THE ROLE OF AUDIO AND TAGS IN MUSIC MOOD PREDICTION: A STUDY USING SEMANTIC LAYER PROJECTION

Pasi Saari*, Tuomas Eerola*, György Fazekas†, Mathieu Barthet†, Olivier Lartillot*, Mark Sandler†
*Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland
†Centre for Digital Music, Queen Mary University of London, United Kingdom
*{firstname.lastname}@jyu.fi, †{firstname.lastname}@eecs.qmul.ac.uk

ABSTRACT

Semantic Layer Projection (SLP) is a method for automatically annotating music tracks according to expressed mood based on audio. We evaluate this method by comparing it to a system that infers the mood of a given track using associated tags only. SLP differs from conventional auto-tagging algorithms in that it maps audio features to a low-dimensional semantic layer congruent with the circumplex model of emotion, rather than training a model for each tag separately. We build the semantic layer using two large-scale data sets – crowd-sourced tags from Last.fm, and editorial annotations from the I Like Music (ILM) production music corpus – and use subsets of these corpora to train SLP for mapping audio features to the semantic layer. The performance of the system is assessed in predicting mood ratings on continuous scales in the two data sets mentioned above. The results show that audio is in general more efficient in predicting perceived mood than tags. Furthermore, we analytically demonstrate the benefit of using a combination of semantic tags and audio features in automatic mood annotation.

1. INTRODUCTION

Our daily experiences with music, together with strongly corroborated research evidence [1], suggest that music has a remarkable ability to induce as well as to express emotions or moods. For this reason, the mood associated with a musical piece is often a key aspect in music listening. This provides clear motivations for creating Music Information Retrieval (MIR) systems to organize, navigate or access music collections based on mood. These systems typically rely on mood models and appropriately selected machine learning techniques [2, 3]. Among several models proposed for emotions, the Circumplex model [4, 5] connecting mood terms to underlying emotion dimensions of valence (positive / negative) and arousal (active / passive) is one of the most popular [6]. On the other hand, Thayers variant [7] of this model suggests dimensions of tension and energy diagonal to arousal and valence. However, training machine learning models that automatically associate musical pieces with moods require high quality human mood annotations that are laborious to create, hence typically limited in amount.

Mood-related tags, i.e., free-form labels applied to artists, albums, tracks, etc., are abundantly available from popular online services such as Last.fm [1], while editorial track-level mood tags are vital in large production music catalogues. However, due to issues related to noise and ambiguity in semantic relations between tags, uncovering reliable mood representations from tag data requires typically filtering and semantic analysis [8, 9]. Previous research showed that semantically processed information using track-level Last.fm tags is congruent with listener ratings of valence, arousal, tension and various mood terms [10]. In a test set of 600 popular music tracks, moderate to high ($.47 < r < .65$) correlation was found using the Affective Circumplex Transformation (ACT) technique, that is based on Latent Semantic Analysis (LSA) and the circumplex model of emotions. These results outperformed several conventional semantic analysis techniques, and notably, raw tag frequency scores ($.16 < r < .47$). The robustness of ACT was also demonstrated in [11], by applying the technique to editorial tags from a production music library of about 250,000 tracks.

In a wider context, modelling mood, and thus estimating mood tags may be seen as a specific form of auto-tagging, which is a popular research topic in MIR. A system is typically trained using audio features extracted from a collection of tracks and their associated tags. Then, the trained model is utilised to label new untagged tracks automatically given their features. Typical auto-tagging studies have trained models independently for each tag [12–14], omitting semantic associations between tags, while results in [15] and [16] showed that post-processing auto-tags according to their semantic similarity increases the performance. These techniques have produced promising results for mood tags, possibly due to the use of cleanly-labeled tag data collected for research purposes. As shown in [10], a considerable semantic gap exists between raw crowd-sourced mood tags and verified listener ratings. However, semantic computing provides a promising direction, not yet exploited to the full extent in auto-tagging, for capturing reliable information from large tag collections.

