MOOD CLASSIFICATION USING LISTENING DATA

Filip Korzeniowski[1] **Oriol Nieto**[1] **Matthew C. McCallum**[1]
Minz Won[2] **Sergio Oramas**[1] **Erik M. Schmidt**[3]
[1] Pandora Media LLC., Oakland, California, USA
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Netflix Inc., Los Gatos, California, USA

ABSTRACT

The mood of a song is a highly relevant feature for exploration and recommendation in large collections of music. These collections tend to require automatic methods for predicting such moods. In this work, we show that listening-based features outperform content-based ones when classifying moods: embeddings obtained through matrix factorization of listening data appear to be more informative of a track mood than embeddings based on its audio content. To demonstrate this, we compile a subset of the Million Song Dataset, totaling 67k tracks, with expert annotations of 188 different moods collected from AllMusic. Our results on this novel dataset not only expose the limitations of current audio-based models, but also aim to foster further reproducible research on this timely topic.

1. INTRODUCTION

The estimation of moods that a given music track might evoke or empathize with is a relevant task that has been active in the Music Informatics Research (MIR) community for years [20]. This task, which is also known as music emotion recognition, has become even more prominent thanks to the advent of streaming music services with massive collections, where understanding the set of moods of each of their tracks could strongly impact the navigation, discovery, and recommendations of such collections [32]. This task has been typically approached in two different ways: i) regressing a continuous mood space such as the Arousal-Valence one [30], and then clustering such space to obtain a specific mood vocabulary [37]; or ii) classifying a given track into one or more moods, thus becoming a multi-label classification problem with a fixed vocabulary [6], which can be seen as a sub-task of the broader audio tagging problem [27]. In this work, we focus exclusively on the second approach, since it can directly impact search-by-mood applications, while methods like metric learning can potentially overcome the limitation of the fixed vocabulary [5].

© Filip Korzeniowski, Oriol Nieto, Matthew C. McCallum, Minz Won, Sergio Oramas, Erik M. Schmidt. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Filip Korzeniowski, Oriol Nieto, Matthew C. McCallum, Minz Won, Sergio Oramas, Erik M. Schmidt. "Mood Classification Using Listening Data", 21st International Society for Music Information Retrieval Conference, Montréal, Canada, 2020.

Framed under the context of music recommendation, mood recognition is particularly interesting. It has been shown that listener personality correlates not only with musical taste [29, 41], but also with genre [11], which makes the development of psychologically inspired approaches one of the most compelling challenges for recommender systems [32]. Thus, several related techniques have been presented: FocusMusicRecommender [40] makes use of the listener's behavior history to play tracks that are appropriate given the current listener's level of concentration. By incorporating the Five Factor Model [7], collaborative filtering [21] is enhanced with personality embeddings [10]. Moreover, emotions from a microblogging service have been exploited to implement an emotion-aware recommendation system [9]. Such techniques employ data beyond the actual audio signal to enhance mood-based recommenders, inspiring us to make use of listening data to classify moods to potentially improve the navigation and recommendation of large music catalogs.

The contribution of this work to the task of mood prediction is two-fold: i) we assemble a set of 67k tracks from the Taste Profile subset from the Million Song Dataset (MSD) [2] and match them with human-annotated moods available from AllMusic. [1] This is, to the best of our knowledge, the largest expert-annotated mood dataset available. And ii) by running several experiments on this proposed dataset we show how listener data are much more accurate at classifying moods than current audio-based approaches. Similarly to [17], where its authors discuss how lyrics can be useful to predict moods better than actual audio, and following the music recommendation approaches described above, we further argue that listening embeddings yield superior results due to their ability to capture information that is not straightforward to be extracted from pure audio content only.

The rest of the article is structured as follows: in Section 2 we give a formal definition of the mood classification problem. In Section 3 the data employed in this work are described. We then detail the mood classification experiments in Section 4. The results of these experiments are discussed in Section 5. Finally, we draw conclusions and consider potential future directions in Section 6.

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

2. MOOD CLASSIFICATION

Predicting moods evoked by music is often treated as an audio classification problem in the MIR community, [2] where audio data are almost exclusively used as input. In this section we give an overview of this task and its current approaches.

