MUSIC MOOD AND THEME CLASSIFICATION - A HYBRID APPROACH

Kerstin Bischoff, Claudiu S. Firan, Raluca Paiu, Wolfgang Nejdl
L3S Research Center,
Appelstr. 4, Hannover, Germany
{bischoff,firan,paiu,nejdl}@L3S.de

Cyril Laurier, Mohamed Sordo
Music Technology Group,
Universitat Pompeu Fabra
cyril.laurier@upf.edu
mohamed.sordo@upf.edu

ABSTRACT

Music perception is highly intertwined with both emotions and context. Not surprisingly, many of the users' information seeking actions aim at retrieving music songs based on these perceptual dimensions – moods and themes, expressing how people feel about music or which situations they associate it with. In order to successfully support music retrieval along these dimensions, powerful methods are needed. Still, most existing approaches aiming at inferring some of the songs' latent characteristics focus on identifying musical genres. In this paper we aim at bridging this gap between users' information needs and indexed music features by developing algorithms for classifying music songs by moods and themes. We extend existing approaches by also considering the songs' thematic dimensions and by using social data from the *Last.fm* music portal, as support for the classification tasks. Our methods exploit both audio features and collaborative user annotations, fusing them to improve overall performance. Evaluation performed against the *AllMusic.com* ground truth shows that both kinds of information are complementary and should be merged for enhanced classification accuracy.

1. INTRODUCTION

General music perception – *i.e.* how we think and talk about music – is heavily influenced by emotions and context. Consequently, users' music information seeking behavior also reflects the importance of opinion/mood and theme associations for music songs. Searching for music usually is an exploratory and social process, in which people make use of collective knowledge, as well as the opinions and recommendations of other people [1]. Related is their need for contextual metadata expressing, for example, which situations/events are often associated with the songs. Thus, besides directly searching or browsing music by artist or title, associated usage, theme/main subject and mood/emotional state are used in every third (navigational) query [1]. Similarly, [2] found that the majority of music queries from a search engine log falls into these categories – 30% of the queries are theme-related (*e.g.* "party music", "wedding songs") and 15% target mood information. Such statistics thus show the necessity of indexing music collections according to mood and theme classes.

Hence, our goal in this paper is to automatically derive mood and theme metadata for music tracks to better cover diverse facets reflecting the complex real-world music information needs of users. With the "mood of a song" we denote the state or the quality of a particular feeling induced by listening to that song (*e.g. aggressive*, *happy*, *sad*, *etc.*). The "theme of a song" refers to the context or situation which fits best when listening to the song, *e.g. at the beach*, *night driving*, *party time*, *etc.*

Currently available state-of-the-art music search engines still do not explicitly support music retrieval based on mood and theme information, and content-based approaches trying to address this problem mainly focus on identifying the moods of songs and do not tackle the thematic aspects of the music resources. Several works in Music Information Retrieval have shown a potential to model the mood from audio content (like [3–6], see [7] for an extensive review). Although this task is quite complex, satisfying results can be achieved if the problem is reduced to simple models [7]. However, an important limitation of these approaches is that they concentrate on the mood only expressed in the audio signal itself and can not capture other sources of emotionality.

Apart from analyzing the low-level features of music resources to identify the songs' corresponding mood or theme, another powerful source of information that can be used are Web 2.0 portals. Collaborative tagging platforms have become extremely popular in recent years – users associate descriptive keywords to various types of content (*e.g.* pictures, Web pages, music). Especially for multimedia data, such as music, the gain provided by the newly available textual information is substantial, since with most prominent search engines on the Web, users are currently still constrained to search for music using textual queries.

The contributions of the paper are twofold:

- We show the feasibility of automatic music classification according to contextual aspects like themes.
- We successfully exploit collective knowledge in form of tags in order to complement the intrinsic information derived from audio features.

The algorithms can be used in various ways: predicted mood and theme labels can be indexed to enrich the metadata index of music search engines enabling a more social and context-aware search (or browsing). Besides, such labels will be valuable for recommendation and playlist generation, *e.g.* for listening to "Party Time"-like songs.

