This is a preprint of an article accepted for publication in Journal of the Association for Information Science and Technology copyright © 2016 (Association for Information Science and Technology)

A Framework for Evaluating Multimodal Music Mood Classification

Xiao Hu*
Faculty of Education, University of Hong Kong, Hong Kong
Email: xiaoxhu@hku.hk
* Corresponding Author

Kahyun Choi
Graduate School of Library and Information Science, University of Illinois, Champaign, IL, 61820
Email: kahyun2@illinois.edu

J. Stephen Downie
Graduate School of Library and Information Science, University of Illinois, Champaign, IL, 61820
Email: jdownie@illinois.edu

Abstract: This research proposes a framework of music mood classification that utilizes multiple and complementary information sources, namely, music audio, lyric text and social tags associated to music pieces. This article presents the framework and a thorough evaluation on each of its components. Experiment results on a large dataset of 18 mood categories show that combining lyrics and audio significantly outperformed systems using audio-only features. Automatic feature selection techniques were further proved to have reduced feature space. In addition, the examination of learning curves shows that the hybrid systems using lyrics and audio needed fewer training samples and shorter audio clips to achieve the same or better classification accuracies than systems using lyrics or audio singularly. Last but not least, performance comparisons reveal relative importance of audio and lyric features across mood categories.

Introduction

Music is an essential information type in people's everyday life. Nowadays, there have been a large number of music collections, repositories, and websites that strive to provide convenient access to music for various users, from musicians to the general public. These repositories and users often use different types of metadata to describe music such as genre, artist, country of source and music mood[1]. (Vignoli, 2004; Hu & Downie, 2007). Many of these music repositories have been relying on manual supply of music metadata, but the increasing amount of music data calls for tools that can automatically classify music pieces. Music mood classification has thus been attracting researchers' attention in the last decade, but many existing classification systems are solely based on information extracted from the audio recordings of music and have achieved suboptimal performances or reached a "glass ceiling" of performance (Lu, Liu, & Zhang, 2006; Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2008; Hu, Downie, Laurier, Bay, & Ehmann, 2008, Yang & Chen, 2012, Barthet, Fazekas, & Sandler, 2013).

At roughly the same time, studies have reported that lyrics and social tags associated with music have important values in Music Information Retrieval (MIR) research. For example, Cunningham, Downie, and Bainbridge (2005) reported lyrics as the most mentioned feature by respondents in answering why they hated a song. Geleijnse, Schedl, and Knees (2007) proposed an effective method of measuring artists similarity using social tags associated with the artists. As lyrics often bear with semantics of human language, they have been exploited in music classification as well (e.g., He et al, 2008; Hu, Chen, & Yang, 2009; Van Zaanen & Kanters, 2010; Dakshina & Sridhar, 2014). Furthermore, based on the hypothesis

that lyrics and music audio[2] are different enough and thus may complement each other, researchers have started to combine lyrics and audio for improved classification performances (Laurier, Grivolla, & Herrera, 2008; Yang et al., 2008; Björn, Johannes, & Gerhard, 2010; Brilis et al., 2012). Such approach of combining multiple information sources in solving classification problems is commonly called multimodal classification (Kim, et al., 2010; Yang & Chen, 2012; Barthet et al., 2013).

Multimodal classification approaches in general are reported to have improved classification performances over those based on a single source. However, there are many options and decisions involved in a multimodal classification approach, and to date, there has not been any general guidance on how to make these decisions in order to achieve more effective classifications. This study proposes a framework of multimodal music mood classification where research questions on each specific stage or component of the classification process could be answered. This is one of the first studies presenting a comprehensive experiment on a multimodal dataset of 5,296 unique songs, which exemplifies every stage of the framework and evidences how the performances of music mood classification can be improved by a multimodal approach. Specifically, novelty and contributions of this study can be summarized as follows:

1. Conceptualize a framework for the entire process of automatic music mood classification using multiple information sources. The framework is flexible in that each component can be easily extended by adding new methods, algorithms and tools. Under the framework, this study systematically answers questions often involved in multimodal classification: feature extraction, feature selection, ensemble methods, etc.
2. Following a previous study evaluating a wide range of lyric features and their combinations (Hu & Downie, 2010), this study further explores feature selection and effect of dimension reduction of feature spaces. Thus, it pushes forward the state-of-the-art on sentiment analysis for music lyrics;
3. Examine the reduction of training data brought by the multimodal approach. This aspect of improvement has rarely been addressed by previous studies on multimodal music classification. Both the effect on the number of training examples and that on the length of audio clips are evaluated in this study.
4. Compare relative advantages of lyrics and audio across different mood categories. To date, there is little evidence on which information source works better for which mood category(ies). Gaining insight on this question can contribute to deeper understanding of sources and components of music mood.
5. Build a large ground truth dataset for the task of multimodal music mood classification. The dataset contains 5,296 unique songs in 18 mood categories. This is one of the largest experimental datasets in music mood classification with both audio and lyrics available (Kim, et al., 2010; Yang & Chen, 2012; Barthet et al., 2013). Results from a large dataset with realistic and representative mood categories are more generalizable and of higher practical values.

The rest of the paper is organized as follows. Related work is reviewed and research questions are stated. After that, a framework for multimodal music mood classification is proposed. We then report an experiment with ternary information and conclude by discussing the findings and pointing out future work on enriching the proposed framework.

Related Work

Audio-based Music Mood Classification

Based on the assumption that the perceptions of music mood are usually associated with various acoustic cues, most existing work on automated music mood classification builds classification models on features extracted from music audio. The development on this topic has been further stimulated by the Audio

Mood Classification (AMC) in the Music Information Retrieval Evaluation eXchange (MIREX), a community-based annual event for the formal evaluation of algorithms and techniques related to MIR development (Downie, 2008). The AMC task was initiated in 2007, and during the years, it has evaluated more than 200 systems developed by research laboratories around the worlds. The datasets used in existing experiments including the MIREX AMC task usually consisted of several hundred to a thousand songs labeled with four to six mood categories.

A number of acoustic features have been used in automated music classification, representing various aspects of music signals such as energy, rhythm, pitch, and timbre. Among them, timbral features capture characteristics of audio signal spectrum and have been widely used in music mood classification (e.g., Pohle, Pampalk, & Widmer, 2005; Lu et al., 2006; Hu et al, 2008; Trohidis et al., 2008, Yang & Chen, 2012, Barthet et al., 2013). As a supervised learning task, music classification often applies standard supervised learning models such as K-Nearest Neighbor (KNN), Gaussian Mixture Models (GMM), and Support Vector Machines (SVM). Among these models, SVM seems to be the most popular model with top performance (Hu et al., 2008, Kim et al., 2010, Yang & Chen, 2012).

Text-based Music Mood Classification

Music related text information has drawn researchers' attention in recent years. Some studies on music mood classification have been based only on music lyrics (He et al., 2008; Hu et al., 2009b). They showed that higher-order bag-of-words features (i.e., bigrams and trigrams) and linguistic features based on affective lexicons helped capture semantics related to song mood. As in audio-based studies, the datasets used in these studies are usually in smaller scales.

Besides lyrics, Bischoff, Firan, Nejdl, and Paiu (2009a) also tried to use social tags to predict mood and theme labels of popular songs. Their experiment with user evaluation showed that social tags, and the combination of social tags and lyrics were able to predict music in a small number of mood categories. Other studies such as (Saari & Eerola, 2013) exploited mood-related social tags in establishing models of music mood representation and/or predicting listener ratings of moods in music tracks.

