Emotion-Based Music Recommendation from Quality Annotations and Large-Scale User-Generated Tags

MARTA MOSCATI, Johannes Kepler University Linz, Austria
HANNAH STRAUSS, University of Innsbruck, Austria
PEER-OLE JACOBSEN, University of Innsbruck, Austria
ANDREAS PEINTNER, University of Innsbruck, Austria
EVA ZANGERLE, University of Innsbruck, Austria
MARCEL ZENTNER, University of Innsbruck, Austria
MARKUS SCHEDL, Johannes Kepler University Linz and Linz Institute of Technology, Austria

Emotions constitute an important aspect when listening to music. While manual annotations from user studies grounded in psychological research on music and emotions provide a well-defined and fine-grained description of the emotions evoked when listening to a music track, user-generated tags provide an alternative view stemming from large-scale data. In this work, we examine the relationship between these two emotional characterizations of music and analyze their impact on the performance of emotion-based music recommender systems individually and jointly. Our analysis shows that (i) the agreement between the two characterizations, as measured with Cohen's κ coefficient and Kendall rank correlation, is often low, (ii) Leveraging the emotion profile based on the intensity of evoked emotions from high-quality annotations leads to performances that are stable across different recommendation algorithms; (iii) Simultaneously leveraging the emotion profiles based on high-quality and large-scale annotations allows to provide recommendations that are less exposed to the low accuracy that algorithms might reach when leveraging one type of data, only.

CCS Concepts: • **Information systems** → **Recommender systems**; Music retrieval; • **Applied computing** → *Psychology*; • **Human-centered computing** → *User studies*.

Additional Key Words and Phrases: Music Recommender Systems, Emotion-based Recommender Systems, Annotation Study, Music, Emotions

ACM Reference Format:
Marta Moscati, Hannah Strauß, Peer-Ole Jacobsen, Andreas Peintner, Eva Zangerle, Marcel Zentner, and Markus Schedl. 2024. Emotion-Based Music Recommendation from Quality Annotations and Large-Scale User-Generated Tags. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP'24), July 1–4, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3627043.3659540

1 INTRODUCTION AND RELATED WORK

Emotions play a pivotal role in the experience of music listening and in the motivations behind it [2, 9, 17, 20]. Therefore, concepts from psychology have been gaining the interest of both academia [7, 14, 15] and industry [3, 22], in particular for their application to music recommender systems (MRSs), which dominate the ways music is consumed nowadays [19]. However, MRSs often rely solely upon past collective user listening behavior or on large-scale user-generated data to characterize music tracks. These data are often noisy and are not annotated within a framework that is common to all users. They are therefore subjective, reflect the conceptualization of the individual users in different ways, and lack the depth and quality that characterizes annotated data from psychologically-informed user studies. Such quality data are, however, available from human annotations of the emotional content of music and collected with well-established psychology scales designed specifically for this task [25]. Compared to user-generated data, annotations are very specific in the question asked in the process of data collection, therefore capturing very well-defined aspects of the emotional content of music. The natural question to ask is therefore whether these two characterizations of music tracks provide the same information and are therefore redundant, or if they provide two different perspectives on the emotional content of music tracks. From the point of view of MRSs, this leads to the question of whether one perspective allows for better recommendations, or if they are complementary and should be leveraged simultaneously. In this work we aim to fill the gap in the current research on MRSs by addressing these questions, which have not been touched upon yet. For this purpose, we define two research questions (RQs). *RQ1: Is the emotional content of music tracks from large-scale user-generated data consistent with that from high-quality annotations? RQ2: For emotion-based MRSs, which representation of the emotional content of music tracks allows to reach a higher accuracy of recommendations?* To address RQ1 we carry out a statistical analysis on a set of 453 music tracks for which both high-level data from a psychologically-informed user study and large-scale user-generated tags that relate to emotions are available. We define the *emotion profile* of a music track as the information regarding which emotions the track evokes when being listened to. We analyze and compare the characteristics of the emotion profiles both from a global point of view, i. e., aggregated over all music tracks, and at the level of the individual tracks. Since our analysis shows that these data are often not consistent, it serves as basis and motivation for RQ2, i. e., on the impact of each emotion profile on the accuracy of emotion-based MRSs. For this purpose, we perform extensive experiments on music recommendation using three hybrid recommender systems, in variants that leverage the emotional profile from user-generated tags, the emotional profile from high-quality psychology-informed user studies, or both simultaneously. For comparison, we also include well-established algorithms for music recommendation that are based on collaborative filtering (CF) only. Our experiments show that leveraging the information regarding the intensities of the evoked emotions, as available from psychology-informed user studies, leads all hybrid recommendation algorithms to a similar performance. In contrast, when leveraging information on the frequency of the evoked emotion, the algorithms perform differently: some reach a better accuracy with the frequency from tags, while others with the frequency from psychology-informed user-studies. Finally, leveraging information on the frequency of evoked emotions simultaneously from large-scale and high-quality human-annotated data improves the accuracy with respect to information on the intensity of the evoked emotion, and leads to results that are more stable across the recommender algorithms compared to the individual representations of the emotion frequency, with a large improvement with respect to the worse performing variant of each algorithm.

