participation? PLOS ONE **9**(4), 1–11 (04 2014). https://doi.org/10.1371/journal.pone.0090375
35. Patel, A.D.: Music as a Transformative Technology of the Mind: An Update. In: Honing, H. (ed.) The origins of musicality, chap. 5. MIT Press (2018)
36. Ponciano, L., Brasileiro, F.: Finding Volunteers' Engagement Profiles in Human Computation for Citizen Science Projects. Human Computation **1**(2), 1–17 (2014). https://doi.org/10.15346/hc.v1i2.12
37. Raddick, M.J., Bracey, G., Gay, P.L., Lintott, C.J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A.S., Vandenberg, J.: Galaxy zoo: Motivations of citizen scientists. Astronomy Education Review **12**(1) (2013). https://doi.org/10.3847/AER2011021
38. Resnik, D.B., Elliott, K.C., Miller, A.K.: A framework for addressing ethical issues in citizen science. Environmental Science & Policy **54**, 475–481 (2015). https://doi.org/10.1016/j.envsci.2015.05.008
39. Rotman, D., Hammock, J., Preece, J., Hansen, D., Boston, C., Bowser, A., He, Y.: Motivations affecting initial and long-term participation in citizen science projects in three countries. in iconference, 110–124. In: iConference 2014 Proceedings (2014). https://doi.org/10.9776/14054
40. Russell, J.A.: A circumplex model of affect. Personality and Social Psychology **39**(6), 1161–1178 (1980)
41. Simperl, E., Reeves, N., Phethean, C., Lynes, T., Tinati, R.: Is virtual citizen science a game? Trans. Soc. Comput. **1**(2) (Jun 2018). https://doi.org/10.1145/3209960
42. Soleymani, M., Aljanaki, A., Yang, Y.H., Caro, M.N., Eyben, F., Markov, K., Schuller, B.W., Veltkamp, R., Weninger, F., Wiering, F.: Emotional analysis of music: A comparison of methods. In: Proceedings of the 22nd ACM International Conference on Multimedia. p. 1161–1164. MM '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2647868.2655019
43. Turnbull, D., Liu, R., Barrington, L., Lanc kriet, G.R.: A game-based approach for collecting semantic annotations of music. In: ISMIR. vol. 7, pp. 535–538 (2007)
44. Vastenburg, M., Romero Herrera, N., Van Bel, D., Desmet, P.: Pmri: Development of a pictorial mood reporting instrument. In: CHI '11 Extended Abstracts on Human Factors in Computing Systems. p. 2155–2160. CHI EA '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1979742.1979933
45. Wald, D.M., Longo, J., Dobell, A.R.: Design principles for engaging and retaining virtual citizen scientists. Conservation Biology **30**(3), 562–570 (2016). https://doi.org/10.1111/cobi.12627
46. West, S.E., Pateman, R.M.: Recruiting and retaining participants in citizen science: What can be learned from the volunteering literature? Citizen Science: Theory and Practice **1**(2) (2016). https://doi.org/10.5334/cstp.8
47. Wiggins, A., Crowston, K.: From conservation to crowdsourcing: A typology of citizen science. In: 2011 44th Hawaii International Conference on System Sciences. pp. 1–10 (2011). https://doi.org/10.1109/HICSS.2011.207
48. Wynn, J.: Citizen science in the digital age: rhetoric, science, and public engagement. The University of Alabama Press (2019)
49. Yang, Y.H., Chen, H.H.: Music Emotion Recognition. CRC Press (2011)
50. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: characterization, classification, and measurement. Emotion **8**(4), 494 (2008). https://doi.org/10.1037/1528-3542.8.4.494