Previous studies in auto-tagging have compared predict-

[1] http://www.last.fm

ed tags with human-labeled tags as a way to assess performance. By considering listener ratings as ground-truth rather than tags, in this paper we analytically compare auto-tags to actual tags as predictors. The two representations relate to two distinct assumptions in tag estimation: either human-labeled tags or audio is available for each new track. Our aim is to challenge these assumptions by highlighting the benefit of semantic computing in the context of music mood auto-tagging. Semantic Layer Projection (SLP) proposed in [17] provides a robust method for projecting the audio feature space to multi-dimensional semantic layer based on ACT. SLP followed by linear regression with the projected feature components outperformed state-of-the-art regression models in predicting listener ratings of valence in 600 tracks from Last.fm. In this paper we evaluate the benefits of SLP in auto-tagging using two corpora: tracks and crowd-sourced mood tags from Last.fm tags as well as tracks and curated editorial mood tags obtained from the I Like Music (ILM) production music catalogue. We predict listener ratings of moods in separate test sets extracted from these corpora.

The rest of the paper is organised as follows: Section 2 describes the tag data and the ACT technique for building the semantic space of moods based on the tags, the set of audio features, and the SLP technique for predicting mood in new tracks based on ACT and the features. Section 3 gives a detailed account of the experimental setup, the data sets used for SLP evaluation, baseline techniques, and the method for comparing mood prediction based on tag or audio information of new tracks. Section 4 shows the results of the experiments and conclusions are drawn in Section 5.

2. METHODOLOGY

2.1 Affective Circumplex Transformation

We used two sources of tracks and tags in the analysis: 259,593 tracks from Last.fm and 226,344 tracks from I Like Music (ILM) production music catalogue, associated with 357 and 288 mood terms, respectively. To create these data sets, tags associated to track sets from the two sources were first lemmatized and identified from a vocabulary of 560 mood terms, aggregated from mood words obtained from selected research papers in affective sciences, music psychology and MIR, as well as from the Allmusic.com web service. In both data sets, tracks with only one tag, and tags associated with less than 100 tracks were then excluded. Finally, the tag data was normalised using term frequency-inverse document frequency (TF-IDF) weights. A detailed account of the data sets and the above process is given in [10, 11].

The following process was applied to Last.fm and ILM sets separately. To uncover semantic similarity between individual mood terms, a low-rank approximation of the TF-IDF matrix was computed using Singular Value Decomposition (SVD) and Multidimensional Scaling (MDS) as in [10]. SVD decomposes a sparse TF-IDF matrix N into orthogonal matrices U and V, and a diagonal matrix S with singular values in decreasing order, such that $N = USV^T$. A rank k approximation of N is then computed by $\bar{N}_k = U^k S^k (V^k)^T$, where each row vector U_i^k represents the terms w_i with k relative weights for each dimension. Similarly, V_j^k represents track t_j as k relative weights. Based on the rank k approximation, dissimilarity between terms w_i and $w_{\hat{i}}$ can be computed using the cosine distance between the $U_i^k S^k$ and $U_{\hat{i}}^k S^k$ vectors. To represent mood terms explicitly in a low-dimensional space that resembles the arousal-valence space, MDS was applied on the term distances to obtain a three-dimensional configuration. The choice of using three dimensions instead of two is motivated by the debate around whether two dimensions is enough to capture relevant variance in moods. Past research have proposed various candidates for the third dimension, such as dominance, potency, or movement.

Next we applied the Affective Circumplex Transformation (ACT) to conform the MDS configuration to the space of *arousal* and *valence* (AV), using AV values of 101 mood terms given in [4, p. 1167] and [5, p. 54]. This technique takes advantage of the Procrustes transformation [18] involving translation, reflection, orthogonal rotation, and isotropic scaling using sum of squared errors as goodness-of-fit. The motivation for this is to *i)* increase the interpretability of the MDS configuration, and *ii)* enable direct prediction of arousal and valence from the semantic space. The technique yields a mood term configuration $x_i = (x_{1,i}, x_{2,i}, x_{3,i}), i = 1, ..., nterms$. A subset of Last.fm and ILM mood term configurations are visualised in Fig. 1 (with $k = 16$). The frequencies of the terms across tracks (co-occurence counts) range from 110 ("vindictive") to 79,524 ("chill") for Last.fm, and 346 ("narrative") to 39,892 ("uplifting") for ILM. Each track j was projected onto the resulting space by taking the Euclidean mean of the term positions, weighted by the sparse TF-IDF vector q_j of the track:

$$t_j = (\Sigma_i q_{j,i} x_i)/(\Sigma_i q_{j,i}). \quad (1)$$

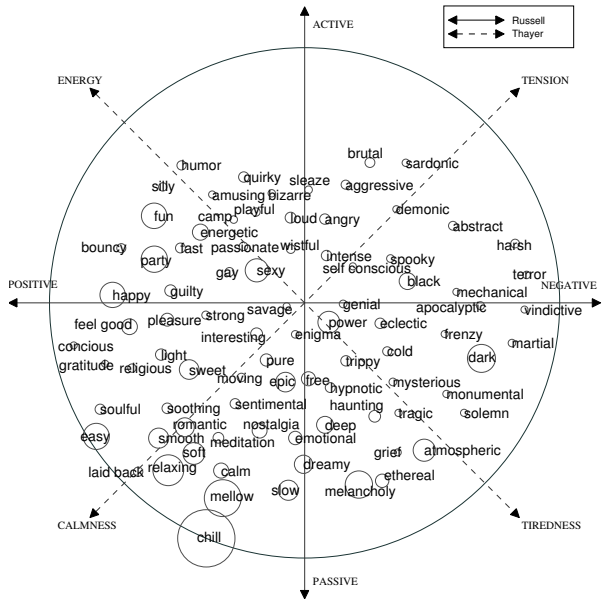
Finally, explicit mood-term-specific weights for the track with position t_j were computed using:

$$P_{j,i} = (x_i/|x_i|) \cdot t_j, \quad (2)$$

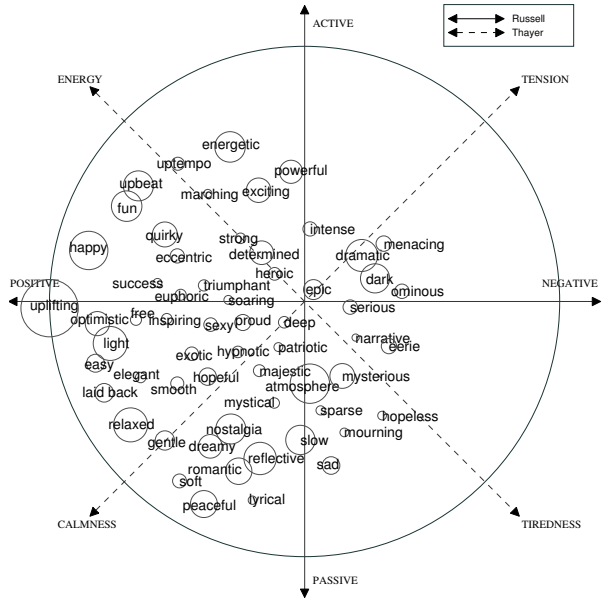
whereas arousal and valence for a track was estimated directly by using the positions along corresponding dimensions. Tension was obtained by projecting tracks along the direction (-1,1,0) as suggested in [7] (see Fig. 1). The analysis in [10] showed that the value of the rank parameter k in SVD computation has minor effect on ACT performance. Therefore we chose to use a heuristically selected value $k = 16$ in our analysis.

2.2 Audio Feature Extraction

Audio features describing dynamics (RMS energy, Low-energy ratio, Attack time, Attack slope), rhythm (Fluctuation pos. & mag., Event density, Pulse clarity, Tempo), pitch (avg. pitch, Chromagram unwrapped centroid), harmony (Key clarity, Mode [majorness], Harmonic change, Roughness), timbre (Brightness, Irregularity, Zerocrossings, Spectral Centroid, Flatness, Skewness, Entropy, Flux



(a) Last.fm.



(b) ILM.

Figure 1. Two first dimensions (valence–arousal) of the three-dimensional mood term configurations obtained with ACT ($k = 16$) for (a) Last.fm and (b) ILM.

and Spread), and structure (Spectral, Rhythmic and Registral repetition) as well as 13 MFCCs, Δ MFCCs, and $\Delta(\Delta)$ MFCCs were extracted from the data sets presented in Table 1 using the MIRtoolbox [19]. To characterise tracks using audio features, statistical means and standard deviations were computed for each feature extracted over short 50% overlapping time frames, yielding a 128 element feature vector for each track. For the features describing the rhythmic repetition and zero crossing rate, we used longer frame lengths of 2s, whereas for chromagram-based features such as the repetition of register, key clarity, centroid, mode, harmonic change, and roughness we used a frame length of 100ms. For other features the frame length was 46.4ms, except for low-energy ratio which is a track-level feature by definition.