2. RELATED WORK

Music enrichment recently focuses on deriving mood information based on extracted acoustic data [3–5]. [3] proposes a content-based method, tailored to classical music, that uses the Thayer's model [8] for classification. For detecting the mood of music, timbre, intensity and rhythm, features are extracted and a Gaussian Mixture Model is used to model each feature set. In [4], the authors propose a schema such that music databases are indexed on four labels of music mood: "happiness", "sadness", "anger" and "fear". The relative tempo of the music tracks, the mean and standard deviation of average silence ratio are used to classify moods, using a neural network as classifier. For automatically detecting mood for music tracks, [5] uses a set of 12 mood classes which are not mutually exclusive. However, the main focus of the paper is creating a ground truth database for music mood classification.

Several existing papers aim at automatically inferring additional information from available content as well as (user generated) metadata. [9] present a music retrieval system that uses supervised multiclass Naïve Bayes classification for learning a relationship between acoustic features and words from expert reviews on songs, thus enabling query-by-text for music. Similarly, [10, 11] aim at enriching songs with textual descriptions for improving music IR. [10] uses a variant of the AdaBoost algorithm, FilterBoost, in order to predict social tags of the songs based on the information captured in the audio features. Nevertheless, the tags learned by the classifier pertain to multiple categories of tags (genres, styles, moods and contexts) and there is no special focus on mood and theme-related tags, like in our case. [11] compares five methods for collecting tags: user surveys, harvesting social tags, annotation games, mining web documents and auto-tagging audio content. Again, here there is no discussion about the performance of the described methods for predicting mood and theme tags. Moreover, both [10, 11] are not comparable with our approach, since there is no clear definition for mood and theme classes and the data sets on which evaluation was performed differ from ours.

[12] and [13] investigate social tags for improving music recommendations – [12] to attenuate the cold-start problem by automatically predicting additional tags based on the learned relationship between existing tags and acoustic features, [13] to make better recommendations based on the latent factors hidden in user-tag-item relations. For this, the authors successfully apply Higher Order Singular Value Decomposition on the triplets. Again, while both approaches make use of *Last.fm* to predict (the likelihood) of all kinds of tags, our work explicitly focuses on inferring mood and theme annotations.

In [14], *Last.fm* user tags have been used together with content-based features for automatic genre classification. Two classification strategies are proposed that make implicit use of tags: A graph of music tracks is constructed that captures their semantic similarity in terms of tags associated. Both the baseline low-level feature only classifier as well as a single-layer classifier, considering audio features and implicit tag similarity simultaneously, are clearly outperformed by a double-layer classifier, which firsts learns genre labels based on audio information and then iteratively updates its models considering the tag-based neighborhood of tracks.

Thus it seems that especially for multimedia user generated tags are valuable, since low-level features may not be expressive enough. [15] found that *Last.fm* tags define a low-dimensional semantic space which - especially at the track level highly organized by artist and genre - is able to effectively capture sensible attributes as well as music similarity. Somewhat complementary to our approach, [16] aims at studying the relationships between moods and artists, genres and usage metadata. As a test set for the experiments, the authors use *AllMusic.com*, *Epinions.com* and a subset of *Last.fm* data. The authors point out an interesting finding: Many of the individual mood terms were highly synonymous, or described aspects of the same underlying mood space. The experiments also showed that decreasing the mood vocabulary size in some ways clarified the underlying mood of the items being described.

We use *Last.fm*'s valuable folksonomy [1] information for inferring mood and theme labels for songs. While in earlier experiments only tags were used for deriving moods, themes and styles/genres [17], in this paper we also investigate fusion with audio-based methods. Extending existing music metadata enrichment studies, we fuse social tags and low-level audio features of the tracks to infer mood or theme labels showing that both sources provide helpful complementary information.