2.1 Problem definition

Mood tagging is a multi-label classification problem, and can be considered a subset of the broader audio tagging task where only those tags that represent moods are considered. Formally, let $\mathbf{x} \in \mathbb{R}^E$ be an embedding representing a given track, where E is the number of dimensions in the embedding. Each track is associated with a set of mood tags from a mood vocabulary $\mathcal{T}$ (e.g., "energetic," "gloomy," "happy"), represented by a binary indicator vector $\mathbf{y} \in \{0,1\}^{|\mathcal{T}|}$. We aim at predicting the set of mood tags associated with the track, using a learnable function f that computes the predicted label vector $\hat{\mathbf{y}} = f(\mathbf{x})$.

Note that $\mathbf{x}$ can be extracted from any source of data representing the track. In our case, we will use audio- and listening-based embeddings.

Other approaches have also framed emotion prediction as a regression problem of an n-dimensional continuous space [37], where the 2D Arousal-Valence model [30] is the most widely used. While this approach has the benefit of considering moods that are not constrained by a specific vocabulary, in this work we focus on the multi-label classification approach due to the direct application to potential user-based scenarios such as search by typing or by voice.

2.2 Current Approaches

The current state of the art largely approaches music mood prediction via audio analysis. Early approaches identified spectral contrast as an informative representation [19], and a number of other authors confirmed this finding as well as a variety of other standard audio features [20, 31, 33, 39]. While the relationship between mood and spectral representations remains non-obvious, previous work has shown that human subjects annotate reconstructions from these representations with reasonable consistency to their original form [35]. Still, the problem remained far from solved.

In moving towards increasing model complexity, most approaches have incorporated deep learning methods that seek to learn their own representations [34]. In addition to prediction, audio-based approaches have also been extended to the problem of segmentation [1]. More recent approaches have expanded to multi-modal representations by combining lyrics [8] and others have focused on interpretability of these complex models [6]. At the time of writing, the authors are not aware of any models which leverage features derived from user interactions to estimate the moods of a music track.

3. DATA

The data we collected for this work are derived from various sources: AllMusic provides mood annotations; The Echo Nest Taste Profile [25], mapped to tracks in the Million Song Dataset, adds listening data; finally, 7-digital contributes 30s previews as audio data.

We link AllMusic data to the MSD by fuzzy string matching of artist and track names, and requiring track lengths to be within $\pm$10s. This results in a dataset of 66 993 matched tracks in total, which we call the AllMusic Mood Subset (AMS). As opposed to other music tagging datasets, such as the LastFM Set [4, 15, 17], AMS provides a large vocabulary of mood tags annotated by music experts. While the AllMusic annotations are proprietary, they can be freely consulted on their website and, moreover, are available to be licensed.

Finally, we randomly split the AMS into 80% training, 10% validation, and 10% test, resulting in 53 585, 6695, 6713 tracks respectively. The splits are available online [3] to ensure comparability of future results.

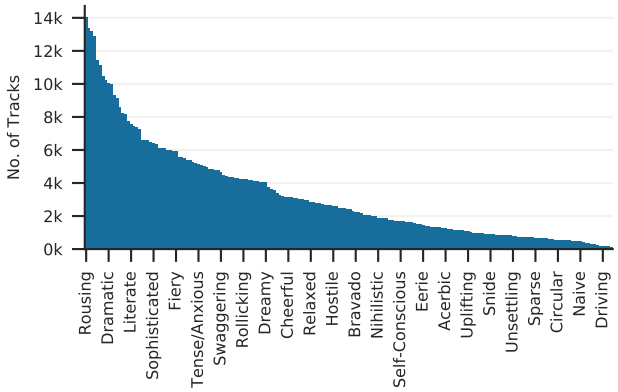


Figure 1: Number of tracks per annotated mood in the AMS. Due to space limitations, only the names of a subset of mood tags are shown.

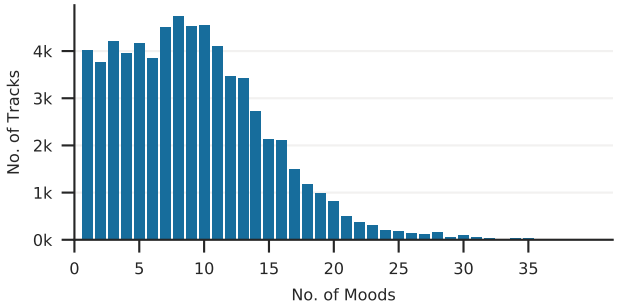


Figure 2: Number of moods annotated per tracks in AMS.

3.1 Mood Data

The mood information that we employ in this work has been human-annotated by experts from AllMusic. These data were previously employed for mood classification [3, 16] and lyrics sentiment detection [24]. The mood tags are annotated at an album level, and we unfold them such that each track is assigned its album-level moods.