Music Mood Classification Combining Audio and Text

The seminal work of Aucouturier and Pachet (2004) pointed out that there appeared to be a "glass ceiling" in audio-based MIR, due to the fact that some high-level music features with semantic meanings might be too difficult to be derived from audio using current technology. With the hope that multiple information sources can compensate for each other, researchers started paying attention to multimodal classification systems that combine audio and text (e.g., Mayer, Neumayer, & Rauber, 2008; Aucouturier, Pachet, Roy, & Beurivé, 2007; Muller, Kurth, Damm, Fremerey, & Clausen, 2007; Björn, Johannes, & Gerhard, 2010) or audio, scores and text (McKay & Fujinaga, 2008). As an earlier attempt, Yang and Lee (2004) used lyric linguistic features to disambiguate categories that audio-based classifiers found confusing. Laurier et al. (2008) combined audio and lyric bag-of-words features and improved classification accuracy on a dataset of 1,000 songs in four categories. Yang et al. (2008) used three fusing methods to combine bag-of-words lyric features and audio features on 1,240 songs and also showed improvement over audio-only classifiers. Moreover, social tags and audio have been combined in music mood classification (e.g., Bischoff et al., 2009b). The experiment results showed that combined classifiers outperformed audio-based ones, suggesting that combining heterogeneous resources helped improve classification performances.

Most of the above studies found audio-based classifiers outperformed lyric-based classifiers, while Bischoff et al. (2009b) found social tag-based classifier outperformed audio-based classifier. Very few studies compared the relative advantages of different information sources across individual mood categories. The study by Schuller and colleagues (2010) was an exception which revealed lyrics were

more helpful on the classification of valence (i.e., songs in positive vs. negative moods) than that of arousal (i.e., songs in relaxing vs. arousing moods).

Most related work on music mood classification used a handful mood categories which were often adapted from classical music psychology models, especially Russell's model of *valence* and *arousal* dimensions (Russell, 1980, Figure 6). It is likely that those mood categories were oversimplified and might not reflect today's reality of the music listening environment (Hu, 2010). Furthermore, the datasets were relatively small, which limited the generalizability of the findings to a variety of music. It is also noteworthy that performances reported in these studies may not be directly comparable as they were evaluated with different datasets.

Research Questions

To fill the aforementioned gaps, this study aims to answer the following research questions within the proposed framework:

1. Are there significant differences between the performances of multimodal systems and single-sourced systems in music mood classification?
   1.1 Which feature selection method works best with the chosen classification model?
   1.2. Which ensemble method is more effective in combining audio and lyrics??
2. Can combining lyrics and audio help reduce the amount of training data needed for effective classification, in terms of the number of training examples and audio length?
3. Are there relative advantages of different information sources across mood categories?

The Framework

The proposed framework of multimodal music mood classification is illustrated in Figure 1. It consists of four major components: dataset construction, feature generation and selection, classification and multimodal combination, evaluation and analysis. This section describes each component in details.

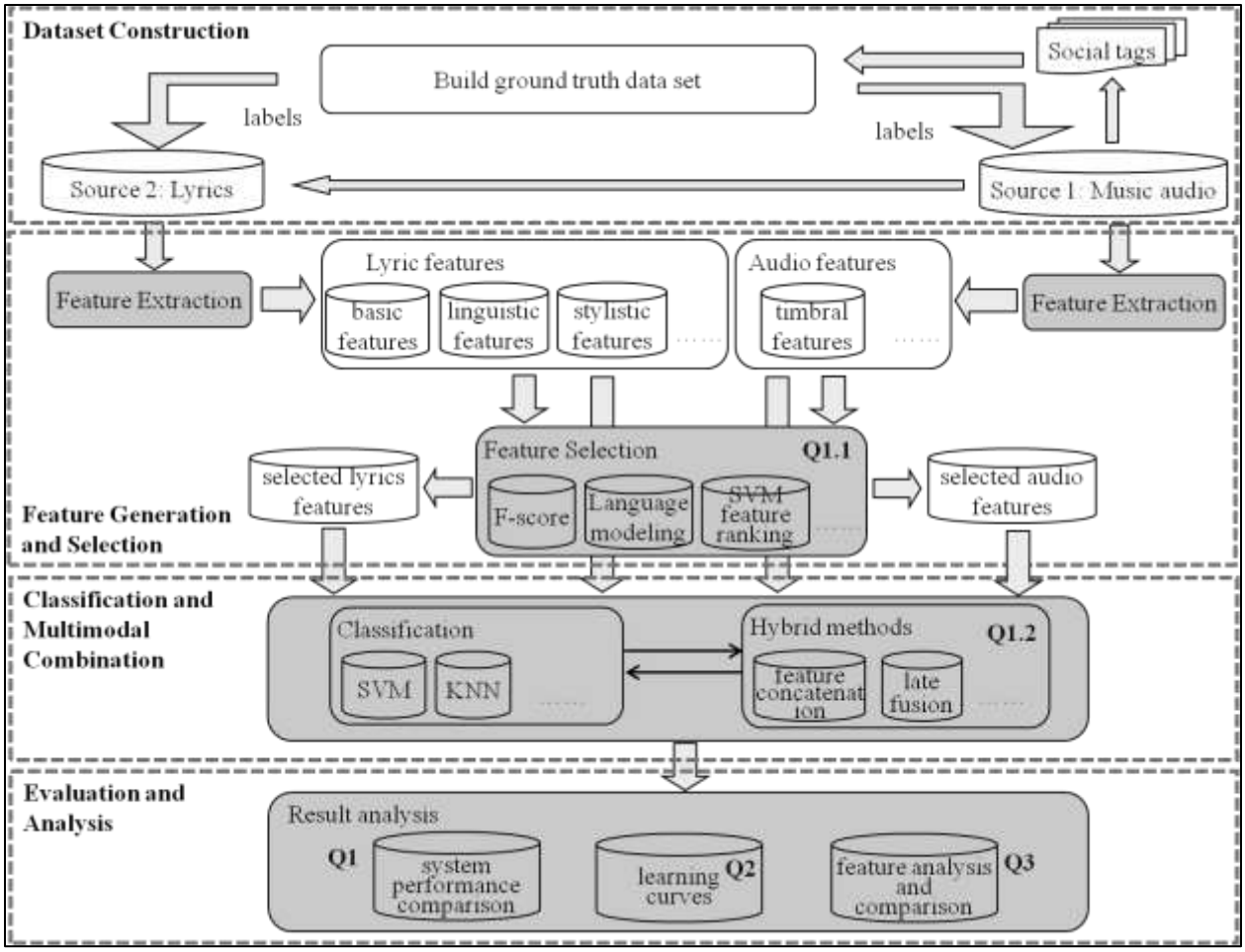

FIG. 1. Proposed framework for multimodal music mood classification (research questions in this study are marked as Q1 to Q3)

Dataset Construction

This component is to collect information sources needed for multimodal classification. As music audio is usually protected by intellectual property laws, the size and types of music in the dataset mainly depend on which music audio files the researchers can gain access to. As an effort to overcome this limitation, some research initiatives have started sharing extracted audio features of large quantity of music with the research community, such as the Million Song Dataset (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011). If the available audio features can satisfy the researchers' needs, then they can also start from these publicly available datasets. Starting from the metadata of the music audio, the associated social tags can be collected from social tagging websites such as last.fm. The social tags can then be used for identifying mood categories and labeling the music pieces with category labels. Cautions have to be paid in deriving groundtruths from social tags, as social tags can be noisy and idiosyncratic (Saari & Eerola, 2014). The method described in Hu et al. (2009a) combined linguistic resources and human inspections to ensure the quality of selected social tags. It also filtered out songs with metadata matching the mood tags. Alternatively, mood labels can be collected from human annotators, but this method is hardly scalable. Besides social tags, lyrics of the music pieces can be obtained from online lyric databases such as lyricwiki.org. Although Figure 1 only shows two information sources (that will be used in the experiment in this study), the framework can include more and diverse sources such as music videos which could then be processed in similar ways in the next components.