Our paper is structured as follows: in Section 2 we describe the data and methodology used to carry out our analysis and experiments. In Section 3 we report our observations on the analysis carried out to address RQ1 and RQ2. We discuss the results, limitations, and possible extensions of our work in Section 4.[1]

2 METHODOLOGY

In this section we describe the data and methodology used to compare the emotional content of music tracks as derived from high-quality data from psychology-informed user studies on the emotions evoked by music, and as estimated

[1]We provide the code for our analysis and experiments at https://github.com/hcai-mms/emo-mrs.

Table 1. GEMS-9 emotions and examples of corresponding GEMS-45 terms. For more details on the GEMS-45 terms we refer the reader to Zentner et al. [26]. The dots indicate that more terms are present in the full set of GEMS-45 terms.

GEMS-9 emotions	Examples of GEMS-45 terms
Tenderness	Tender, Sensual, ...
Joyful Activation	Joyful, Stimulated, ...
Transcendence	(Feeling of) Transcendence, Fascinated, ...
Peacefulness	Calm, Relaxed, ...
Nostalgia	Nostalgic, Sentimental, ...
Wonder	(Filled with) Wonder, Allured, ...
Power	Energetic, Triumphant, ...
Sadness	Sad, Sorrowful, ...
Tension	Tense, Nervous, ...

Table 2. Characteristics of the set of listening events used in the recommendation experiments.

# Tracks	# Users	# Interactions
453	38,601	428,613

based on large-scale user-generated data from music streaming platforms. We also describe the experimental setup for carrying out our experiments on emotion-based music recommendation.

We consider a set of n_t music tracks and n_e emotions and represent the emotional profile of a music track over the n_e emotions as an n_e-dimensional vector. Consequently, we describe the emotional profiles of the n_t tracks as an $n_t \times n_e$ matrix. From the Emotion-to-Music Mapping Atlas (EMMA) database [1, 5, 21, 23],[2], we use the $n_t = 453$ music tracks that were annotated in 2023. The emotional effects of these tracks were rated with the Geneva Emotion Music Scale (GEMS) [26]. The GEMS-9 [26] is a scale to assess the emotions evoked while listening to a music track. The scale consists of nine dimensions (Tenderness, Joyful Activation, Transcendence, Peacefulness, Nostalgia, Wonder, Power, Sadness, Tension). In user studies, annotators assign a value to each dimension based on their emotional experience. We refer to the GEMS-9 dimensions as GEMS-9 emotions throughout the paper, and therefore $n_e = 9$ in our work. To extract the emotion profile of a track from large-scale user-generated data, we use the tags provided by the Music4All-Onion dataset [13], which is a large-scale multi-modal dataset for content-based MRSs. The tags available in the dataset were extracted with the Last.fm API[3] with the method `track.getTopTags`, which provides the most frequent tags attached to the track by the users of Last.fm. Alongside each of the most frequent tags, the API provides an integer weight ranging from 1 to 100 and representing the frequency with which each of the most frequent tags was associated to the track; 100 is associated with the most frequent tag, and the remaining weights are rescaled to the frequency of the most frequent tag. To extract the emotion profile of a track from the tags, we first define a set of terms that refer to the GEMS-9 emotions. This set consists of the GEMS-9 emotions themselves, as well as the 45 terms that have been reported to refer to those emotions [26], and which we refer to as GEMS-45 terms. The set of GEMS-9 emotions and examples of the corresponding GEMS-45 terms are displayed in Table 1.[4] We stem these terms as well as the Last.fm tags using Porter stemmer, and select the tags that contain at least one word stem referring to any of the GEMS emotions or terms. To get the weight of each GEMS-9 emotion, we sum the weights of the tags that contain a stemmed version of either the GEMS-9 emotion itself, or of one of the GEMS-45 terms corresponding to that emotion. We normalize the weights of each track such that they add up to 1 and rescale them by 100. We refer to these data as **Tags**, in short, and to the corresponding profile as $P^{\mathbf{Tags}}$. For the emotion profile from high-quality psychology-informed user studies, we use the annotations from the EMMA database: tracks were annotated by 15 annotators on average, who reported on a scale