The total number of mood tags available is 188. As previous work noted [16], many tags may describe similar

[2] https://www.music-ir.org/mirex/wiki/2019:Audio_Classification_(Train/Test)_Tasks

[3] https://github.com/fdlm/listening-moods

Top	Count	Bottom	Count
Rousing	14 018	Melodic	95
Reflective	13 330	Animated	140
Energetic	13 153	Powerful	148
Earnest	12 873	Driving	163
Passionate	11 438	Introspective	176
Confident	11 092	Flowing	218
Amiable	10 424	Positive	307
Intimate	10 188	Stately	310
Dramatic	10 014	Giddy.	315
Playful	9952	Thoughtful	340

Table 1: 10 top and bottom mood tags based on the number of tracks they have been annotated in the AMS.

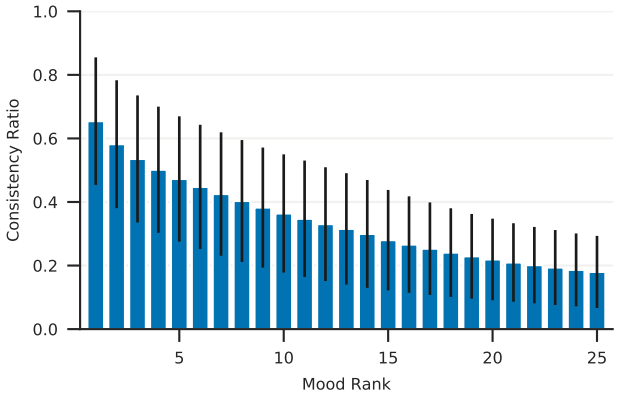


Figure 3: Consistency ratios for the top 25 most popular moods for each user in the AMS.

moods (such as "Romantic" and "Sensual"), which tend to co-occur, and can be clustered into a smaller number of groups. While we can confirm this by performing manual and/or data-driven explorations on the co-occurrence matrix, we intentionally kept the original annotations. For one, we expect modern machine learning methods to cope with large and possibly overlapping vocabularies. For another, these tags were curated by expert annotators to specifically describe how music feels; while they might characterize similar concepts, they could also provide a more nuanced view of a song's mood.

To give a better notion of the moods in this dataset, in Figure 1 we depict the histogram of number of tracks per mood tag, which follows a typical long-tail distribution. The 10 top and bottom annotated mood tags can be seen in Table 1. As we can see, "Rousing" is the most frequent mood, which appears in 14 018 tracks. On the other hand, "Melodic" is the least frequent one, associated with only 95 tracks. On average across the dataset, there are 3258.6 ± 2961.3 tracks for each tag, with a median of 2385. Furthermore, Figure 2 shows the distribution of number of mood tags per track. It can be seen that most tracks have 13 moods or less, with an average of 9.1 ± 5.7 tags per track and the median centered at 9.

3.2 Audio Data

Since the AMS is a subset of the MSD, we gather the audio data by obtaining the 7-digital 30 second previews associated with all MSD tracks. These are 128kbps mp3 stereo files sampled at 44.1kHz.

3.3 Listening Data

We make use of the Taste Profile from the MSD to obtain listening data. These data contain over 28 million play counts from undisclosed partners associated with $L = 1\,019\,318$ listeners and $S = 384\,546$ tracks.

We motivate the usage of such data in the context of mood classification by showing the relationship between listening habits and the moods of the tracks played, thus arguing that such embeddings are likely to contain relevant data when predicting moods. By mapping the tracks in this set with the moods from the AMS (and thus reducing the set of listeners down to 1 012 825, the track set down to 66 993, and the play counts down to $\sim$9 million) we observe that listeners play music that tends to be consistent in terms of its mood. We define the consistency ratio of a mood as the fraction of times it appears in the listening history of a given user. Figure 3 shows the consistency ratio of the n^{th} most popular moods aggregated across users. More specifically, 65.4% of all plays by a given user contain the most popular mood tag for that user; similarly, around 50.1% of a user's plays are annotated with their 4^{th} most popular mood; etc. This exhibits the potential benefits of using listening data, as we confirm in the results of our experiments described next.