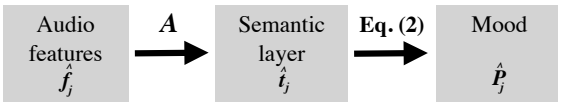


Figure 2. Mapping process in SLP for a novel track represented by audio features $\hat{f}_j$.

	Last.fm		ILM	
	SET10K	SET600	SET5K	SET205
# Tracks	9,662	600	4,692	205
# Terms	357	357	288	288
Term density (%)	1.59	2.43	1.86	2.38

Table 1. Statistics of the mood term sets.

2.3 Semantic Layer Projection

Semantic Layer Projection (SLP), originally proposed in [17], is a technique for the automatic annotation of audio tracks with mood using audio features. SLP is trained on a large collection of audio features and associated tag information. The difference between SLP and conventional auto-tagging is that audio features are not directly mapped to individual tags, but to three-dimensional semantic representation of the mood space obtained by ACT. The mapping is determined by a training stage using the Partial Least Squares (PLS) method. PLS has been found efficient at handling high dimensional and collinear input variables, and it is shown to be robust when using a large number of observations [20].

Given a set F of m audio features related to n tracks $F^{n\times m} = (f_1, f_2, ..., f_n)$, and a semantic layer representation $T^{n\times 3} = (t_1, t_2 ..., t_n)$ for the corresponding tracks (see Eq. 1), the mapping matrix A between F and T is determined using PLS so that $T \approx AF$. We optimize the number of components in PLS by applying (50, 100)-fold cross-indexing [21]. Cross-indexing tackles problems of model overfitting when choosing the optimal parameterisation from several candidates.

The explicit mood of a previously unseen track represented by audio features $\hat{f}_j$ are estimated by first $\hat{t}_j = A\hat{f}_j$, and then by $\hat{P}_{j,i} = (x_i/|x_i|) \cdot \hat{t}_j$ as in Eq. 2. This process is summarised in Fig. 2.

3. EXPERIMENTAL SETUP

3.1 Data Sets

For both data sources, Last.fm and ILM, the semantic models are built from the full set of tracks (250,000 approx.), whereas mappings between audio features and the semantic space in SLP are trained on subsets of these large corpora, namely SET10K and SET5K. The performance of the model is evaluated using listener ratings of perceived moods in separate test sets: SET600 and SET205. Statistical measures of these data sets in terms of semantic mood content are summarised in Table 1.

SET10K consists of 9,662 tracks and was also used in [17]. The set was sampled from the Last.fm corpus in a balanced manner by *i)* optimising mood variance in terms

of track projections in the ACT space, *ii)* favouring tracks with many listeners according to Last.fm, and *iii)* including only unique artists. The audio content of SET10K consists of 15-30s Last.fm preview clips. The clips are typically samples of full tracks in the 128kB/s mp3 format, starting from 30s-60s into the beginning. We assume that these samples are sufficiently representative of the whole tracks. SET600 collected in [10] consists of 15s excerpts of 600 popular music tracks containing no overlapping artists with SET10K, and no tracks overlapping with the large Last.fm corpus. The set was fetched from Last.fm in a similar balanced manner as SET10K, with additional balancing across multiple popular music genres (jazz, pop, rock, electronic, folk and metal), and favouring tracks with many associated mood tags. SET600 was annotated in a listening test [10], with 59 participants rating the excerpts in terms of perceived mood expressed by music. Moods were rated in nine point bipolar Likert-scales for the mood dimensions of valence (negative / positive), arousal (calm / energetic), and tension (relaxed / tense), as well as in unipolar scales for individual mood terms atmospheric, happy, dark, sad, angry, sensual, and sentimental.

The 4,692 tracks from SET5K were picked up randomly from the ILM production music catalogue by *i)* keeping tracks with a duration of at least 60s (in order to discard short instances of the tracks), and *ii)* discarding instrumental stems, i.e. individual tracks from multitrack recordings. Six main genres were represented (jazz, dance, rock, electronic, folk and orchestral). 30s audio clip versions of the tracks were produced in the 128kB/s mp3 format. SET205, described in [11], consists of 205 clips of 30s duration from SET5K. The tracks were sampled in a similar fashion as for the Last.fm test set, but without taking listener statistics into account. The set was annotated by 46 participants in a similar manner as SET600, but for bipolar scales of valence (negative / positive), arousal (calm / energetic), tension (relaxed / tense), dominance (submissive / dominant), romance (cold / romantic), and humour (serious / funny) [11].