[2]https://musemap-tools.uibk.ac.at/emma/
[3]https://www.last.fm/api
[4]The full set of GEMS-45 terms as well as specific instructions on how to use it is available upon request to the authors of Zentner et al. [26]. Since it is mandatory to obtain a permission to use the set, we only provide a few examples of the GEMS-45 terms for each GEMS-9 emotion.

from 0 to 100 the amount of each GEMS-9 emotion evoked by the track. We refer to these data as **EMMA**. The database also contains track emotion profiles consisting of values from 0 to 100, representing the average value assigned by annotators. We refer to this profile as $P_{\mathrm{i}}^{\mathbf{EMMA}}$. Since $P^{\mathbf{Tags}}$ contains information on the frequency, rather than the intensity of evoked emotions, we also compute a profile $P_{\mathrm{f}}^{\mathbf{EMMA}}$ for which entries consist of the percentage of times that raters annotated a specific track and for a specific emotion with a value greater than 0, relative to the number of times that the track was annotated. We also obtain binary emotion profiles, i. e., treating the emotions as labels. To this purpose, we convert the profile matrices P to binary matrices $P_{\mathrm{bin}} \in \{0,1\}^{n_t \times n_e}$. For $P^{\mathbf{Tags}}$ we associate the emotion to the track if at least one of the terms corresponding to the emotion is present in at least one of the tags from the Music4All-Onion dataset. For $P^{\mathbf{EMMA}}$, we considered two approaches. First, we apply majority voting to $P_{\mathrm{f}}^{\mathbf{EMMA}}$ by setting a threshold of 0.5 for binarization. We refer to the resulting binary profile as $P_{\mathrm{f,\,bin}}^{\mathbf{EMMA}}$. Alternatively, we set the threshold for $P_{\mathrm{i}}^{\mathbf{EMMA}}$ to the value for which $P_{\mathrm{i,\,bin}}^{\mathbf{EMMA}}$ has the same sparsity as $P_{\mathrm{bin}}^{\mathbf{Tags}}$. To analyze the difference in global patterns of the emotion profiles, we look at the frequency of each emotion for the binarized profiles $P_{\mathrm{i,\,bin}}^{\mathbf{EMMA}}$, $P_{\mathrm{f,\,bin}}^{\mathbf{EMMA}}$, and $P_{\mathrm{bin}}^{\mathbf{Tags}}$. Then, to quantify the difference of the distribution of emotions over all tracks, for the binarized profile $P_{\mathrm{bin}}^{\mathbf{Tags}}$ and either $P_{\mathrm{i,\,bin}}^{\mathbf{EMMA}}$ or $P_{\mathrm{f,\,bin}}^{\mathbf{EMMA}}$, we rank the GEMS-9 emotions in descending order of frequency and compute the Kendall rank correlation coefficient τ. We then analyze the agreement of the profiles at the level of the individual tracks. To this purpose we first compute Cohen's κ coefficient between the binarized profile $P_{\mathrm{bin}}^{\mathbf{Tags}}$ and either $P_{\mathrm{i,\,bin}}^{\mathbf{EMMA}}$ or $P_{\mathrm{f,\,bin}}^{\mathbf{EMMA}}$. We compute this over all entries, as well as over each GEMS-9 emotion, i. e., between the entries of the columns of the binarized profile matrices. The coefficient is defined as $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the fraction of entries that assume the same value in the two binarized profile matrices, and p_e is the probability of agreement by chance, estimated as $p_e = f_0^{EMMA} f_0^{tags} + f_1^{EMMA} f_1^{Tags}$, where $f_{0,1}$ represent the frequency of 0's and 1's, respectively. Finally, to quantify the agreement in the amount to which an emotion was evoked, relative to the others, we compute the Kendall τ correlation coefficient for each track and between pairs of profiles $P_{\mathrm{i,f}}^{\mathbf{EMMA}}$ and $P^{\mathbf{Tags}}$.