4. EXPERIMENTS

As described in Section 2.1, we treat mood prediction as a multi-label classification problem, with a function f predicting mood tags $\hat{\mathbf{y}}$ from an input embedding $\mathbf{x}$. The input embedding can stem from different sources, such as listening- or audio-based features; we will refer to this as the *embedding type*. We will use mostly open models trained on publicly available datasets in this work. As we will see, our conclusions follow from these results alone. Furthermore, we will show results for proprietary models trained on in-house listener feedback data. While we acknowledge that these additional results are hardly reproducible without access to our data and methods, they demonstrate how our findings translate to an industrial scale, and are thus a meaningful addition to this work.

4.1 Evaluation Metrics

Our goal is to compare the predictive performance of each embedding type, *i.e.*, given an input embedding of a certain type, how well the predicted moods $\hat{\mathbf{y}}$ resemble the true moods $\mathbf{y}$ associated with a track. To quantify this, we will use macro-averaged *average precision* as the main evaluation metric, as is commonly used in multi-label classification. Average precision summarizes the precision-recall curve in a single number, and is defined as

$$AP = \sum_n (R_n - R_{n-1}) \cdot P_n, \tag{1}$$

where R_n and P_n are the recall and precision at the n^{th} threshold at which the recall changes. Note that we use *macro averaging*—we first compute AP for each mood tag, and then average them to calculate the final result.

In our mood prediction setup, there are two main questions we need to consider: how do we arrive at the input embedding $\mathbf{x}$, and how we model and train f. Let us first explore the various embedding types, before we take a detailed look at f.

4.2 Audio-Based Models

Current mood prediction systems typically use audio-based features as input. In this work, we use several audio models, pre-trained on different datasets with varying sizes. This ensures that our results are not specific to a type of model.

4.2.1 Musicnn

We employ Musicnn [28]—a spectrogram-based convolutional neural network (CNN) for audio tagging—as the main pre-trained audio-based baseline. It is openly available [4] and achieves state-of-the-art results. We compare two variants of this model: a smaller one, trained on ∼19k tracks from the MagnaTagATune dataset [22], which we will refer to as *MCN-MTT-A*; and a larger one, trained on ∼200k tracks from the Million Song dataset, which we will name *MCN-MSD-A*. Both variants come pre-trained to predict 50 tags, a subset of which can be associated with moods. We refer to Musicnn's documentation for further details on its training scheme.

Musicnn is trained to predict tags for 3-second snippets of audio; however, our setup requires a single *embedding per track*. Thus, instead of the final output, we extract the activation of the penultimate layer of the model as embedding. We first compute embeddings of consecutive non-overlapping audio snippets of 3 seconds, and then average all snippet embeddings to form the track-level embedding. This results in a 200-dimensional vector for MCN-MTT-A, and a 500-dimensional vector for MCN-MSD-A. Such global averaging operations are common for music tagging [27].

4.2.2 Short-Chunk CNN

We train a short-chunk CNN [26] from scratch on the 54k training tracks in the AMS. This simple but powerful model feeds a Mel-spectrogram through a 7-layer CNN with 3×3 filters, 2×2 max-pooling layers, and a fully connected layer before the output. For a detailed look into the training regime and architecture, we refer to the original paper.

Since this model was trained directly for mood prediction on the AMS, there is no need for transfer learning as described in Section 4.4. This is a double-edged sword: although the model is focused on the task at hand, it has to learn a large vocabulary of tags from the limited data provided by our dataset. We will refer to this model as *SCC-A*

[4] https://github.com/jordipons/musicnn

4.3 Listening-Based Models

In contrast to audio-based models, listening-based ones consider user-song interaction as source data. This *listening data* comes in the form of a sparse feedback matrix $Y \in \mathbb{N}^{L \times S}$, where $y_{l,s}$ is a cell in Y representing the number of times the listener l has either played or rated the song s. The former is called *implicit* feedback, while the latter is referred to as *explicit* feedback. Factorizing Y using factorization rank E (corresponding to the desired embedding dimensionality) yields dense track embeddings $\mathbf{x} \in \mathbb{R}^E$: the input to our mood prediction model.

4.3.1 Taste-Profile Factorization

We use listening data from the complete Taste Profile of 28M play counts to obtain song embeddings by applying weighted matrix factorization using alternating least squares [18] with a rank of $E = 200$ (chosen empirically). These data contain relevant information about the track defined exclusively with implicit feedback: how many times which listeners have listened to which songs. We will call these embeddings *TP-L*.

4.3.2 Proprietary Factorization

Large music streaming services possess much larger and more detailed listening data than openly available resources. To see how the results on open datasets translate to industrial settings, we derive 200-dimensional embeddings from more than 100B in-house explicit user ratings over the whole music catalog, by applying a weighted matrix factorization algorithm. These embeddings will be referred to as *P-L*.