Feature Generation and Selection

Automatic classification models are built upon features extracted from information sources. While features extracted from all available information sources are important for multimodal classification, in this study we pay more attention to lyric features as most existing studies have focused on analyzing audio features (Saari, Eerola, & Lartillot, 2011; Song, Dixon, & Pearce, 2012; Baume, Fazekas, Barthet, Marston, & Sandler, 2014) . The various lyric feature types evaluated in this study belong to four major categories:

1. basic text features that are commonly used in text categorization tasks (content words, part-of-speech, function words);
2. linguistic features based on psycholinguistic resources, including i) General Inquirer (GI), a psycholinguistic lexicon mapping English words to psychological categories (Stone, 1966), ii) WordNet (Fellbaum, 1998), iii) WordNet-Affect, an extension of WordNet in the affect domain (Strapparava & Valitutti, 2004), and iv) Affective Norm of English Words (ANEW), a specialized English lexicon mapping common English words to scores in emotion scales (Bradley & Lang, 1999);
3. text stylistic features including interjection words (e.g., "ooh", "ah"), special punctuation (e.g., "!", "?") and text statistics (e.g., number of unique words, length of words, etc.); and
4. the various combinations of two or more of these feature types. For detailed description of these lyric features, please refer to (Hu & Downie, 2010).

Studies in text categorization have evidenced that feature selection can improve generalizability of results and computational efficiency (Yu, 2008). The dimensionalities of most lyric feature types and their combinations can be high, which provides a good opportunity for feature selection. In this study, we compare three methods in selecting the most salient lyric features. The first two are generic measures independent of classification algorithms, whereas the third one is derived from the Support Vector Machines (SVM) algorithm.

1. F-scores. F-score measures the discrimination power of a feature between two data sets (Chen & Lin, 2006). Given a set of training vectors, $x_k, k = 1, \ldots, m$, if the number of positive and negative examples are n_+ and n_-, the F-score of the i-th feature is calculated as:

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (1)$$

   where $\bar{x}_i^{(+)}, \bar{x}_i^{(-)}, \bar{x}_i$ are the average of the *i*-th feature of the positive, negative and the whole training datasets, respectively; $x_{k,i}^{(+)}, x_{k,i}^{(-)}$ are the *i*-th feature of the *k*-th positive and the *k*-th negative example. The higher a feature's F-score is, the more likely it is to be discriminative.
2. Chi-square (χ^2). In Statistics the Chi-square test is to test the independence of two events. In the case of text classification, it is used to test whether the occurrence of a term is independent from the occurrence of a category (Forman, 2003). Therefore the Chi-square function can rank the features by their likelihood of being dependent of class and thus is helpful for classification. Features with higher Chi-square scores are regarded as more helpful for the classification.
3. SVM feature weighting. A trained decision function in a linear SVM contains a weight for each feature which indicates the importance of the feature to the classifier. Researchers in a variety of domains have used SVM as their feature selection method to reduce dimensionality of feature spaces (Guyon, Weston, Barnhill, & Vapnik, 2002; Chapelle, Vapnik, Bousquet, & Mukherjee, 2002; Mladenic, Brank, Grobelnik, & Milic-Frayling, 2004; Yu, 2008). Features with higher absolute weights are more important to the classifier.

Classification and Multimodal Combination

Except for (Hu et al., 2009b), which used fuzzy clustering, most existing studies on music classification used standard supervised learning models such as K-Nearest Neighbors (KNN), Gaussian Mixture Model (GMM) and Support Vector Machines (SVM). Although the proposed framework can include multiple classification models (which can be then compared using a common feature set), many studies have found that SVM are effective in both music mood classification (Hu et al., 2008; Kim et al., 2010; Yang & Chen, 2012; Barthet, Fazekas, & Sandler, 2013) as well as text categorization (e.g., Yu, 2008). Therefore this study will use SVM in the experiment without focusing on the evaluation of classifiers.

As there are multiple information sources, fusion methods are employed to flexibly integrate heterogeneous data sources to improve classification performance. Fusion methods work best when the sources are sufficiently diverse and thus can possibly compensate for each other, as is the case of music audio and lyrics. Two methods have been adopted in music classification: feature concatenation and late fusion. The former concatenates the feature sets and runs the classification algorithms on the combined feature vectors (e.g., Laurier et al., 2008; Mayer et al., 2008). The latter combines the outputs of individual classifiers built on individual information sources, either by (weighted) averaging (e.g., Bischoff et al., 2009b; Whitman & Smaragdis, 2002) or multiplying (e.g., Li & Ogihara, 2004). In the case of combining two classifiers for binary classification as in this study (see below), the two late fusion variations, averaging and multiplying are equivalent (Tax, van Breukelen, Duin, & Kittler, 2000), and the final estimation probability of each testing instance is calculated as:

$$p_{hybrid} = \alpha\, p_{lyrics} + (1-\alpha)\, p_{audio} \qquad (2)$$

where α is the weight given to the posterior probability estimated by the lyric-based classifier, with the range from 0 to 1. A song was classified as positive when the hybrid posterior probability was no less than 0.5.

Evaluation and Analysis

Classification performance can be evaluated on the aforementioned conditions including information sources (lyrics, audio or both), feature sets, feature selection methods, as well as fusion methods. In addition, further analysis can be conducted on the impact of multimodal classification on the amount of training data needed (i.e., so called "learning curves" (Yu, 2008)). In addition, for multimodal classifications, it is particularly interesting to investigate relative advantages of each information source on classifying music pieces across mood categories.

Learning curves. A learning curve describes the relationship between classification performance and the number of training examples. Performance usually increases with the number of training examples, and the point where performance stops increasing indicates the minimum number of training examples needed for achieving the best performance. In addition to classification performances, the learning curve is also an important measure of the effectiveness of a classification system. The comparison on learning curves of the hybrid systems and single-source-based systems can reveal whether combining multiple information sources helps reduce the number of training examples needed for achieving comparable or better performances as single-source-based systems.

The concept of learning curves can be extended to describe the relationship between classification performance and the length of audio pieces in the training data. Due to the time complexity of audio processing, music retrieval systems often process audio clips of *x* seconds truncated from the original tracks instead of the complete tracks, where *x* often equals 30, 15 or 10. As text processing is much faster than audio processing, it is also of practical value to find out whether combining complete lyrics with short audio excerpts can help compensate the (possibly significant) information loss due to the approximation of complete tracks with short clips.

Feature analysis and comparison. Previous studies have mixed findings on whether audio or text were more effective in predicting music mood, or which source was better for certain mood classes (He et al., 2008; Laurier et al., 2008; Bischoff et al., 2009b; Hu et al., 2009a). Even fewer studies examine how multiple sources might interact with each other (McVicar, Freeman, & De Bie, 2011). Under the proposed framework, analyses not only include comparison of classification performances but also that of feature spaces of multiple sources. Classification performances tell us whether a certain experiment setup works, while feature analysis can shed light on why it works.

Experiments

A series of experiments were conducted under the proposed framework to find out answers to our research questions. The experimental setup is described as follows.

Dataset

The dataset used in the experiments contain 5,296 unique Western Pop songs in 18 mood categories. Each of the songs has both music audio and lyrics collected, and could belong to multiple mood categories. The dataset and mood categories were built from an in-house set of audio tracks and the social tags associated with those tracks, using linguistic resources and human expertise (Hu & Downie, 2010). Table 1 reprints the mood categories and the number of positive songs in each category. We adopted a binary classification approach for each of the mood categories and balanced the positive and negative set sizes for each category. As categories can share samples, the total number of samples in all categories is 12,980.