3. DATA SETS

AllMusic.com. In 1995, the *AllMusic.com* (AMG) website was created as a place and community for music fans. Almost all music genres and styles are covered, ranging from the most commercial/popular to very obscure ones. Not only genres can be found on *AllMusic.com*, but also reviews of albums and artists within the context of their own genres, as well as classifications of songs and albums according to themes, moods or instruments. All these reviews and classifications are manually created by music experts from the *AllMusic.com* team, therefore the data found here serves as a good ground truth corpus.

For our experiments, we collected *AllMusic.com* pages corresponding to music themes and moods, finding 178 different moods and 73 themes. From the pages corresponding to moods and themes, we also gathered information related to which music tracks fall into these categories. This way, we ended up with 5,770 songs. Looking at the

[1] folk + taxonomy: collaboratively created classification scheme

songs identified in each of the categories, we have 8,158 track-mood and 1,218 track-theme assignments. On average songs are annotated with 1.73 moods and 1.21 themes respectively, with maximum number of annotations of 12 and 6 respectively.

Last.fm. For the tracks collected from *AllMusic.com*, we obtained the *Last.fm* tags users had assigned to these songs together with the corresponding frequencies. *Last.fm* is a popular UK-based Internet radio and music community website. In a comparative study on tagging [2] found that the majority of the generally accurate and reliable user tags on *Last.fm* fall into the genre category (60%). Considerably less frequent are tags referring to moods/opinions/qualities (20%) or themes/context/usage (5%) of the music songs. According to [15], at the track level the tags often name the genre and artist of a song. As not all *AllMusic.com* songs have user tags in *Last.fm*, our set of tracks is reduced to 4,737. Using the AudioScrobbler API, we collected in total 59,525 different tags for this set of songs.

Audio. For each track from the previous collections found in our audio database, we have a 30 seconds excerpt in mp3 format with a bitrate of 192 kbps. From these audio tracks, we automatically extracted several state-of-the-art MIR audio features of different type: timbral, tonal, rhythmic including MFCCs, BPM, chroma features, spectral centroid and others. Please refer to [7] for a complete list. For each excerpt of the data set, its 200ms frame-based extracted features were summarized with their component-wise means and variances. At the end of the process, we obtained 240 low-level and mid-level audio features.

4. MOOD AND THEME CLASSIFICATION

For predicting themes and moods, we base our solution on social knowledge – *i.e.* collaboratively created tags associated to music tracks – extracted from *Last.fm*, as well as on audio information. Building upon already provided user tags, on the audio content of music tracks, or on combinations of both, we build multiclass classifiers to infer additional annotations corresponding to moods and themes.

4.1 AllMusic.com Class Clustering

Given that the number of classes existing in *AllMusic.com* is quite large (*e.g.* 178 different moods) with many of the individual terms being highly synonymous or denoting the same concept in well known models of emotions [2] [16], clustering was applied to the initial set of *AllMusic.com* moods as well as the themes.

Mood Clustering. For comparison reasons, we choose the five mood categories used for the MIREX Audio Music Mood Classification Track (see Table 1). Each of the clusters is a collection of five to seven *AllMusic.com* mood labels that together define the cluster. These categories were proposed in [16], derived from a popular set (of Top Songs, Top Albums). The MIREX mood clusters effectively reduce the diverse mood space, and yet root in the social-cultural context of pop music [3]. Restricting our data set to tracks whose assigned moods fall into exactly one of these categories, we had 1192 distinct songs left for machine learning. To balance cluster size for our multiclass classifiers the cutoff was set to 200 instances per cluster.