4.4 Transfer Learning

Having computed track-level embeddings $\mathbf{x}$ from various sources, we need to map them to mood tags using a learnable function f. This is a transfer-learning scenario: the input embeddings are obtained from a model trained to solve a different (but related) task, such as collaborative filtering or general audio tagging, and then applied for mood prediction by learning f.

We model f as a multi-layer perceptron (MLP) with a binary indicator vector as output, such that $\hat{\mathbf{y}} = f(\mathbf{x})$, where $\hat{\mathbf{y}} \in [0,1]^{|\mathcal{T}|}$. Thresholding $\hat{\mathbf{y}}$ gives us the set of predicted moods. We train f for each embedding type by minimizing the binary cross-entropy between predicted vectors $\hat{\mathbf{y}}$ and target vectors $\mathbf{y}$ obtained from the true mood tags.

The performance of MLPs heavily depends on the choice of hyper-parameters. To enable a fair comparison, we optimized hyper-parameters for each embedding type individually using Bayesian optimization [36], monitoring average precision on the validation set. To limit the computational cost, we only used TP-L and MCN-MSD-A as input embeddings, since they are the main points of comparison. Each setup enjoyed the same, fixed computational budget of 2 days on a single Tesla M40 GPU, which translates to around 200 trials per setup. Table 2 shows details on the search space and the best found configurations. We

	Domain	TP-L	MCN-MSD-A
Nº layers	$[2..4]$	4	4
Nº units	$[1500..4000]$	3909	3933
learning rate	$[0.0001, 0.005]$	4×10^{-4}	5×10^{-4}
dropout [38]	$[0, 0.5]$	0.25	0.25
weight decay	$[0, 0.0001]$	0	1×10^{-6}

Table 2: Hyper-parameters optimized with Bayesian optimization, and best found configurations for each embedding type. Search ranges were defined based on limited initial experiments. For dropout and weight decay, we quantized the interval by 0.125 and 1×10^{-6}, respectively.

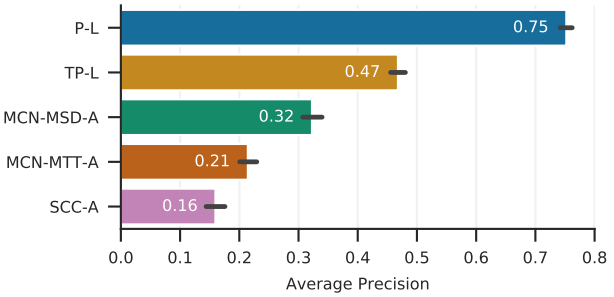


Figure 4: Overall results for each model.

see that for both embedding types, the best models reach the upper limit of our search space, which indicates that even larger models might lead to better results. However, we saw diminishing improvements for large models, so we do not expect much further improvement.

We initialize the MLP weights using Kaiming's method [14], and use a rectifier activation function [13] after each layer (the output layer uses a sigmoid). The input is standardized using mean and standard deviation estimated on the training set. We then train f for 100 epochs using a cosine-annealed learning rate [23] (without restarts) and a 1-epoch warm-up phase. During training, we monitor average precision on the validation set to select the best performing model parameters.

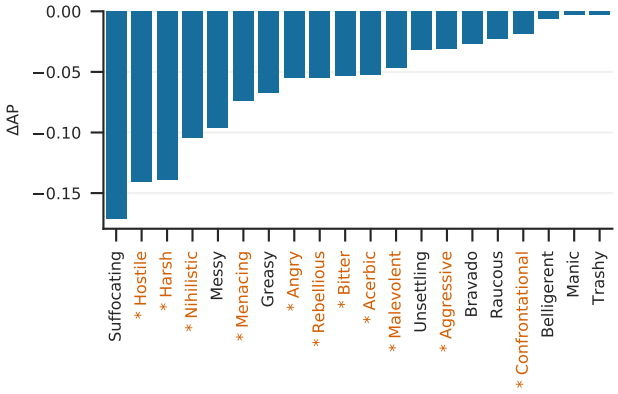
The code to reproduce these experiments is available online. [5]