The recommendation experiments are performed on the subset of the listening events of the Music4All-Onion that consists of the 453 tracks included in our analysis, without restricting to any time window. As commonly done in the domain of MRSs [11, 12], we convert the listening events to binary implicit feedback with a threshold of 2 on the listening counts and apply 5-core filtering, i. e., we only consider users that listened to at least 5 different tracks and tracks listened to by at least 5 different users. Notice that this removes none of the 453 tracks considered. The characteristics of the resulting dataset are reported in Table 2. We split the data into a training, a validation, and a test set respectively consisting of 60%, 20%, and 20% of the total number of interactions, randomly selected. We perform our experiments with the recommendation library RecBole [27, 28] and consider three recommendation algorithms that allow leveraging content information on the items. We select Factorization Machines (FM) [16] since it is a generalization of standard CF algorithms to allow the inclusion of item content information. To analyze variants of this algorithm based on deep neural networks and on graph neural network, we further select Deep Factorization Machines (DeepFM) [8] and Directed Acyclic Graph Factorization Machines via Knowledge Distillation (KD_DAGF) [24]. We leverage these algorithms as emotion-based MRSs by providing either $P_{\mathrm{bin}}^{\mathbf{Tags}}$, $P_{\mathrm{i,\,bin}}^{\mathbf{EMMA}}$, $P_{\mathrm{f,\,bin}}^{\mathbf{EMMA}}$ or a concatenation of $P_{\mathrm{f,\,bin}}^{\mathbf{Tags}}$ and $P_{\mathrm{f,\,bin}}^{\mathbf{EMMA}}$, as content information on the music tracks. Therefore, each algorithm is optimized in four variants. Since we are interested in the comparison among the emotion profiles, rather than of the underlying recommendation algorithm, and due to the limited number of tracks available, we do not perform any hyperparameter optimization on the algorithms. Since the dimensionality of the emotion profiles is the same, we believe that our choice does not affect the comparison among emotion profiles, although it hinders the comparison among recommendation algorithms, which

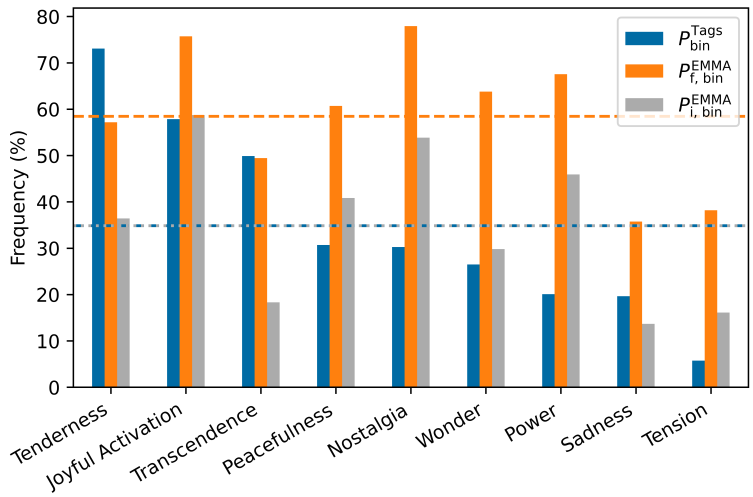

Fig. 1. Frequency of occurrence of each of the GEMS-9 emotions in the binarized profile matrices. The dashed lines indicate the frequency over all GEMS-9 emotions, in blue and grey for $P_{\text{i, bin}}^{\textbf{EMMA}}$ and $P_{\text{bin}}^{\textbf{Tags}}$, and in orange for $P_{\text{f, bin}}^{\textbf{EMMA}}$.

is not the focus of this work. We also include two well-established RSs selected for their good performance in terms of accuracy of recommendations [6], including the music domain [12]: item k-nearest-neighbors (Item-kNN) [4] and variational autoencoders for collaborative filtering (MultVAE) [10]. As baselines for comparison, we also include a RS recommending random tracks (Random) and a RS that always recommends the most popular music tracks (MostPop), defining popularity in terms of distinct listeners in the training set. All algorithms are trained with early stopping for a maximum of 500 epochs, selecting the model for which the normalized discounted cumulative gain at 10 (NDCG@10) on the validation set does not improve in the following 10 epochs.

3 RESULTS

In this section, we report our observations regarding RQ1, i. e., on the agreement between the emotions evoked by music as measured with annotations from psychology-informed user studies and as estimated from user-generated tags, and RQ2, i. e., on the impact of the emotion profiles on the accuracy of emotion-based MRSs.