Table 1. Mood categories and number of positive songs.

Category ID	Category	No. of songs	Category ID	Category	No. of songs	Category ID	Category	No. of songs
G1	calm	1,680	G7	Angry	254	G13	anxious	80
G2	sad	1,178	G8	mournful	183	G14	confident	61
G3	glad	749	G9	dreamy	146	G15	hopeful	45
G4	romanti	619	G10	cheerful	142	G16	earnest	40
G5	gleeful	543	G11	brooding	116	G17	cynical	38
G6	gloomy	471	G12	aggressive	115	G18	exciting	30

Audio-based Features and Classifiers

The audio-based system used in this study is a leading audio-based classification system evaluated in the Audio Mood Classification (AMC) task of MIREX: MARSYAS (Tzanetakis & Lemstrom, 2007). MARSYAS has taken part in the AMC task from 2007 to 2012, with consistently top-ranked performances. Using MARSYAS sets a representative and challenging baseline of audio-based classification performance to which multimodal classification is to be compared. MARSYAS used 63 timbre features, including means and variances of Spectral Centroid (indicating the "brightness" of a musical signal), Rolloff (measuring the skewness of the frequencies in a musical signal), Flux (pertinent to the amount of change in sound component), Mel-Frequency Cepstral Coefficients (MFCC) (reflecting the spectral shape of a music signal). Complete audio tracks were used in our experiments unless otherwise specified. All the audio tracks were converted into 44.1kHz stereo .wav files before audio features were extracted.

Evaluation Measures and Classifiers

For each of the experiments, we report the macro average accuracy which gives equal importance to all categories. Within each category, accuracy was calculated with a 10-fold cross validation. Non-parametric Kruskal-Wallis test was applied to compare performances, as the accuracy data may not conform to normal distribution (Downie, 2008). When comparing performances of different systems, the samples used in the tests were accuracies on individual mood categories.

Experiments in this study were implemented using the scikit-learn machine learning tool (Pedregosa et al., 2011). It has been found in the literature that the linear kernel of SVM outperforms other kernels in text

categorization (Aggarwal & Zhai, 2012) due to the redundancy in text data. To verify this, we conducted pilot runs on two randomly selected categories to compare linear kernel to the radial basis function (RBF) kernel with default parameter settings as well as those optimized parameters with grid search. Table 2 shows their performances on categories G2 ("sad") and G16 ("earnest"). The results indicate no significant difference among the three classifiers (p = 0.58 for G, p = 0.06 for G16). As linear kernel achieved higher accuracies and is computationally efficient, it is used with the default parameter throughout the experiments.

Table 2. Accuracies of linear and RBF kernels

	Linear, C=1 (default)	**Linear, optimized C**	**RBF, optimized C and gamma**
G2	60.50%	60.40%	56.83%
G16	65.08%	64.65%	60.83%

In all experiments, optimizations of parameters (i.e., size of feature set, interpolation coefficient in late fusion, etc.) were conducted using an inner 10 fold cross-validation within the training data in each fold of the aforementioned, outer cross-validation.

Results

This section presents experimental results in the order described in the proposed framework.

Best Lyric Features

As reported in (Hu & Downie, 2010), there were seven types of lyric features compared against one another, including 1) content words, 2) Part-of-speech, 3) function words, 4) affective words, 5) psychological categories in General Inquirer (GI), 6) scores derived from the ANEW, and 7) text stylistic features (i.e., interjection words, punctuation marks and text statistics). In addition, the various combinations of the individual feature sets were evaluated and the best performing one was the combination (concatenation) of all the above feature types except for Part-of-speeches (Hu & Downie, 2010). In this study, we will use this combined feature set (denoted as "BEST-all"). Its performance and number of features in each mood category are reported in Table 3.

Best Feature Selection Method

The high dimensionality of the BEST-all feature combination provides room for feature selection and reduction. Using each of the three feature selection methods described above, we selected the top 10% to 90% features from the BEST-all feature set. The features were ranked using an internal cross-validation within the training dataset in each fold, and the results were then averaged across 10 cross validation folds. Table 3 presents the accuracies across mood categories using each of the feature selection methods. The *n%* indicates the percentages of selected features averaged across the 10 folds in the internal cross-validation.

On average, F-score and SVM feature selection methods did not perform as well as the original BEST-all feature set, while the Chi-square method (chi^2) achieved the same average accuracy as the full BEST-all feature set using 65% features on average. A Kruskal-Wallis test using the performance of 18 categories shows that the four systems had no significant difference (H = 1.45, df = 3, p = 0.70), whereas on average 35-44% features were reduced using feature selection. In subsequent experiments, both the full BEST-all feature set and the selected BEST feature set using Chi-square feature selection (denoted as "BEST-chi^2") will be evaluated and compared.

Table 3. Accuracies of feature selection methods across categories.

Mood category	F-score		Chi-square		SVM		BEST-all	
	Accuracy	*n%*	Accuracy	*n%*	Accuracy	*n%*	Accuracy	*N*
calm	0.601	81%	0.604	86%	**0.574***	75%	0.612	11,061

sad	0.659	90%	0.659	89%	0.641	86%	0.669	9,012
glad	0.611	74%	0.613	71%	0.608	73%	0.613	40,897
romantic	0.687	77%	0.687	76%	0.671	74%	0.688	3,661
gleeful	0.582	80%	0.595	79%	0.578	65%	0.604	57,538
gloomy	0.629	65%	0.627	75%	0.662	71%	0.649	11,768
angry	0.684	59%	0.685	71%	0.672	63%	0.706	4,831
mournful	0.694	49%	0.663	70%	0.678	43%	0.685	7,448
dreamy	0.587	51%	0.618	75%	0.644	36%	0.631	12,039
cheerful	0.587	55%	0.588	59%	0.626	47%	0.611	34,067
brooding	0.536	44%	0.547	28%	0.523	70%	0.538	27,145
aggressive	0.780	34%	0.736	61%	0.741	47%	0.749	4,041
anxious	0.625	51%	0.638	71%	0.594	56%	0.606	6,268
confident	0.450	41%	0.540	28%	0.574	21%	0.542	17,545
hopeful	0.578	55%	0.620	55%	0.570	38%	0.630	32,759
earnest	0.667	58%	0.754	69%	0.738	51%	0.750	4,175
cynical	0.600	78%	0.638	69%	0.625	34%	0.650	13,044
exciting	0.533	36%	0.600	38%	0.500	49%	0.500	79,084
Average	**0.616**	**60%**	**0.635**	**65%**	**0.623**	**56%**	**0.635**	**100%**

*: the accuracy is significantly different than that of BEST-all at p <0.05.

Hybrid Systems

The performances of the audio-based classifier are shown in Table 4. Before comparing the two fusion methods, feature concatenation and late fusion, we first need to determine the value of the linear interpolation parameter, α in late fusion (equ. 2). Specifically, we optimized α within the training samples in each fold. The resultant accuracies and corresponding α (averaged across 10 folds) are shown in Table 4. Ties between multiple α values were broken by selecting the largest α, as the lyric-only system outperformed audio-only system on most categories (Table 4). On average, the α value chosen in BEST-chi^2 + Audio (α = 0.67) is larger than that in BEST-all + Audio system (α = 0.33), indicating higher weights to the lyric classifiers were used for BEST-chi^2 + Audio. Table 4 also shows that the late fusion systems significantly outperformed the audio-only systems in "sad", "romantic", "angry" and "hopeful" categories. This can be explained by two possible reasons: 1) the lyrics in these categories contain words semantically representative to the category such as "*kiss*" in "romantic" song and "*fight*" in "angry" songs, and 2) there are few acoustic features known to be related to the mood "hopeful," making it hard to predict "hopeful" songs based on the audio-only classifier.