5. RESULTS

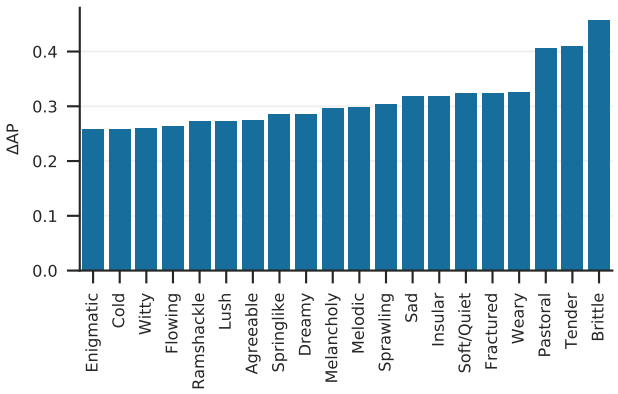
Figure 4 shows the overall results of each embedding type. As mentioned before, our main analysis will be based on the results of open models on publicly available data. We will discuss the results of P-L later.

We see that listening-based embeddings easily outperform audio-based ones (TP-L vs. MCN-MSD-A). We also see a variation within audio-based models. Our experiments were not designed to explain this variation, and the usual suspects offer insufficient clues: for example, dataset size might be an issue (200k for MCN-MSD-A vs. 19k for MCN-MTT-A), but SCC-A was trained on the 54k training tracks from AMS with worse results—here, dataset size relative to vocabulary size might have been the issue. Further experiments, out of scope of this paper, are necessary to understand this in depth.

[5] https://github.com/fdlm/listening-moods



(a) Tags favoring audio-based models.



(b) Tags favoring listening-based models.

Figure 5: Difference of average precision between the best audio-based model (MCN-MSD-A), and the best listening-based model on open data (TP-L). Negative ΔAP means the audio-based embedding performed better. The highlighted tags in (a) belong to the same mood cluster.

5.1 Tag-Wise Results

Even though the overall results are clear, some tags might be easier to predict from audio than from listening data. To explore this, we subtract the tag-wise average precision of TP-L and MCN-MSD-A, and show the results in Figure 5. Indeed, we find 20 tags for which MCN-MSD-A out-performs TP-L. Moreover, these tags seem to describe related moods. To verify this, we clustered the 188 moods using affinity propagation [12], resulting in 13 clusters. We see that 11 out of the 20 mood tags belong to the same cluster, as highlighted in Figure 5a. In contrast, the tags in Figure 5b come from a wider variety of clusters (not highlighted). This indicates that it is a single, coherent "mood subspace" on which audio data is better suited.

5.2 Results by Tag Frequency

As shown earlier, mood tags in the AMS are unevenly distributed: the least popular tag counts only 95 annotations, while the most popular track 14k. It is reasonable to assume that uncommon tags are more difficult to predict than common ones. To evaluate this, we plot the average precision per tag depending on the tag frequency in Figure 6.

Although we see a direct relation between tag frequency and average precision, the extent is less than we expected.

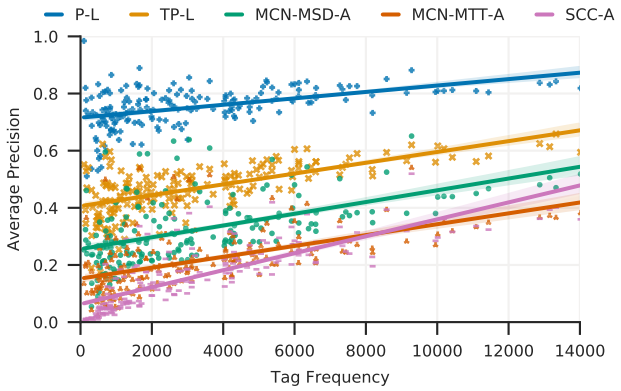


Figure 6: Results per tag frequency. Dots represent the average precision obtained by a tag, that occurs at a frequency shown on the x axis. Lines represent linear regression models, with shades indicating 95% confidence intervals.

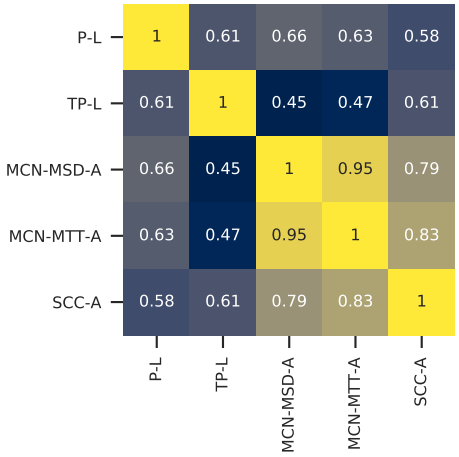


Figure 7: Correlation between tag-wise results of different embeddings. We see that audio-based ones correlate strongly with each other, compared to weaker correlations between listening-based ones.