We first look at the emotion profiles from a global point of view, i. e., considering the distribution of emotions over all tracks. Figure 1 shows the frequency of occurrence of each of the GEMS-9 emotions in the binarized profile matrices $P_{\text{bin}}^{\textbf{Tags}}$, $P_{\text{i, bin}}^{\textbf{EMMA}}$, and $P_{\text{f, bin}}^{\textbf{EMMA}}$. The histogram reveals that the frequency of emotions in profiles from **Tags** as measured with the binarized profile $P_{\text{bin}}^{\textbf{Tags}}$, tends to be more skewed than the frequency of emotions in profiles from **EMMA**, as measured with both binarized profiles $P_{\text{i,bin}}^{\textbf{EMMA}}$ and $P_{\text{f,bin}}^{\textbf{EMMA}}$. The discrepancy in the frequency of occurrence of emotions is also highlighted by the value of Kendall rank correlation coefficient τ measured after ranking the GEMS-9 emotions in descending order of frequency. The coefficient reaches a value of $\tau = 0.33$ with $P_{\text{i,bin}}^{\textbf{EMMA}}$ and of $\tau = 0.16$ with $P_{\text{f,bin}}^{\textbf{EMMA}}$. Although not statistically significant ($p > 0.05$), the fact that the correlations are positive but rather low indicates that even across several tracks, emotions occur with different frequency in $P_{\text{bin}}^{\textbf{Tags}}$ compared to $P_{\text{i,bin}}^{\textbf{EMMA}}$ and $P_{\text{f, bin}}^{\textbf{EMMA}}$. Table 3 shows the values of Cohen's κ coefficient computed between $P_{\text{bin}}^{\textbf{Tags}}$ and either $P_{\text{i,bin}}^{\textbf{EMMA}}$ or $P_{\text{f,bin}}^{\textbf{EMMA}}$, for each GEMS-9 emotion, and over all emotions. The coefficient takes a low value when computed over all emotions, and generally

Table 3. Cohen's κ coefficient between $P^{\textbf{Tags}}_{\text{bin}}$ and either $P^{\textbf{EMMA}}_{\text{i,bin}}$ or $P^{\textbf{EMMA}}_{\text{f,bin}}$, for each GEMS-9 emotion, and over all emotions.

Emotion	$P^{\textbf{EMMA}}_{\text{i,bin}}$	$P^{\textbf{EMMA}}_{\text{f,bin}}$
Tenderness κ	-0.036	0.045
Joyful Activation κ	0.010	0.025
Transcendence κ	-0.004	-0.007
Peacefulness κ	0.069	-0.003
Nostalgia κ	0.113	0.002
Wonder κ	-0.063	0.011
Power κ	-0.026	-0.040
Sadness κ	-0.003	-0.041
Tension κ	-0.048	-0.010
Overall κ	0.055	0.035

Table 4. NDCG@10 reached by the emotion-based MRSs leveraging $P^{\textbf{Tags}}_{\text{bin}}$, $P^{\textbf{EMMA}}_{\text{i, bin}}$, $P^{\textbf{EMMA}}_{\text{f, bin}}$, or the concatenation of $P^{\textbf{Tags}}_{\text{bin}}$ and $P^{\textbf{EMMA}}_{\text{f, bin}}$, as well as by the algorithms used for comparison.

	NDCG@10			
	$P^{\textbf{Tags}}_{\text{bin}}$	$P^{\textbf{EMMA}}_{\text{i, bin}}$	$P^{\textbf{EMMA}}_{\text{f, bin}}$	$P^{\textbf{EMMA}}_{\text{f, bin}} \bigcup P^{\textbf{Tags}}_{\text{bin}}$
FM	0.200	0.154	0.135	0.173
DeepFM	0.168	0.151	0.084	0.159
KD_DAGFM	0.145	0.154	0.214	0.227
MultVAE	0.351			
Item-kNN	0.342			
MostPop	0.025			
Random	0.013			