Features extracted from SET10K and SET5K were normalised using the z-score transform. All feature values with more than 5 standard deviations from zero were considered outliers and truncated to the extremes $[-5, 5]$. The features associated with SET600 and SET205 were then normalised according to the means and standard deviations of the larger feature sets.

3.2 Modelling Techniques

To show the efficiency of the mappings from audio features to the semantic layer, we compare SLP to two baseline techniques (BL1 and BL2) aiming at predicting mood ratings of e.g. valence, arousal, and tension in the test corpora. Prediction rates are computed as squared correlation coefficients (R^2) between the estimates and ratings over the test sets. The difference between the three techniques lies in how the semantic relationships between mood terms are exploited in the modelling. **BL1** uses mappings between audio features and individual mood terms directly, in order to predict mood ratings for the corresponding terms in the test corpora. This is analogous to the techniques used in [12–14]. **BL2** uses mappings between audio features and individual mood terms to predict each (term-track) pair in the test corpora. Test tracks are then projected using Eq. 1 and Eq. 2 based on the inferred tags. This is analogous to the techniques presented in [15, 16]. The **SLP** technique has been described in Section 2.3.

In short, BL1 does not use information about mood term relationships at all, while BL2 exploits the semantic information after producing a mapping from audio features to mood terms. SLP, on the other hand, maps audio features directly to the semantic layer.

Mappings in BL2 were trained for terms appearing at least ten times in SET10K and SET5K, amounting to 287 and 201 terms, respectively. Since valence, arousal, or tension are not explicitly modeled by BL1 (and no tags "valence" or "arousal" exist in either of the tag corpora), we use terms corresponding to the bipolar labels of the mood scales in the listening tests for modelling these ratings. Tags "positive", "energetic", and "relaxing" / "relaxed" were applied more often than tags "negative", "calm", and "tense" in both SET10K and SET5K, so we use the aforementioned tags to model the corresponding mood dimensions. Similarly, for dominance, romance, and humour that were rated in bipolar scales in SET205, we use tags "powerful", "romantic", and "funny".

Evaluating the role of tags and audio in predicting moods is achieved by comparing SLP and ACT prediction rates. While both of these techniques rely on the same semantic representation of moods, for each novel track, SLP uses only audio features and automatically inferred moods. ACT however uses actual tags associated with the track. We use these techniques in conjunction by computing the weighted mean of these two estimates for each track, and comparing that to the mood ratings. We vary the weights $[w, 1-w]$ ($w \in [0, 1]$) for the techniques so that the case $w = 0$ corresponds to using ACT, whereas the case $w = 1$ corresponds to using SLP.

4. RESULTS AND DISCUSSION

4.1 Evaluation of SLP

Table 2 presents the comparison of SLP with the baseline methods. In case of Last.fm, prediction rates of SLP span from moderate ($R^2 = 0.248$ for happy) to considerably high ($R^2 = 0.710$ for arousal). SLP consistently outperforms both baseline methods, except in one case, where BL1 gives marginally higher performance for sad ($R^2 = 0.313$). The differences between the baseline techniques and SLP are however small for the arousal, angry, and sensual dimensions. We also note that valence and related moods (happy, sad, and angry) are the most difficult to predict with all of the models, and in turn, arousal is the easiest to predict. This is consistent with past studies in music emotion recognition [22]. Although BL1 suffers from the lack of explicit tags for valence, arousal, and tension to infer explicit predictions, results for the seven mood

		BL1	BL2	SLP
Last.fm	Valence	0.045	0.244	**0.322**
	Arousal	0.693	0.662	**0.710**
	Tension	0.198	0.469	**0.560**
	Atmospheric	0.075	0.541	**0.581**
	Happy	0.073	0.183	**0.248**
	Dark	0.264	0.314	**0.370**
	Sad	**0.313**	0.295	0.310
	Angry	0.475	0.465	**0.497**
	Sensual	0.505	0.523	**0.546**
	Sentimental	0.218	0.354	**0.390**
	Mean	0.286	0.405	**0.453**
ILM	Valence	0.156	0.330	**0.486**
	Arousal	0.680	0.672	**0.718**
	Tension	0.478	0.501	**0.588**
	Dominance	**0.461**	0.376	0.352
	Romance	0.274	0.301	**0.351**
	Humour	0.209	0.362	**0.502**
	Mean	.376	.424	**.499**

Table 2. Prediction rates (R^2) for the Last.fm and ILM test sets using SLP and two baseline methods (BL1 and BL2). For each dimension, best scores are reported in bold.