Furthermore, all embedding types seem to be equally affected: with the exception of SCC-A, the regression slopes of both audio- and listening-based models are notably similar. The exception of SCC-A indicates that tag sparsity may be an issue when training audio models from scratch, but not so when transfer-learning a model that has been trained on more balanced data.

5.3 Results of Proprietary Algorithms

So far, we have discussed the results of open methods on publicly available datasets. However, the attentive reader has noticed that Figure 4 and 6 demonstrate how P-L performs even better than TP-L. To explain the gap between TP-L and P-L, we can point to the different nature and amount of data they were trained on—28M implicit plays for the former, but more than 100B explicit ratings for the latter. The sheer amount of data (a factor of $\sim$3500) and the stronger signal provided by explicit feedback seem to be remarkably beneficial.

5.4 Consistency of Audio-Based Models

We have shown that listening-based models clearly outperform audio-based models in mood prediction. To demonstrate this, we selected a wide variety of audio models that differed in multiple aspects: network architecture, training datasets, and training regime (pre-trained and trained from scratch). Given these differences, we can ask if there are aspects of mood that current audio models are not capable to capture, but listening-based models can. We try to answer this question by exploring which embeddings capture similar mood information. If an embedding captures similar aspects of mood as another embedding, their tag-wise performance should be correlated—but not necessarily similar in magnitude, as one embedding might just perform better than the other.

We show the correlation in tag-wise performance in Figure 7. The remarkable result is that regardless of their differences, the tag-wise results of all audio-based models are much more correlated than between audio- and listening-based embeddings. This indicates that audio-based models do capture similar aspects, even if they might not capture it equally well (as the difference between MCN-MSD-A and SCC-A shows). This does not mean that the aspects current audio-based models are missing are not present in the audio at all—just that current models are not able to extract them.

We do not observe a similar pattern for listening-based embeddings: TP-L and P-L show weaker correlation. At this time, we cannot provide a better explanation than referring to the different nature of explicit and implicit feedback data and the sizes of the two datasets.

6. CONCLUSIONS

In this work we have associated 66 993 tracks from the Million Song Dataset with the AllMusic set to yield the AMS, the largest dataset available with the following data modalities: high quality human mood annotations, audio content, and listening data. Furthermore, we have shown how listening data surpass audio-based embeddings when classifying moods in the proposed dataset. The notable differences in performance between listening- and audio-based models suggest that either i) current state-of-the-art audio models are not capable of successfully extracting certain mood information about a given track; and/or ii) such mood information is not necessarily present in the audio content, and thus the usage of other signals such as listening information may be required to obtain more accurate results. With these findings, we encourage researchers to employ data beyond audio content when estimating the mood of a track. In the future, we look to further scrutinize the tags to better understand which moods might be more suitable to be extracted by which type of input representation. Moreover, and along these lines, we would like to address this task in a multi-modal manner, combining different sources to potentially improve performance of this compelling and timely problem.

7. ACKNOWLEDGEMENTS

The authors F. Korzeniowski and O. Nieto contributed equally to this work.

8. REFERENCES

[1] Anna Aljanaki, Frans Wiering, and Remco C. Veltkamp. Emotion Based Segmentation of Musical Audio. In *Proc. of the 16th Conference of the International Society for Music Information Retrieval (ISMIR)*, Málaga, Spain, October 2015.

[2] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Anssi Klapuri and Colby Leider, editors, *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, USA, October 2011.

[3] Kerstin Bischoff, Claudiu S. Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo. Music Mood and Theme Classification - a Hybrid Approach. In *Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, October 2009.

[4] Erion Çano and Maurizio Morisio. Music Mood Dataset Creation Based on Last FM Tags. In *Proc. of the 4th International Conference on Artificial Intelligence and Applications (AIAP)*, Vienna, Austria, May 2017.

[5] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam. Zero-shot Learning for Audio-based Music Classification and Tagging. In *Proc. of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, November 2019.

[6] Shreyan Chowdhury, Andreu Vall, Verena Haunschmid, and Gerhard Widmer. Towards Explainable Music Emotion Recognition: The Route via Mid-level Features. In *Proc. of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, November 2019.

[7] Paul T. Costa and Robert R. McCrae. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Psychological Assessment Resources, 1992.