even lower or negative for the individual emotions. This is especially true for $P^{\textbf{EMMA}}_{\text{f, bin}}$ and indicates that on the level of individual tracks, the binarized profile $P^{\textbf{Tags}}_{\text{bin}}$ often disagrees with the binarized profile from **EMMA**, both as $P^{\textbf{EMMA}}_{\text{i, bin}}$ and as $P^{\textbf{EMMA}}_{\text{f, bin}}$. As for the ranking of the emotions for each music track, the average value of Kendall rank correlation coefficient τ computed between the rows of $P^{\textbf{Tags}}$ and the rows of either $P^{\textbf{EMMA}}_{\text{i}}$ or $P^{\textbf{EMMA}}_{\text{f}}$ is $\tau = 0.210$ for $P^{\textbf{EMMA}}_{\text{i}}$ and $\tau = 0.199$ for $P^{\textbf{EMMA}}_{\text{f}}$ (in both cases $p < 0.05$). This small positive value indicates that for the same track ranking the emotions according to $P^{\textbf{Tags}}$ often results in a different ranking than that obtained by ranking them according to $P^{\textbf{EMMA}}_{\text{i}}$ or $P^{\textbf{EMMA}}_{\text{f}}$.

Moving on to RQ2, we analyze the impact of the emotion profiles from psychologically-informed user studies and from large-scale user-generated tags on the accuracy of emotion-based MRSs. The first three columns of the first block of Table 4 show the values of the NDCG@10 reached by the emotion-based MRSs leveraging either $P^{\textbf{Tags}}_{\text{bin}}$, $P^{\textbf{EMMA}}_{\text{i, bin}}$, or $P^{\textbf{EMMA}}_{\text{f, bin}}$. The second and third block show the values reached by MultVAE, Item-kNN, MostPop, and Random. We observe that MultVAE and Item-kNN reach the highest NDCG@10 values, which confirms the observation from previous work that these algorithms are very accurate RSs [6]. Comparing the performance of the emotion-based MRSs, we observe that all algorithms show a similar performance when leveraging $P^{\textbf{EMMA}}_{\text{i, bin}}$; Therefore, $P^{\textbf{EMMA}}_{\text{i, bin}}$ seems to lead to more stable results compared to the other two representations of the emotion profile. Additionally, for each model, leveraging $P^{\textbf{Tags}}_{\text{bin}}$ or $P^{\textbf{EMMA}}_{\text{i, bin}}$ constantly reaches the most pronounced results, i. e., either $P^{\textbf{Tags}}_{\text{bin}}$ reaches the highest NDCG@10 and $P^{\textbf{EMMA}}_{\text{f, bin}}$ the lowest, or vice-versa. This behavior is in agreement with the results from the analysis of the emotion profiles, both seeming to indicate that $P^{\textbf{Tags}}_{\text{bin}}$ and $P^{\textbf{EMMA}}_{\text{f, bin}}$ contain complementary information. We therefore concatenate these two profiles to test the performance of emotion-based MRSs simultaneously leveraging large-scale user-generated data and high-quality data from psychology-informed user studies. The NDCG@10 reached by the corresponding MRSs is shown in the last column of Table 4 as $P^{\textbf{Tags}}_{\text{bin}} \bigcup P^{\textbf{EMMA}}_{\text{f, bin}}$. For all algorithms, this variant always outperforms the one based on $P^{\textbf{EMMA}}_{\text{i, bin}}$. For *KD_DAGFM*, leveraging $P^{\textbf{Tags}}_{\text{bin}} \bigcup P^{\textbf{EMMA}}_{\text{f, bin}}$ improves the performance both with respect to $P^{\textbf{Tags}}_{\text{bin}}$ and to $P^{\textbf{EMMA}}_{\text{f, bin}}$. For FM and DeepFM, $P^{\textbf{Tags}}_{\text{bin}} \bigcup P^{\textbf{EMMA}}_{\text{f, bin}}$ reaches an intermediate accuracy; However, the increase in NDCG@10 with respect to the variant of the same algorithm performing worst is always higher than the loss with respect to the variant performing best. We therefore conclude that leveraging $P^{\textbf{Tags}}_{\text{bin}}$ and $P^{\textbf{EMMA}}_{\text{f, bin}}$ simultaneously allows to reach results that are better compared to $P^{\textbf{EMMA}}_{\text{i, bin}}$, and less susceptible to the underlying model compared to $P^{\textbf{Tags}}_{\text{bin}}$ and $P^{\textbf{EMMA}}_{\text{f, bin}}$.