Nr.	MOOD CLUSTERS – MIREX
MM1	Passionate, Rousing, Confident, Boisterous, Rowdy
MM2	Rollicking, Cheerful, Fun, Sweet, Amiable/Good natured
MM3	Literate, Poignant, Wistful, Bittersweet, Autumnal, Brooding
MM4	Humorous, Silly, Campy, Quirky, Whimsical, Witty, Wry
MM5	Aggressive, Fiery, Tense/Anxious, Intense, Volatile, Visceral

Nr.	MOOD CLUSTERS – THAYER
MT1	**high energy / high stress:** Tense/Anxious, Angst-Ridden, Spooky, Eerie, Rowdy, Fiery, Angry, Fierce, Provocative, Boisterous, Hostile, Aggressive, Volatile, Rebellious, Confrontational, Paranoid, Outrageous, Unsettling, Brittle
MT2	**high energy / low stress:** Rollicking, Exuberant, Happy, Sexy, Exciting, Energetic, Party/Celebratory, Intense, Gleeful, Lively, Cheerful, Fun, Rousing, Freewheeling, Carefree, Passionate, Playful, Gritty, Joyous,
MT3	**low energy / low stress:** Calm/Peaceful, Sentimental, Cathartic, Soft, Romantic, Springlike, Warm, Precious, Laid-Back/Mellow, Confident, Hypnotic, Naive, Intimate, Innocent, Relaxed, Soothing, Dreamy, Smooth, Gentle
MT4	**low energy / high stress:** Sad, Melancholy, Detached, Whimsical, Gloomy, Ironic, Snide, Somber, Autumnal, Wry, Wintry, Plaintive, Yearning, Austere, Bittersweet, Fractured, Bleak, Cynical/Sarcastic, Bitter, Acerbic

Table 1. (Samples from) Mood clusters

Since many *AllMusic.com* mood labels and thus the corresponding songs classified by human experts are not used in MIREX, we as well experimented with the well known two-dimensional models of emotion/mood. In the Thayer energy-stress model [8], emotions are classified along the two axes of (low - high) energy and (low - high) stress. Thus, the two factors divide the mood space into the four clusters "exuberance", "anxious/frantic", "depression" and "contentment". Similarly, Russell/Thayer's bipolar model differentiates emotions based on arousal and valence. In the psychological literature there is little agreement on the number of basic emotional categories or dimensions. However, the Thayer model has been proven useful for music classification and the four categories resulting seem a fair compromise: reducing the mood space to enable clear classificatory distinction and still providing valuable extra-musical metadata for exploratory information needs. During clustering all *AllMusic.com* labels were manually mapped into the two-dimensional mood space by the authors adopting a similarity sorting method as described below for themes. The four resulting clusters are shown together with some example *AllMusic.com* labels in Table 1. Again, clusters were balanced by randomly choosing 403 instances for each cluster.

Theme Clustering. Since *AllMusic.com* themes do not directly correspond to human emotions, mapping the 73 theme terms into the mood spaces used before was not possible (though themes may often be strongly related to specific moods). For manual clustering, we adopted a similarity sorting procedure, in which all *AllMusic.com* themes written on cards were sorted by the authors into as many and as high piles as appropriate. Co-occurrence matrices

[2] Moods are considered to be similar to emotions, but being longer in duration, less intensive and missing object directedness

[3] `http://www.music-ir.org/mirex2007/index.php/Audio_Music_Mood_Classification`

Nr.	THEME CLUSTERS
T1	Party Time, Birthday Party, Celebration, Prom, Late Night, Guys Night Out, Girls Night Out, At the Beach, Drinking, Cool & Cocky, TGIF, Pool Party, Club, Summertime
T2	Sexy, Seduction, Slow Dance, Romantic Evening, In Love New Love, Wedding, Dinner Ambiance
T3	Background Music, Exercise/Workout Music, Playful The Sporting Life, Long Walk, The Great Outdoors, Picnic, Motivation, Empowering, Affirmation, The Creative Side, Victory, Day Driving, Road Trip, At the Office
T4	D-I-V-O-R-C-E, Heartache, Feeling Blue, Breakup, Regret, Loss/Grief, Jealousy, Autumn, Rainy Day, Stay in Bed, Solitude, Reminiscing, Introspection, Reflection, Winter, Sunday Afternoon

Table 2. Theme clusters

were built and added to find good groupings by analyzing the clusters. Unclear membership of singular labels was resolved after discussion. Applying this method resulted in a theme list with 13 labels. Classes containing too few songs are discarded in order to have a minimal representative learning corpus for the classifier, such that the remaining four theme clusters (Table 2) contain 74 songs each.