Table 4. Accuracies of late fusion across categories.

Mood category	Audio-only	BEST-chi^2 + Audio		BEST-all + Audio		Lyric-only (BEST-all)
	Accuracy	Accuracy	$\bar{\alpha}$	Accuracy	$\bar{\alpha}$	Accuracy
calm	0.657	**0.679***	0.43	**0.681***	0.63	0.612
sad	0.676	0.716	0.53	**0.719***†	0.51	0.669
glad	0.590	0.637	0.52	0.640	0.47	0.613
romantic	0.617	**0.699**†	0.57	**0.721**†	0.46	0.688
gleeful	0.620	0.611	0.69	0.633	0.33	0.604
gloomy	0.619	0.642	0.36	0.651	0.71	0.649
angry	0.595	0.692	0.73	**0.723**†	0.23	0.706
mournful	0.630	0.702	0.67	0.699	0.28	0.685

dreamy	0.665	0.666	0.53	0.666	0.44	0.631
cheerful	0.513	0.573	0.85	0.626	0.17	0.611
brooding	0.602	0.574	0.53	0.570	0.44	0.538
aggressive	0.637	0.731	0.83	0.750	0.18	0.749
anxious	0.488	0.600	0.9	0.600	0.13	0.606
confident	0.542	0.582	0.76	0.543	0.15	0.542
hopeful	0.388	**0.640†**	0.87	**0.593†**	0.15	0.630
earnest	0.629	0.679	0.67	0.725	0.32	0.750
cynical	0.575	0.638	0.89	0.650	0.11	0.650
exciting	0.350	0.567	0.78	0.433	0.17	0.500
Average	**0.577**	**0.646**	0.67	**0.646**	0.33	**0.635**

*: the accuracy is significantly higher than that of BEST-all at $p < 0.05$. †: the accuracy is significantly higher than that of Audio-only at $p < 0.05$.

To illustrate the trend of how α values affect the classification performances, we conducted a separate experiment using fixed α values ranging from 0.1 to 0.9 with an increment step of 0.1. The results with different α values are shown in Figure 2. Note these experiments were not aimed to evaluate the effectiveness of the late fusion method, but instead to compare α values which can reflect the relative importance of the audio-based and lyric-based classifiers. In both cases, performances improved quickly in a steady manner when α was increased from 0.1 to 0.6, indicating that even modest involvement of lyric-based classifiers can compensate for the audio-based classifier. The highest average accuracy was achieved when α equaled 0.8 for the hybrid system with the full BEST-all feature set; while the BEST-chi^2 + Audio system performed the best when α was 0.6, giving a small increase in weight to the audio-based classifier.

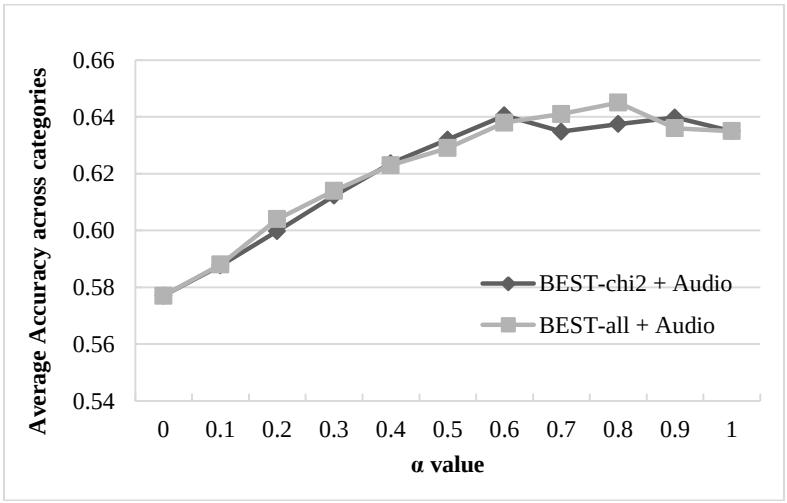

FIG. 2. Effect of α value in late fusion on averaged accuracy.

Table 5 presents the average accuracies of single-source-based systems and hybrid systems with late fusion and feature concatenation. Feature concatenation was not helpful for BEST-chi^2, but was helpful for BEST-all. Late fusion was a good method for both lyric feature sets, improving accuracy over the audio-only system by 7%. In fact, for both lyric feature sets, the hybrid systems using late fusion were significantly better than the audio-only system ($p < 0.05$). For the BEST-all feature set, feature concatenation also outperformed the audio-only system. These again demonstrated the usefulness of lyrics in complementing music audio in the task of mood classification. It is also noteworthy that the

lyric-only systems outperformed the audio-only system by 6%. Previous studies have rarely shown lyric-only systems outperform audio-only systems in terms of averaged accuracy across all mood categories. We surmise that this difference could be attributed to the new lyric features and effective feature selection method applied in this study.

Table 5. Accuracies of single-source and hybrid systems.

Feature set	Audio-only	Lyric-only	Feature concatenation	Late fusion
BEST-chi^2	0.577	0.635	0.613	0.646*
BEST-all	0.577	0.635	0.647*	0.646*

*: The performance is significantly better than that of the audio-only system at $p < 0.05$ level.

Effects on Training Data Size

Number of training examples. In order to find out whether lyrics can help reduce the amount of training examples required for achieving certain performance levels, we examined the learning curves of the single-source-based systems and the late fusion hybrid system for the BEST-chi^2 and BEST-all lyric feature sets. Presented in Figure 3 are the accuracies of the systems when the number of training examples varied from 10% to 100% of all available training samples in each mood category.

Figure 3 shows a general trend that all system performances increased with more training data. It is clear that the performance of the audio-based system increased much more slowly than the other systems. With 20% training examples, the accuracies of the hybrid and the lyric-only systems were already better than that of the audio-only system with any number of available training examples. With 40% (BEST-all) and 70% (BEST-chi^2) training examples, the hybrid systems achieved comparable performances to that of lyric-only system using all training examples. This validates the premise that combining lyrics and audio can reduce the number of training examples needed to achieve the same classification performance levels by single-source-based systems.

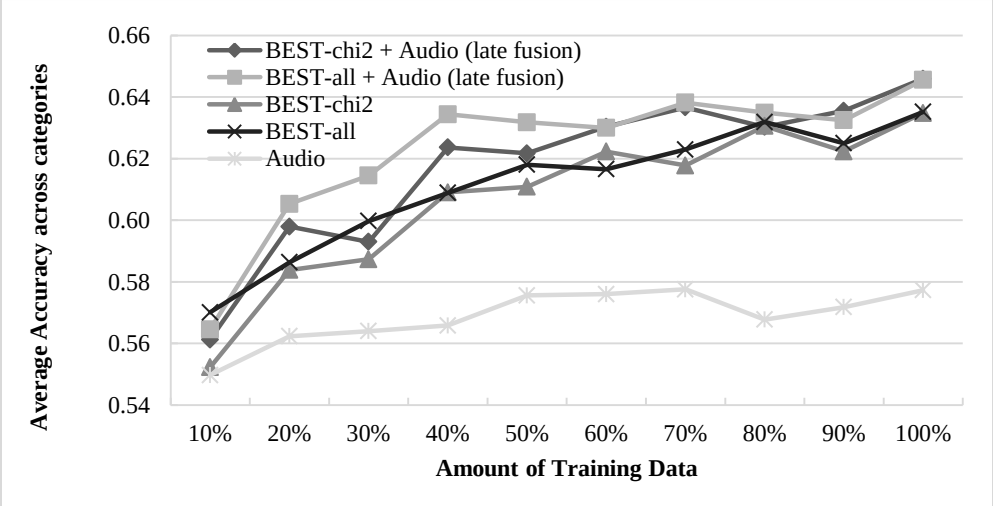

FIG. 3. Learning curves of hybrid and single-source systems.