4 DISCUSSION AND CONCLUSION

In this work, we investigated the relationship between the emotional content of music tracks as measured with annotations using tools grounded in psychological research on music and emotions, and with large-scale user-generated tags from music streaming platforms. We then analyzed their impact on emotion-based music recommendation. Our analysis showed that there is a discrepancy between the profiles, as highlighted by considering the frequency of occurrence of each emotion over several tracks, as well as for individual tracks, both in terms of binary occurrence of emotions, and in terms of ranking of emotions according to their intensity or frequency. One of the reasons underlying the discrepancy between the annotations from psychology-informed user studies and the user-generated tags might be the intrinsic difference in the aspect of emotional content they capture. While the annotations from the EMMA database refer specifically to the emotions *evoked* when listening to a music track, user-generated tags might be more sensitive to emotions *perceived* for a track, instead. This distinction was clearly highlighted by Zentner et al. [26] when developing the GEMS. When leveraging the emotion profiles from **Tags** and from EMMA for emotion-based music recommendation, we observed that across several hybrid architectures, leveraging the emotion profile based on the intensity of evoked emotions from high-quality annotations ($P_{\text{i, bin}}^{\textbf{EMMA}}$) leads to comparable performances irrespective of the algorithm. In addition, simultaneously leveraging data from **Tags** and **EMMA** allows to provide recommendations that are less exposed to the low accuracy that algorithms might reach when leveraging one type of data, only. To be more specific, when leveraging both profiles simultaneously one algorithm reaches the highest accuracy and two algorithms reach an intermediate accuracy between the accuracies reached when leveraging one profile at the time. For these two algorithms, the improvement in NDCG@10 compared to the worst-performing variant is larger than its decrease compared to the best-performing variant.

When analyzing the performance of emotion-based MRSs, we focused on the overall accuracy of recommendations. Since the emotion profiles are intrinsic characteristics of music tracks, they do not directly relate to the track representations in terms of user-item interactions. Therefore, it would be interesting to analyze if leveraging the emotion profiles for music recommendation allows to mitigate well-known issue of RSs relying solely on collaborative data, such as cold-start scenarios or popularity bias. We leave for future work an analysis of emotion-based RSs that goes beyond accuracy. Additionally, due to the currently limited amount of annotations regarding the emotions *evoked* by music, the number of music tracks considered in this work is limited compared to the number of music tracks typically available in datasets for music recommendation, often consisting of several millions distinct tracks [18]. Although ideal, extending high-quality annotations to such large datasets with psychology-informed user studies is unfeasible. Therefore, it would be interesting to devise algorithms that allow to extend the emotion profile from high-quality psychology-informed user-studies from a small number of tracks to a set of tracks several orders of magnitudes larger, i. e., by means of semi-supervised learning and using the characteristics of the audio signal of the music tracks. The quality of the automatic annotations could then be tested – although on a smaller subset – with a user study. Finally, the full set of automatically-annotated tracks could be used as dataset for large-scale emotion-based music recommendation. We leave the development and evaluation of algorithms for automatically annotating the emotions *evoked* by a music track, as well as the use of their annotations for emotion-based music recommendation, for future work.

ACKNOWLEDGMENTS

This research was funded in whole, or in part, by the Austrian Science Funds (FWF): https://doi.org/10.55776/P33526 and https://doi.org/10.55776/DFH23.

REFERENCES

[1] Vivienne Biedermann, Hannah Strauß, and Marcel Zentner. 2019. Different genres, different emotions? An analysis of the influence of individual preferences and genre on musically evoked emotions. In *Poster presented at ICPS*.

[2] Diana Boer and Ronald Fischer. 2011. Towards a holistic model of functions of music listening across cultures: A culturally decentred qualitative approach. *Psychology of Music* 40 (03 2011), 179–200.

[3] Théo Bontempelli, Benjamin Chapus, François Rigaud, Mathieu Morlon, Marin Lorant, and Guillaume Salha-Galvan. 2022. Flow Moods: Recommending Music by Moods on Deezer. In *Proc. of ACM RecSys* (Seattle, WA, USA). 452–455.

[4] Mukund Deshpande and George Karypis. 2004. Item-Based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems* 22, 1 (2004), 143–177.

[5] Maximilian Dick, Hannah Strauß, and Marcel Zentner. 2019. How Musical Experience Influences Emotional Perception of Classical Music.. In *Poster presented at ICPS*.

[6] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proc. of ACM RecSys* (Copenhagen, Denmark). 101–109.

[7] Bruce Ferwerda, Andreu Vall, Marko Tkalcic, and Markus Schedl. 2016. Exploring Music Diversity Needs Across Countries. In *Proc. of ACM UMAP* (Halifax, Nova Scotia, Canada). 287–288.

[8] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proc. of IJCAI* (Melbourne, Australia). 1725–1731.