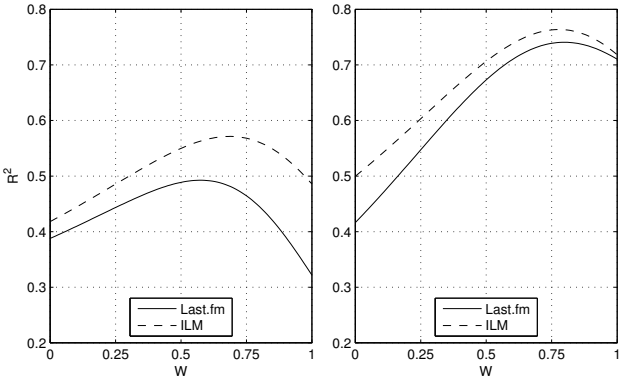
terms show that exploiting semantic associations between tags is highly beneficial. Moreover, as SLP outperforms BL2 for all mood dimensions, mapping tags to the semantic layer directly rather than projecting individual auto-tags to the layer is efficient.

In the case of the ILM data sets, our results show patterns that are highly consistent with those of Last.fm – in general SLP outperforms the baseline methods, while BL1 obtains the lowest performance, on average. However, the performance for valence is considerably higher ($R^2 = 0.486$) than for the Last.fm data set. A clear exception to this pattern is the higher performance of BL1 for dominance ($R^2 = 0.461$) compared to the other techniques. Since dominance is not directly captured either by the tags or the semantic layer dimensions, using other tags than "powerful" would have changed the modelling. In fact, tags "airy", "intimate", and "soft" yielded the highest performance for SLP ($R^2 > 0.57$), the tag "relaxed" yielded the highest performance for BL1 ($R^2 = 0.493$), and the tag "airy" yielded the highest performance for BL2 ($R^2 = 0.543$).

Overall, the results show the advantage of mapping audio features directly to a semantic layer to train predictive models for moods. This solution provides increased performance over methods not exploiting semantic associations at all, or projecting auto-tags to the semantic layer in a later stage, after mapping from audio features to mood tags.

4.2 Tags vs. Audio Features in Mood Prediction

To assess the importance of tags and audio in conjunction, systematic evaluation of using SLP and ACT separately or in conjunction using the weights was carried out. Overall, the results of such comparisons (see Fig. 3 and Table 3) first suggest that the predictions driven by audio features alone yield better performance. However, the combination of audio features and tags lead to a notable increase, especially for moods that are the most difficult for SLP



(a) Valence (b) Arousal

Figure 3. Prediction rate obtained when relying on different information in Last.fm and ILM test sets: tags ($w = 0$), audio ($w = 1$), or combination ($0 < w < 1$).

(valence, happy, and sad). For Last.fm, the mean of the maximum performance when using audio and tags in conjunction is higher ($R^2 = 0.531$) compared to the individual use of tags ($R^2 = 0.334$) and audio ($R^2 = 0.453$). Similar patterns can be observed with the ILM data, for which audio content-based methods outperform tag-based methods for all mood scales (0.062 and 0.164 increases in mean R^2 when both audio and tags are used in conjunction, compared to audio and tags alone, respectively). As can be seen on Fig. 3, the optimal weight for the combination varies to a small degree in both data sets, but lies around 0.70 (mean). In other words, the best prediction of mood is achieved when the acoustic features are attributed a higher weight and are supplemented by tag data, both projected via a semantic space.

However, there are significant exceptions to the conclusions drawn from simply tallying up the prediction rates across the models and data sets. In the Last.fm data set, audio features are actually worse than tags in explaining the ratings of the valence and happy dimensions. This is in line with a number of previous studies in mood prediction with audio features, such as [22], and may have to do with the fact that valence is an elusive concept in music, and maybe particularly dependent on music genres. Further research that extends the mutual patterns between mood and genre is required to untangle such specific results.

5. CONCLUSIONS

In this study, we demonstrated that mood prediction is efficient when relying on large-scale music tag data and audio features, and is boosted by exploiting semantic modelling. The results suggest that higher prediction rates are achievable using the semantic layer projection (SLP) technique when compared to baseline techniques related to conventional auto-tagging that do not incorporate semantic modelling into mappings from audio features.