It is also noteworthy that the BEST-all + Audio system seems to have an advantage over BEST-chi^2 + Audio in reaching better performances with fewer training data samples, although after 60% training data, the two hybrid systems performed very similarly. There seems to be a trade-off between sample size and feature size. In applications where training samples are scarce, it may be favorable to use the entire lyric feature set; whereas when samples are sufficient but processing speed is critical, using feature reduction would be more desirable.

Length of audio clips. This experiment compared the performances of the audio-only and the late fusion hybrid systems on datasets with audio clips of various lengths extracted from the song tracks. In MIR research, most audio clips were extracted from the middle of the songs, as the middle part has been deemed as more representative for the whole song than beginning or ending parts (Silla, Kaestner, & Koerich, 2008). In this experiment, we extracted the audio clips from the middle of the tracks, and set the lengths of the clips to 5, 10, 15, …, 120 seconds as well as the full lengths of the tracks. The hybrid systems also used the complete lyrics throughout the experiment. The results are shown in Figure 4.

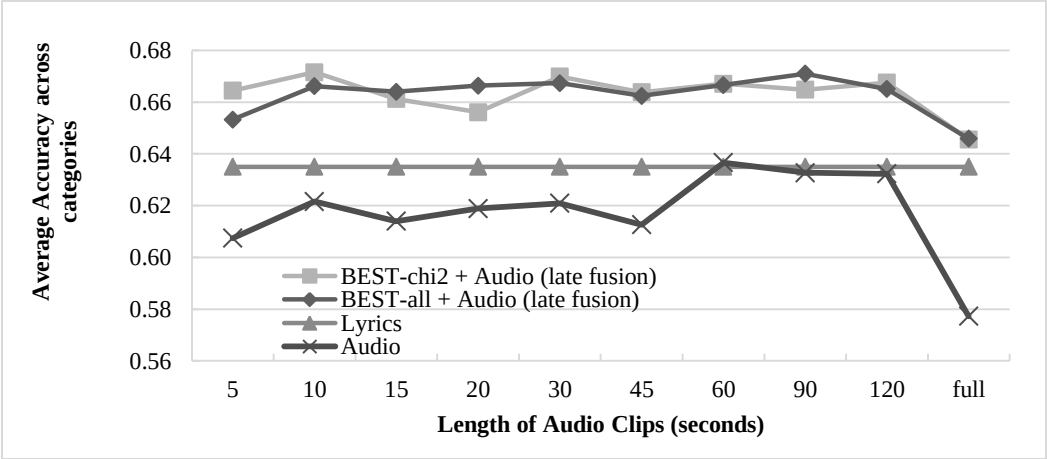

FIG. 4. System performances with varied audio lengths.

The hybrid systems outperformed audio-based systems consistently. With the shortest audio clips (5 seconds), the hybrid systems already outperformed the audio-only system using clips of any length. Therefore, combining lyric and audio can help reduce the length of audio needed and at the same time improve classification performances.

For the two hybrid systems, the performances within each system did not make any significant difference ($p < 0.05$). That is, for the hybrid systems, short audio clips worked as well as long clips. The fact that full lyrics were always used might have at least partially compensated for the shortened audio clips. Also, with the help of lyrics, the performances of hybrid systems did not change as dramatically as that of the audio-only system. This result has an important implication for designing real-time systems: instead of spending precious time processing the audio of an entire song, it is much more efficient to process a very short audio clip sliced from the original song and combine it with the song lyrics.

It is interesting that for all three systems with audio input, the full track did not perform well, which suggests that shorter audio clips would work better with for music mood classification. One possible reason could be that music mood can vary during the time of a song (Yang & Chen, 2012). The beginning and ending parts of a music track may be quite different from the dominant mood of the song, and thus may contribute confusing or distracting information to the classifiers. While how to select parts of song tracks for more accurate predictions is an ongoing research question, our results suggest the sensitivity of audio part selection could be mitigated by combining audio with song lyrics.

Feature Comparison

In this subsection, we examine the relative advantages of lyric and audio features across mood categories. Table 6 lists the categories where performances of the two sources differ significantly. It can be seen that lyrics and audio have their respective advantages in *different* mood categories. Audio timbral features significantly outperformed both lyric feature sets in only one mood category: “calm,” whereas lyric

features achieved significantly better performance than audio in five divergent categories: "romantic", "hopeful", "angry", "aggressive", and "exciting."

Table 6. Comparison of lyric and audio-based classifiers across categories.

Category	Better	Worse	*p*	Category	Better	Worse	*p*
hopeful	BEST-all	Audio	0.005	calm	Audio	BEST-chi^2	0.001
hopeful	BEST-chi^2	Audio	0.009	calm	Audio	BEST-all	0.010
angry	BEST-all	Audio	0.012	romantic	BEST-all	Audio	0.001
angry	BEST-chi^2	Audio	0.045	romantic	BEST-chi^2	Audio	0.001
aggressive	BEST-all	Audio	0.045	exciting	BEST-chi^2	Audio	0.011

To facilitate the comparison and contrast among the categories, we plotted the 18 mood categories in a 2-dimensional space using Multidimensional Scaling (Figure 5). The relative distances between the 18 mood categories were calculated based on the co-occurrence of songs in the positive examples in the dataset. Each mood category is represented by a bubble whose size is proportional to the number of songs in this category. The positions of the mood categories are optimized using classical Procrustes analysis (Saari & Eerola, 2013), with reference to the positions of the six overlapped terms in the well-adopted Russell's model (Russell, 1980; Figure 6). Russell's model consists of two dimensions, *valence* (i.e., level of pleasure) and *arousal* (level of energy). It is the most widely adopted model in music mood recognition (Yang & Chen, 2012; Kim et al., 2010; Barthet et al., 2013) and shares commonality with many other mood models such as equally influential Hevner's adjective circles and the Geneva Emotional Music Scale (GEMS) (Hevner, 1936; Zentner, Grandjean & Scherer, 2008; Gabrielsson & Lindström, 2001; Vuoskoski & Eerola, 2010). It is noteworthy that the distribution of categories in Figure 5 is similar to Russell's model (Figure 6). Categories with lyric features outperforming audio features are scattered in all quadrants but the bottom left one (with negative arousal and negative valence). This seems to suggest lyrics are relatively less helpful for moods with negative valence and negative arousal.

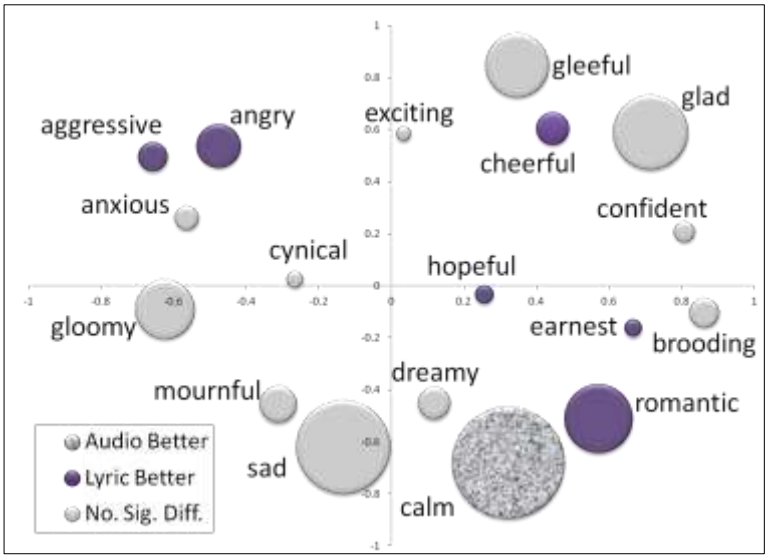

FIG. 5. The 18 mood categories plotted in a 2-dimensional space.