4.2 Classification

The core of our mood and theme classification methods are multiclass classifiers trained on the *AllMusic.com* ground truth using tags or audio information as features. We experiment both with classifiers created separately for the two different types of features we consider, which are then combined in order to produce for each song a final mood/ theme classification, as well as with a classifier taking as input a combination of audio and tag features. After several experiments, we could observe that SVM classifiers with Radial Basis Function (RBF) kernel performed best for the case of audio input features (it outperformed Logistic Regression, Random Forest, GMM, K-NN and Decision Trees), whereas in the case of tag features, Naïve Bayes Multinomial achieved the best performance. Additionally, the linear combination of the separate classifiers for audio and tag features performed better than the classifier trained on the combination of audio and tag features. Only the best obtained classification results are presented in this paper. We have classifiers trained for the whole set of classes (*i.e.* either for moods or themes) and these classifiers produce for every song in the test set a probability distribution over all classes (*e.g.* over all moods). The highest probability is considered in order to assign the songs to the corresponding class. We experimented with feature selection based on automatic methods (*e.g.* Information Gain) but the results showed that the full set is better suitable for learning, even though it contains some noise.

Algorithm 1 presents the main steps of our classification approach, where classifiers are trained separately for the two different types of input features – tags and audio information. We show the algorithm for mood classification, the case of themes classification being similar.

Step 1 (optional) of the algorithm above aims at reducing the number of mood classes to be predicted for the songs. If two classes are clustered, the resulted class will contain all songs which have been originally assigned to any of the composing classes. As we need a certain amount of input data in order to be able to consistently train the classifiers, we discard those classes containing less than a certain number of songs [4] assigned (step 2).

Alg. 1. Mood classification
Input:ftype – feature type
$ftype = \begin{cases} 0, \text{for tag features;} \\ 1, \text{for audio features.} \end{cases}$
M – mood classes to be learned
S_{total} *– set of songs*

1: *Apply clustering method to cluster moods (see Section 4.1)*
2: *Select classes of moods M to be learned*
For each mood class
If the class does not contain at least *X* songs
Discard class
3: *Classifier learns a model*
3a: Split song set S_{total} into
S_{train} = songs used for training the classifier
S_{test} = songs used for testing the classifiers' learned model
3b: Select features for training the classifier
If $(ftype = 0)$ // tag features
For each song $s_i \in S_{train}$
Create feature vector $F_t(s_i) = \{t_j | t_j \in T\}$, where
T = set of tags from all songs in all mood classes
$t_j = \begin{cases} log(freq(t_j) + 1), \text{if } s_i \text{ has tag } t_j; \\ 0, \text{otherwise.} \end{cases}$
Else // audio features
For each song $s_i \in S_{train}$
Create feature vector $F_a(s_i) = \{a_j | a_j \in A\}$, where
A = set of audio features from all songs in all mood classes
$a_j = standardize(a_j)$
3c: Train and test classifier
If $(ftype = 0)$ // tag features
Train Naïve Bayes (NB) on S_{train} using $\{F_t(s_i); s_i \in S_{train}\}$
Test Naïve Bayes (NB) on S_{test}
Else // audio features
Train SVM on S_{train} using $\{F_a(s_i); s_i \in S_{train}\}$
Test SVM on S_{test}
4: *Classify songs into mood classes*
For each song $s_i \in S_{total}$
If $(ftype = 0)$ // tag features
Compute probability distribution $P_t(s_i)$ as
$P_t(s_i) = \{p_{NB}(m_j|s_i); m_j \in M\}$
Assign s_i to m_j, where $max(p_{NB}(m_j|s_i))$
Else // audio features
Compute probability distribution $P_a(s_i)$ as
$P_a(s_i) = \{p_{SVM}(m_j|s_i); m_j \in M\}$
Assign s_i to m_j, where $max(p_{SVM}(m_j|s_i))$

After selecting separate sets of songs for training and testing in step 3a, we build the feature vectors corresponding to each song in the training set (step 3b). In the case of features based on tags, the vectors have as many elements as the total number of distinct tags assigned to the songs belonging to the mood classes. The elements of a vector will have values depending on the frequency of the tags occurring along with the song. We experimented with different variations for computing the vector elements, but the formula based on the logarithm of the tag frequency provided best results. Audio features are standardized for better suitability with the SVM classifier. Here, a one-vs-one multiclass approach was taken with the parameters selected via grid search (C and gamma with 3-fold cross validation method). Probability estimations are made by pairwise coupling [18].