We conclude that building large-scale predictive models for moods in music can be done more efficiently for certain mood dimensions by relying on audio features rather than

		Tags	max(R^2)	w	Audio
Last.fm	Valence	0.388	0.492	0.57	0.322
	Arousal	0.416	0.741	0.80	0.710
	Tension	0.392	0.618	0.71	0.560
	Atmospheric	0.298	0.607	0.83	0.581
	Happy	0.357	0.429	0.53	0.248
	Dark	0.328	0.506	0.71	0.370
	Sad	0.300	0.393	0.58	0.310
	Angry	0.221	0.518	0.84	0.497
	Sensual	0.371	0.584	0.73	0.546
	Sentimental	0.271	0.422	0.72	0.390
	Mean	0.334	0.531	0.70	0.453
ILM	Valence	0.418	0.571	0.69	0.486
	Arousal	0.500	0.764	0.78	0.718
	Tension	0.497	0.667	0.69	0.588
	Dominance	0.271	0.386	0.73	0.352
	Romance	0.261	0.386	0.75	0.351
	Humour	0.437	0.590	0.67	0.502
	Mean	0.397	0.561	0.72	0.499

Table 3. Prediction rate for the Last.fm and ILM test sets using tags (ACT), audio (SLP), or a weighted combination.

associated tags. This is supported by the higher overall performance of audio compared to tags, and by the overall stable performance of the predictions between the models in two different data sets, crowd-sourced tags from Last.fm and a curated production music corpus (ILM). These data sets consisted of nearly 250,000 tracks each, out of which different subsets were carefully utilized in model training and evaluation. The results also imply that mood tags for novel tracks are not crucial for the automatic annotation of tracks along most mood dimensions. However, for moods related to valence, the use of tags yields a considerable increase in the predictive performance when combined with audio feature-based estimations. In the future we will factor in music genre to the approach presented here.

Acknowledgements This work was partly funded by the Academy of Finland (The Finnish Centre of Excellence in Interdisciplinary Music Research) and the TSB project 12033 - 76187 Making Musical Mood Metadata (TS/J002283/1).

6. REFERENCES

[1] J. A. Sloboda and P. N. Juslin. *Music and Emotion*, chapter Psychological Perspectives on Music and Emotion, pages 71–104. Oxford University Press, New York, 2001.

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303 –319, 2011.

[3] M. Barthet, G. Fazekas, and M. Sandler. Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models. In *Proc. of the 9th Int. Symposium on Computer Music modelling and Retrieval (CMMR)*, pages 492–507, 2012.

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

[5] K. R. Scherer. *Emotion as a multicomponent process: A model and some cross-cultural data*, pages 37–63. CA: Sage, Beverly Hills, 1984.

[6] T. Eerola and J. K. Vuoskoski. A review of music and emotion studies: Approaches, emotion models and stimuli. *Music Perception*, 30(3):307–340, 2012.

[7] R. E. Thayer. *The Biopsychology of Mood and Arousal.* Oxford University Press, New York, USA, 1989.

[8] C. Laurier, M. Sordo, J. Serra, and P. Herrera. Music mood representations from social tags. In *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR)*, pages 381–86, 2009.

[9] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[10] P. Saari and T. Eerola. Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, In press 2013.

[11] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler. Semantic models of mood expressed by music: Comparison between crowd-sourced and curated editorial annotations. In *IEEE International Conference on Multimedia and Expo (ICME 2013): International Workshop on Affective Analysis in Multimedia*, 2013.

[12] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[13] M. I. Mandel and D. PW Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008.

[14] M. I. Mandel and D. P. Ellis. Multiple-instance learning for music information retrieval. In *Proceedings of 9th International Conference of Music Information Retrieval (ISMIR)*, pages 577–582, 2008.

[15] R. Miotto and G. Lanckriet. A generative context model for semantic music annotation and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1096–1108, 2012.

[16] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, 2008.

[17] P. Saari, T. Eerola, G. Fazekas, and M. Sandler. Using semantic layer projection for enhancing music mood prediction with audio features. In *Sound and Music Computing Conference*, 2013.

[18] J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*, volume 3. Oxford University Press Oxford, 2004.

[19] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects*, 2007.

[20] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[21] P. Saari, T. Eerola, and O. Lartillot. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Speech and Audio Processing*, 19(6):1802 –1812, 2011.

[22] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, 2008.