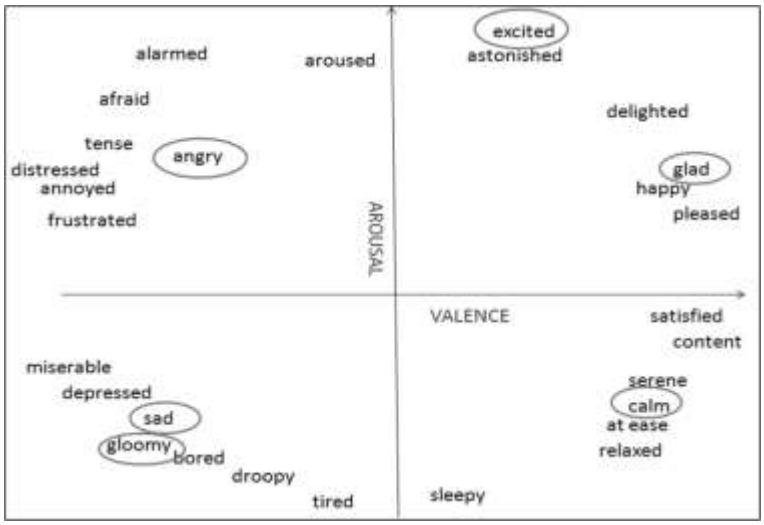

FIG. 6. Russell's 2-dimensional model of music mood (Russell, 1980, p.1168). The terms matching those in FIG. 5 are circled.

Summary of Experiment Results

The experiment results have answered the proposed research questions:

1. Multimodal systems combining lyric and audio features significantly outperformed audio-only classifier. Among the three commonly used feature selection methods, the Chi-square univariate feature selection was the most effective and achieved the same averaged accuracy as the full lyric feature set, using 65% features on average. For ensemble methods, late fusion worked well for both lyric feature sets evaluated while feature concatenation only worked for the BEST-all lyric set.
2. Experiments on learning curves discovered that complementing audio with lyrics could reduce the number of training samples as well as length of audio clips required to achieve the same or better performance than single-source-based systems.
3. Features analysis and comparisons revealed that different information sources have relative advantages in different mood categories with certain valence and arousal configurations.

Conclusions and Future Work

This study proposed a framework for evaluating multimodal music mood classification systems which can flexibly accommodate variations in each of the components. Under this framework, this study systematically evaluated a series of technical options involved in building a multimodal music mood classification system, including a number of novel lyric text feature types, feature selection methods, and fusion methods, all against the same ground truth dataset of a significant scale. In addition, the study also analyzed the effects of multimodal approach on the number of training examples and length of audio clips, as well as the relative advantages of different information sources across different mood categories.

The findings have practical implications on designing and implementing music mood classification and recommendation systems. The proposed framework will help future studies by streamlining system design and evaluation from a holistic point of view. Together, this study contributes to making mood not only a desirable but a practical access point in music repositories. The multimodal framework of combining and compensating multiple information sources could be applied to other domains involving more than one information source, such as multimedia learning resources retrieval using both audiovisual channels and social tags/bookmarks provided by user communities.

Admittedly this study cannot evaluate all variations included in the proposed framework, and thus we plan on extending this work by considering other types of audio features such as rhythmic, harmonic and

psychoacoustic features. Based on the findings of this study, a closer examination on the correlations between multimodal features and mood categories will be conducted to find out why certain sources are more helpful for certain mood categories.

Acknowledgement

This research is partially supported by the Andrew W. Mellon Foundation and a Seed Fund for Basic Research in University of Hong Kong.

Footnotes:

1. In the literature, music mood is also referred as music emotion. Although mood and emotion have different meanings in Psychology, the two terms are often interchangeable in the Music Information Retrieval literature. In this article we do not tell the difference between the two and use "mood" throughout to refer the affect aspect of music information.
2. It is noteworthy that in vocal music, singing of lyrics is recorded in the audio media files as well, but audio engineering technology has yet to be developed to correctly and reliably transcribe lyrics from media files, and thus "audio" at current stage of Music Information Retrieval (MIR) research is regarded as independent of lyrics.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. C. Aggarwal & C. Zhai, C. (Eds.),*Mining Text Data* (pp. 163-222). Springer US.

Aucouturier, J.-J., & Pachet, F. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences , 1* (1), Retrieved from http://ayesha.lti.cs.cmu.edu/ojs/index.php/jnrsas/article/viewFile/2/2.

Aucouturier, J.-J., Pachet, F., Roy, P., & Beurivé, A. (2007). Signal + Context = Better Classification. *In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, (pp. 425-430). Vienna, Austria.

Barthet, M., Fazekas, G., & Sandler, M. (2013). Music Emotion Recognition: From Content-to Context-Based Models. In M. Aramaki, M. Barthet, R. Kronland-Martinet, & S. Ystad, *From Sounds to Music and Emotions* (pp. 228-252). Springer Berlin Heidelberg.

Baume, C., Fazekas, G., Barthet, M., Marston, D., & Sandler, M. (2014). Baume, C., Fazekas, G., Barthet, M., Marston, D., & Sandler, M. (2014, January). Selection of Audio Features for Music Emotion Recognition Using Production Music. *53rd International Conference: Semantic Audio.*

Bertin-Mahieux, T., Ellis, D., Whitman, B., & Lamere, P. (2011). The million song dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 591-596). Miami, Florida: University of Miami.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2009). How do you feel about "Dancing Queen"? Deriving mood and theme annotations from user tags. *In Proceedings of Joint Conference on Digital Libraries (JCDL)* (pp. 285-294). Austin, TX: New York: ACM Press.

Bischoff, K., Firan, C., Paiu, R., Nejdl, W., Laurier, C., & Sordo, M. (2009). Music mood and theme classification - a hybrid approach. *In Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, (pp. 285-294). Kobe, Japan.

Björn, S., Johannes, D., & Gerhard, R. (2010). Determination of nonprototypical valence and arousal in popular music: features and performances. *Journal on Audio, Speech, and Music Processing* .

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings.* Technical report C-1. University of Florida.

Brilis, S., Gkatzou, E., Koursoumis, A., Talvis, K., Kermanidis, K. L., & Karydis, I. (2012). Mood classification using lyrics and audio: A case-study in greek music. In L. Iliadis, I. Maglogiannis, H. Papadopoulos, K. Karatzas, & S. Sioutas, *Artificial Intelligence Applications and Innovations* (pp. 421-430). Springer Berlin Heidelberg.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple param- eters for support vector machines. *Machine Learning , 46*, 131–159.

Chen, Y.-W., & Lin, C.-J. (2006). Combining SVMs with various feature selection strategies. In I. Guyon, S. Gunn, M. Nikravesh, & L. Zadeh (Eds.), *In Feature Extraction, Foundations and Applications* (pp. 315-324). Springer.

Cunningham, S. J., Downie, J. S., & Bainbridge, D. (2005). "The Pain, The Pain": Modeling music information behavior and the songs we hate. *In Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR).*, (pp. 474-477). London, UK.

Dakshina, K., & Sridhar, R. (2014). LDA Based Emotion Recognition from Lyrics. In M. Kumar Kundu, D. P. Mohapatra, A. Konar, & A. Chakraborty, *Advanced Computing, Networking and Informatics* (pp. 187-194). Springer International Publishing.

Downie, J. S. (2008). The Music Information Retrieval Evaluation eXchange (2005-2007): a window into music information retrieval research. *Acoustical Science and Technology , 29* (4), 247-255.