[9] Patrik Juslin and Petri Laukka. 2004. Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research* 33 (09 2004), 217–238.

[10] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proc. of ACM WWW* (Lyon, France). 689–698.

[11] Alessandro B. Melchiorre, Navid Rekabsaz, Christian Ganhör, and Markus Schedl. 2022. ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations. In *Proc. of RecSys*. Association for Computing Machinery, New York, NY, USA.

[12] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating Gender Fairness of Recommendation Algorithms in the Music Domain. *Information Processing & Management* 58, 5 (2021), 102666.

[13] Marta Moscati, Emilia Parada-Cabaleiro, Yashar Deldjoo, Eva Zangerle, and Markus Schedl. 2022. Music4All-Onion - A Large-Scale Multi-faceted Content-Centric Music Recommendation Dataset. In *Proc. of ACM CIKM* (Atlanta, GA, USA). 4339–4343.

[14] Marta Moscati, Christian Wallmann, Markus Reiter-Haas, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2023. Integrating the ACT-R Framework with Collaborative Filtering for Explainable Sequential Music Recommendation. In *Proc. of ACM RecSys* (Singapore, Singapore). 840–847.

[15] Markus Reiter-Haas, Emilia Parada-Cabaleiro, Markus Schedl, Elham Motamedi, Marko Tkalcic, and Elisabeth Lex. 2021. Predicting Music Relistening Behavior Using the ACT-R Framework. In *Proc. of ACM RecSys* (Amsterdam, Netherlands). 702–707.

[16] Steffen Rendle. 2010. Factorization Machines. In *Proc. of IEEE ICDM* (Sydney, Australia). 995–1000.

[17] Suvi Saarikallio, Sirke Nieminen, and Elvira Brattico. 2013. Affective reactions to musical stimuli reflect emotional use of music in everyday life. *Musicae Scientiae* 17 (03 2013), 27–39.

[18] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *Proc. of ACM SIGIR CHIIR* (Regensburg, Germany). 337–341.

[19] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[20] Thomas Schäfer, Peter Sedlmeier, Christine Städtler, and David Huron. 2013. The psychological functions of music listening. *Frontiers in psychology* 4 (08 2013), 511.

[21] Jana Serebriakova, Hannah Strauß, and Marcel Zentner. 2019. Influence of familiarity and tempo on the emotions evoked across different music genres. In *Poster presented at ICPS*.

[22] Bruno Sguerra, Viet-Anh Tran, and Romain Hennequin. 2023. Ex2Vec: Characterizing Users and Items from the Mere Exposure Effect. In *Proc. of ACM RecSys* (Singapore, Singapore). Association for Computing Machinery, New York, NY, USA, 971–977.

[23] Hannah Strauss, Julia Vigl, Peer-Ole Jacobsen, Francesca Talamini, Wolfgang Vigl, Eva Zangerle, and Marcel Zentner. 2024, in press. The Emotion-to-Music Mapping Atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts. *Behavior Research Methods* (2024, in press).

[24] Zhen Tian, Ting Bai, Zibin Zhang, Zhiyuan Xu, Kangyi Lin, Ji-Rong Wen, and Wayne Xin Zhao. 2023. Directed Acyclic Graph Factorization Machines for CTR Prediction via Knowledge Distillation. In *Proc. of ACM WSDM 2023* (Singapore, Singapore). 715–723.

[25] Marcel Zentner and Tuomas Eerola. 2010. Self-Report Measures and Models. In *Handbook of Music and Emotion: Theory, Research, Applications*.

[26] Marcel Zentner, Didier Grandjean, and Klaus Scherer. 2008. Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement. *Emotion* 8 (2008), 494–521.

[27] Wayne Xin Zhao, Yupeng Hou, Xingyu Pan, Chen Yang, Zeyu Zhang, Zihan Lin, Jingsen Zhang, Shuqing Bian, Jiakai Tang, Wenqi Sun, Yushuo Chen, Lanling Xu, Gaowei Zhang, Zhen Tian, Changxin Tian, Shanlei Mu, Xinyan Fan, Xu Chen, and Ji-Rong Wen. 2022. RecBole 2.0: Towards a More Up-to-Date Recommendation Library. In *Proc. of ACM CIKM* (Atlanta, GA, USA). 4722–4726.

[28] Wayne Xin Zhao, Shanlei Mu, Houm Yupeng, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *Proc. of ACM CIKM*. 4653–4664.