Once the feature vectors are constructed, they are fed into the classifier and used for training. The assignment of a song to a class is done based on the maximum predicted probability for a song among all possible classes

[4] The threshold depends on the type of clustering and class type. The exact numbers are given in Section 4.1.

Classifier	Class	R	P	F1	Acc
SVM (audio)	Mood MIREX	0.450	0.442	0.420	0.450
NB (tags)	Mood MIREX	0.565	0.566	0.564	0.565
Comb ($\alpha = 0.7$)	Mood MIREX	0.575	0.573	0.572	0.575
SVM (audio)	Mood THAYER	0.517	0.515	0.515	0.517
NB (tags)	Mood THAYER	0.539	0.542	0.539	0.539
Comb ($\alpha = 0.8$)	Mood THAYER	0.570	0.569	0.569	0.569
SVM (audio)	Themes clustered	0.528	0.581	0.522	0.527
NB (tags)	Themes clustered	0.595	0.582	0.575	0.595
Comb ($\alpha = 0.9$)	Themes clustered	0.625	0.617	0.614	0.625

Table 3. Experimental results: P, R, $F1$, Acc for the different classifiers and mood/theme classes

(step 4). As already mentioned, we also experiment with a linear combination of the predictions of the two separately trained classifiers (details are presented in Algorithm 2).

Alg. 2. Mood classification – classifiers' linear combination
Input: M – mood classes to be learned
S_{total} – set of songs

1: For each song $s_i \in S_{total}$
Compute $P_a(s_i) = \{p_{SVM}(m_j|s_i)\} = \{p_a(m_j|s_i)\}$
and $P_t(s_i) = \{p_{NB}(m_j|s_i)\} = \{p_t(m_j|s_i)\}$ (see Alg. 1, step 4)
2: For each α=0.1,...,0.9, $step$=0.1
For each song $s_i \in S_{total}$
For each mood $m_j \in M$
$p_{at}(m_j|s_i) = \alpha \cdot p_a(m_j|s_i) + (1-\alpha) \cdot p_t(m_j|s_i)$
Assign s_i to m_j, where $max(p_{at}(m_j|s_i))$
Compute P, R, Acc, $F1$
3: Select $\alpha = \alpha_{best}$ that produces best results for P, R, Acc, $F1$
4: Classify songs into mood classes, using α_{best} for weighting the probabilities outputted by the audio-based classifier and $(1 - \alpha_{best})$ for weighting the probabilities predicted by the tag-based classifier.

The two different classifiers are first trained to make predictions for all songs in the collection. For producing a linear combination of the classifiers as final output, we then experiment with different values of the α parameter.

5. EVALUATION

For measuring the quality of our theme and mood predictions, we compare our output against the AllMusic experts' assignments, using Precision (P), Recall (R), Accuracy (Acc) and F1-Measure ($F1$) for the evaluation. We present the best results achieved among all our experimental runs (10-fold cross validations) in Table 3. These runs correspond to the different combinations of classifiers (audio-based, tag-based, or linear combinations of the two) and classes to be predicted (themes or moods clustered according to Mirex or Russell/Thayer resp.).