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

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge: The MIT Press.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, *3*, 1289-1305.

Gabrielsson, A., & Lindström, E. (2001). The influence of musical structure on emotional expression. In P. N. Juslin and J. A. Sloboda (Eds.), *Music and Emotion*: *Theory and Research*. New York: Oxford University Press, pp. 223-248.

Geleijnse, G., Schedl, M., & Knees, P. (2007). The quest for ground truth in musical artist tagging in the social web era. *In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, (pp. 525-530). Vienna, Austria.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning , 46* (1-3), 389-422.

He, H., Jin, J., Xiong, Y., Chen, B., Sun, W., & Zhao, L. (2008). Language feature mining for music emotion classification via supervised learning from lyrics. *Advances in Computation and Intelligence, Lecture Notes in Computer Science (LNCS) , 5370*, 426-435.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 246-268.

Hu, X. (2010). Music and Mood: Where Theory and Reality Meet. *iConference*, (pp. 1-8). Champaign, IL.

Hu, X., & Downie, J. S. (2007). Exploring mood metadata: relationships with genre, artist and usage metadata. *In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, (pp. 462-467). Vienna, Austria.

Hu, X., & Downie, J. S. (2010). Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. *Proceedings of the Joint Conference on Digital Libraries* (pp. 159-168). Surfers Paradise, Australia: ACM New York, NY, USA.

Hu, X., & Downie, J. S. (2010). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. *Proceedings of the 11th International Conference on Music Information (ISMIR)*, (pp. 619 - 624). Utrecht, Netherlands.

Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. *In Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, (pp. 411–416). Kobe, Japan.

Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 MIREX Audio Music Classification task: lessons learned. *In Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, (pp. 462-467). Philadelphia, PA.

Hu, Y., Chen, X., & Yang, D. (2009). Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. . *In Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, (pp. 123-128). Kobe, Japan.

Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., et al. (2010). Music emotion recognition: a state of the art review. *In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 255-266). Utrecht, Netherland.

Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. *In Proceedings of the 7th International Conference on Machine Learning and Applications (ICMLA)* (pp. 688-693). San Diego, CA: IEEE Computer Society.

Li, T., & Ogihara, M. (2004). Music artist style identification by semi-supervised learning from both lyrics and content. *Proceedings of the 12th annual ACM international conference on Multimedia. 8*, pp. 364 - 367. New York, USA: ACM New York, NY, USA.

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. *In Proceedings of the 8th International Conference on Intelligent User Interfaces* (pp. 125-132). Miami, FL: ACM New York.

Lu, L., Liu, D., & Zhang, H. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing , 14* (1), 5-18.

Mayer, R., Neumayer, R., & Rauber, A. (2008). Rhyme and style features for musical genre categorisation by song lyrics. *In Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, (pp. 337-342). Philadelphia, PA.

McKay, C., & Fujinaga, I. (2008). Combining features extracted from audio, symbolic and cultural sources. *In Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR)*, (pp. 597-602). Philadelphia, PA.

McVicar, M., Freeman, T., & De Bie, T. (2011). Mining the Correlation between Lyrical and Audio Features and the Emergence of Mood. *Proceedings of the12th Conference of International Society for Music Information Retrieval*, (pp. 783-788).

Mladenic, D., Brank, J., Grobelnik, M., & Milic-Frayling, N. (2004). Feature selection using linear classifier weights: Interaction with classification models. *In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 234–241). Sheffield, UK: ACM Press.

Muller, M., Kurth, F., Damm, D., Fremerey, C., & Clausen, M. (2007). Lyrics-based audio retrieval and multimodal navigation in music collections. *In Proceedings of the 11th European Conference on Digital Libraries (ECDL)* (pp. 112-123). Budapest, Hungary: Springer Verlag.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval. , 2* (1-2), 1-135.

Pedregosa et al. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830.

Pohle, T., Pampalk, E., & Widmer, G. (2005). *Evaluation of frequently used audio features for classification of music into perceptual categories.* Technical Report, Österreichisches Forschungsinstitut für Artificial Intelligence, Wien, Austria.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities , 31*, 351–365.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology , 39*, 1161-1178.

Saari, P., & Eerola, T. (2013). Semantic Computing of Moods Based on Tags in Social Media of Music. *IEEE Transactions on Knowledge and Data Engineering* , 1-14.

Saari, P., Eerola, T., & Lartillot, O. (2011). Generalizability and simplicity as criteria in feature selection: application to mood classification in music. *Audio, Speech, and Language Processing, IEEE Transactions on , 19* (6), 1802-1812.

Schuller, B., Hage, C., Schuller, D., & Rigoll, G. (2010). Mister D.J., cheer me up!: Musical and textual features for automatic mood classification. *Journal of New Music Research , 39* (1), 13-34.

Silla, C. N., Kaestner, C. A., & Koerich, A. L. (2008). The Latin Music Database. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, (pp. 451-456).

Song, Y., Dixon, S., & Pearce, M. (2012). Evaluation of Musical Features for Emotion Classification. *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)*, (pp. 523-528). Porto, Portugal.

Stone, P. J. (1966). *General Inquirer: a computer approach to content analysis.* Cambridge, MA: The MIT Press.

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. *In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 1083-1086). Lisbon, Portugal: European Language Resources Association.

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems, Special Issue, , 9*, 483-496.

Tax, D. M., van Breukelen, M., Duin, R. P., & Kittler, J. (2000). Combining multiple classifiers by averaging or by multiplying. *Pattern Recognition , 33*, 1475-1485.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich Part-of-Speech tagging with a cyclic dependency network. *In Proceedings of HLT-NAACL*, (pp. 252 - 259). Edmonton, Canada.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multi-label classification of music into emotions. *In Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, (pp. 325-330). Philadelphia, PA.

Tzanetakis, G. (2007). Marsyas submissions to MIREX 2007. *MIREX 2007 Extended Abstract* , Retrieved April 20, 2010 from http://www.music-ir.org/mirex/abstracts/2007/AI_CC_GC_MC_AS_tzanetakis.pdf.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing , 10* (5), 293-302.

Tzanetakis, G., & Lemstrom, K. (2007). Marsyas-0.2: a case study in implementing music information retrieval systems. *Intelligent Music Information Systems* .

Van Zaanen, M., & Kanters, P. (2010). Automatic Mood Classification Using TF* IDF Based on Lyrics. *In Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, (pp. 75-80). Utrecht, Netherlands.

Vignoli, F. (2004). Digital Music Interaction concepts: a user study. *In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, (pp. 415-420). Barcelona, Spain.

Vuoskoski, J. K., & Eerola, T. (2010). Domain-specific or not? The applicability of different emotion models in the assessment of music-induced emotions. In *Proceedings of the 10th international conference on music perception and cognition* (pp. 196-199).

Whitman, B., & Smaragdis, P. (2002). Combining musical and cultural features for intelligent style detection. *In Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, (pp. 47–52). Paris, France.

Yang, D., & Lee, W. (2004). Disambiguating music emotion using software agents. *In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR*, (pp. 52-58). Barcelona, Spain.

Yang, Y. H., & Chen, H. H. (2011). *Music Emotion Recognition.* CRC Press.

Yang, Y., & Chen, H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST) , 3* (3).

Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., & Chen, H. H. (2008). Toward multi-modal music emotion classification. *In Proceedings of Pacific Rim Conference on Multimedia (PCM)* (pp. 70-79). Tainan, Taiwan: Springer.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing , 23* (3), 327-343.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008).Emotions evoked by the sound of music: Characterization,classification, and measurement. *Emotion*, 8(4), 494-521. doi:10.1037/1528-3542.8.4.494