[8] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music Mood Detection Based on Audio and Lyrics with Deep Neural Net. In *Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, September 2018.

[9] Shuiguang Deng, Dongjing Wang, Xitong Li, and Guandong Xu. Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications*, 42(23):9284–9293, 2015.

[10] Ignacio Fernández-Tobías, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Iván Cantador. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction*, 26(2-3):221–255, 2016.

[11] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. Personality Traits and Music Genres: What Do People Prefer to Listen To? In *Proc. of the 25th Conference on User Modeling, Adaptation, and Personalization*, Bratislava, Slovakia, July 2017.

[12] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, February 2007.

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In *Proc. of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, USA, June 2011.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 2015.

[15] Xiao Hu and J Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proc. of the 10th ACM International Conference on Digital Libraries (JCDL)*, Surfer's Paradise, Australia, June 2010.

[16] Xiao Hu and J. Stephen Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, September 2007.

[17] Xiao Hu and J. Stephen Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *Proc. of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, August 2010.

[18] Yifan Hu, Chris Volinsky, and Yehuda Koren. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, December 2008.

[19] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Lausanne, Switzerland, August 2002.

[20] Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon G. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull.

Music emotion recognition: A state of the art review. In *Proc. of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, August 2010.

[21] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):42–49, 2009.

[22] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, October 2009.

[23] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations (ICLR)*, Toulon, France, April 2017.

[24] Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. Classification and regression of music lyrics: Emotionally-significant features. In *Proc. of the 8th International Conference on Knowledge Discovery and Information Retrieval (KDIR)*, Porto, Portugal, November 2016.

[25] Brian McFee, Thierry Bertin-Mahieux, Daniel P.W. Ellis, and Gert R.G. Lanckriet. The Million Song Dataset Challenge. In *Proc. of the 21st International Conference on World Wide Web (WWW)*, Lyon, France, April 2012.

[26] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of CNN-based Automatic Music Tagging Models. In *Proc. of the Sound and Music Computing Conference (SMC)*, Torino, Italy, June 2020.

[27] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In *Proc. of the 19th International Society for Music Information Retrieval Conference*, Paris, France, September 2018.

[28] Jordi Pons and Xavier Serra. MusiCNN: Pre-trained convolutional neural networks for music audio tagging. In *Late Breaking of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, November 2019.

[29] Peter J. Rentfrow and Samuel D. Gosling. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236–1256, 2003.

[30] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[31] Markus Schedl, Emilia Gomez, Erika S. Trent, Marko Tkalcic, Hamid Eghbal-Zadeh, and Agustin Martorell. On the interrelation between listener characteristics and the perception of emotions in classical orchestra music. *IEEE Transactions on Affective Computing*, 9(4):507–525, 2018.

[32] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, 2018.

[33] Erik M. Schmidt and Youngmoo E. Kim. Modeling Musical Emotion Dynamics with Conditional Random Fields. In *Proc. of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, USA, October 2011.

[34] Erik M. Schmidt and Youngmoo E. Kim. Learning Rhythm and Melody Features with Deep Belief Networks. In *Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, November 2013.

[35] Erik M. Schmidt, Matthew Prockup, Jeffrey Scott, Brian Dolhansky, Brandon G. Morton, and Youngmoo E. Kim. Relating Perceptual and Feature Space Invariances in Music Emotion Recognition. In *9th Int. Symp. Computer Music Modeling and Retrieval (CMMR)*, London, UK, June 2012.

[36] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, USA, December 2012.

[37] Mohammad Soleymani, Michael N. Caro, Erik M. Schmidt, Cheng Ya Sha, and Yi Hsuan Yang. 1000 Songs for Emotional Analysis of Music. In *Proc. of the 2nd ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM)*, Barcelona, Spain, October 2013.

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[39] Ju-Chiang Wang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng. Modeling the Affective Content of Music with a Gaussian Mixture Model. *IEEE Transactions on Affective Computing*, 6(1):56–68, 2015.

[40] Hiromu Yakura, Tomoyasu Nakano, and Masataka Goto. FocusMusicRecommender: A system for recommending music to listen to while working. In *Proc. of the 23rd International Conference on Intelligent User Interfaces (IUI)*, Tokyo, Japan, March 2018.

[41] E. Zangerle, C. Chen, M. Tsai, and Y. Yang. Leveraging affective hashtags for ranking music recommendations. *IEEE Transactions on Affective Computing (Early-Access)*, 2018.