For both moods and themes, we observe that the classifiers relying solely on audio features perform worse than the pure tag-based classifiers. However, combining the two types of classifiers leads to improved overall results. For the moods clustered according to Mirex, Russell/Thayer and themes manually clustered, the best values of α are 0.7, 0.8 and 0.9 respectively. These values indicate a higher weight for the audio-based classifiers, though their achieved performance is poorer than that of the tag-based classifiers. This fact is easily explainable, due to the different types of classifiers considered: SVM for audio features and Naïve Bayes for tag features. It is known that Naïve Bayes produces probabilities close to 1 for the most likely identified class, whereas for the rest of classes, the probabilities are closer to 0. On the other hand, SVM produces more even probability distributions, therefore the high probabilities outputted by Naïve Bayes need to be evened out through a lower α weight. The variations of the $F1$ measure with α are depicted in Figure 1. The biggest variations are to be found in the case of moods clustered according to the Russell/Thayer model, where for values of $alpha$ starting with 0.7 we observe a sharp drop of the $F1$ value. For Mirex mood classes the $F1$ values start to deprecate with α values greater than 0.8.

The baseline accuracy for a random classifier trying to assign songs to the Russell/Thayer mood classes or to the theme clusters is 0.25, while for the Mirex mood classes it would be 0.2. The linear combination of the classifiers improves accuracy in the range of 10 to 27.7% for moods and 18.5% for themes over audio-based classifiers. Overall, results are better for theme classification, indicating that themes are easier to distinguish.

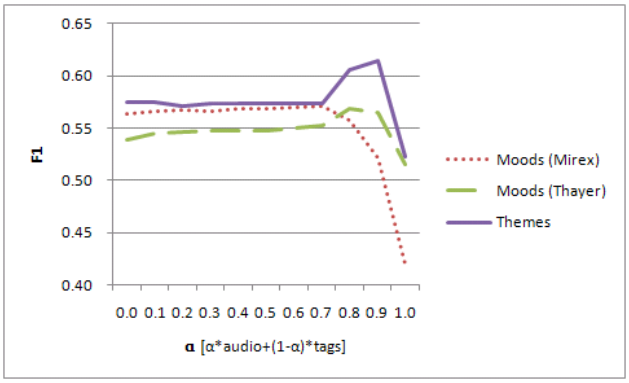

Figure 1. F1 values when varying the α parameter

Analyzing the confusion matrices for the best performing approaches (Figure 2), we observe some prominent confusion patterns: in the case of Mirex clustering, instances belonging to class $MM1$ are often misclassified into $MM2$, $MM4$ instances into $MM3$. Similarly, $MT3$ instances are wrongly classified into $MT4$ for the case of Russell/Thayer clustered moods; also $MT1$ and $MT4$ are often confused. For the latter, the energy dimension does not seem to ease differentiation, given that high stress (negative valence) is characteristic for both classes. $T3$ and $T1$ are the difficult theme classes. Further refinement of these classes should be considered for future work, in order to eliminate this kind of ambiguities (*e.g.* Exercise/Workout music might be as well considered Party-like music).

It is difficult to directly compare our results to the related work cited, as each paper uses a different number of classes. Moreover, experimental goals, ground truth and evaluation procedures vary as well, or detailed descriptions are missing. Comparing to the best algorithms submitted to the MIREX task, we achieve results with lower accuracy. However, knowing that the algorithm used in this paper for audio classification is the same as submitted to MIREX in

A) Moods (Mirex)

Correct Class \ Predicted Class	MM1	MM2	MM3	MM4	MM5
MM1	83	49	14	24	30
MM2	26	118	33	21	2
MM3	6	24	134	27	9
MM4	14	23	38	101	24
MM5	25	6	10	20	139

B) Moods (Thayer)

Correct Class \ Predicted Class	MT1	MT2	MT3	MT4
MT1	269	50	30	54
MT2	57	229	48	69
MT3	40	80	198	85
MT4	61	38	82	222

C) Themes

Correct Class \ Predicted Class	T1	T2	T3	T4
T1	39	8	13	14
T2	3	60	6	5
T3	19	12	30	13
T4	4	9	5	56

Figure 2. Confusion matrix for the best approaches

2007 [19] (obtaining 60.5% accuracy), our conclusion is that the difference comes from the ground truth data. The hypothesis is that our results here are lower because we did not filter the training and test instances using listeners. Moreover for the MIREX collection, listeners were asked to focus on audio only (not lyrics, context or other), which makes it much easier then to classify using audio-based classifiers. In that context, the classification task on our MIREX-like *AllMusic.com* ground truth is more difficult.

6. CONCLUSION

Previous attempts to associate mood labels to music songs often rely on lyrics or audio information for clustering or classifying song corporas. Our algorithms exploit both audio information and social annotations from *Last.fm* for automatically classifying songs according to moods and themes and thus enriching music tracks with information often queried for by users. Themes capturing contextual aspects of songs, in particular, is a facet not considered so far in the literature. The algorithms proposed in this paper rely either on user tags, on audio features, or on combinations of both. Results of an evaluation performed against *AllMusic.com* experts' ground truth indicate that providing such mood and theme information is feasible. The results show that audio and tag information is complementary and should be merged in order to achieve improved overall classification performance. Using our algorithm, music also becomes searchable by associated themes and moods, providing a first step towards effectively searching music by textual, descriptive queries.

For future work, some of the promising ideas to be further investigated refer to refinements of the moods and themes clusters, as well as to other possible combinations of the audio and tag-based classifiers, *i.e.* metaclassifiers.

7. ACKNOWLEDGMENTS

This work was partially supported by the PHAROS project funded by the European Commission under the 6th Framework Programme (IST Contract No. 045035).

8. REFERENCES

[1] J. H. Lee and J. S. Downie: "Survey of music information needs, uses, and seeking behaviours: Preliminary findings," *ISMIR*, 2004.

[2] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu: "Can all tags be used for search?," *CIKM*, pp. 193–202, 2008.

[3] D. Liu, L. Lu, and H.-J. Zhang: "Automatic mood detection from acoustic music data," *ISMIR*, 2003.

[4] Y. Feng, Y. Zhuang, and Y. Pan: "Popular music retrieval by detecting mood," *SIGIR*, 2003.

[5] J. Skowronek, M. McKinney, and S. van de Par: "A demonstrator for automatic music mood estimation," *ISMIR*, 2007.

[6] 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*, Vol. 16, No. 2, pp. 448–457, 2008.

[7] C. Laurier and P. Herrera: "Automatic detection of emotion in music: Interaction with emotionally sensitive machines," *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, pp. 9–32, 2009.

[8] R. E. Thayer: *The biopsychology of mood and arousal*, Oxford University Press, 1989.

[9] D. Turnbull, L. Barrington, and G. Lanckriet: "Modeling music and words using a multi-class nave bayes approach," *ISMIR*, 2006.

[10] 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*, Vol. 37, No. 2, pp. 115–135, 2008.

[11] D. Turnbull, L. Barrington, and G. Lanckriet: "Five approaches to collecting tags for music," *ISMIR*, 2008.

[12] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green: "Automatic generation of social tags for music recommendation," *NIPS*, 2007.

[13] P. Symeonidis, M. Ruxanda, A. Nanopoulos, and Y. Manolopoulos: "Ternary semantic analysis of social tags for personalized music recommendation," *ISMIR*, 2008.

[14] L. Chen, P. Wright, and W. Nejdl: "Improving music genre classification using collaborative tagging data," *WSDM*, pp. 84–93, 2009.

[15] M. Levy and M. Sandler: "A semantic space for music derived from social tags," *ISMIR*, 2007.

[16] X. Hu and J. S. Downie: "Exploring mood metadata: Relationships with genre, artist and usage metadata," *ISMIR*, 2007.

[17] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu: "How do you feel about "dancing queen"?: deriving mood & theme annotations from user tags," *JCDL*, pp. 285–294, 2009.

[18] T.-F. Wu, C.-J. Lin, and R. C. Weng: "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, Vol. 5, pp. 975–1005, 2004.

[19] C. Laurier and P. Herrera: "Audio music mood classification using support vector machine," *MIREX Audio Music Mood Classification contest, ISMIR*, 2007.