

A Framework for Evaluating Multimodal Music Mood Classification

Xiao Hu

Faculty of Education, University of Hong Kong, Hong Kong. E-mail: xiaoxhu@hku.hk

Kahyun Choi and J. Stephen Downie

Graduate School of Library and Information Science, University of Illinois, Champaign, IL 61820. E-mail: {ckahyu2, jdownie}@illinois.edu

This research proposes a framework for music mood classification that uses multiple and complementary information sources, namely, music audio, lyric text, and social tags associated with music pieces. This article presents the framework and a thorough evaluation of each of its components. Experimental results on a large data set 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 the relative importance of audio and lyric features across mood categories.

Introduction

Music is an essential information type in people's everyday life. There are a large number of music collections, repositories, and websites that 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 mood¹ (Hu & Downie, 2007; Vignoli, 2004). Many of these music repositories have been relying on a 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 past decade, but many existing classification systems are solely based on information extracted from the audio recordings of music and have achieved suboptimal performance or reached a "glass ceiling" of performance (Barthet, Fazekas, & Sandler, 2013; Hu, Downie, Laurier, Bay, & Ehmann, 2008; Lu, Liu, & Zhang, 2006; Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2008; Yang & Chen, 2012).

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 carry semantics of human language, they have been exploited in music classification as well (e.g., Dakshina & Sridhar, 2014; He et al., 2008; Hu, Chen, & Yang, 2009b; Van Zaanen & Kanters, 2010). Furthermore, based on the hypothesis that lyrics and music audio² are different enough and thus may complement each other, researchers have started to combine lyrics and audio for improved classification performance (Björn, Johannes, & Gerhard, 2010; Brilis et al., 2012; Laurier, Grivolla, & Herrera, 2008; Yang et al., 2008). Such an approach of combining multiple information sources in solving classification problems is commonly called multimodal

Received October 22, 2014; revised August 14, 2015; accepted August 17, 2015

© 2016 ASIS&T • Published online 13 March 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23649

classification (Barthet et al., 2013; Kim et al., 2010; Yang & Chen, 2012).

Multimodal classification approaches in general are reported to have improved classification performance 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 and 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 data set of 5,296 unique songs, which exemplifies every stage of the framework and shows how the performance of music mood classification can be improved by a multimodal approach. Specifically, the novelty and contributions of this study can be summarized as follows:

1. Conceptualizes 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, and so on.
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 the effect of dimension reduction on feature spaces. Thus, it pushes forward the state-of-the-art on sentiment analysis for music lyrics.
3. Examines 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. Compares relative advantages of lyrics and audio across different mood categories. To date, there is little evidence as to 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. Builds a large ground truth data set for the task of multimodal music mood classification. The data set contains 5,296 unique songs in 18 mood categories. This is one of the largest experimental data sets in music mood classification with both audio and lyrics available (Barthet et al., 2013; Kim et al., 2010; Yang & Chen, 2012). Results from a large data set 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 designed to enrich 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 over the years it has evaluated more than 200 systems developed by research laboratories around the world. The data sets 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 the audio signal spectrum and have been widely used in music mood classification (e.g., Barthet et al., 2013; Hu et al., 2008; Lu et al., 2006; Pohle, Pampalk, & Widmer, 2005; Trohidis et al., 2008; Yang & Chen, 2012). As a supervised learning task, music classification often applies standard supervised learning models such as K-Nearest Neighbor (KNN), Gaussian Mixture Models (GMMs), and Support Vector Machines (SVMs). 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 data sets used in these studies are usually on smaller scales.

Besides lyrics, Bischoff, Firan, Nejdil, 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 and Eerola (2014) 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, & Beuriv , 2007; Muller, Kurth, Damm, Fremerey, & Clausen, 2007; Bj rn et al., 2010) or audio, scores, and text (McKay & Fujinaga, 2008). 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 data set 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 performance.

Most of the above studies found audio-based classifiers outperformed lyric-based classifiers, whereas Bischoff et al. (2009b) found social tag-based classifiers outperformed audio-based classifiers. Very few studies compared the relative advantages of different information sources across individual mood categories. The study by Schuller, Hage, Schuller, and Rigoll (2010) was an exception that revealed that 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 of mood categories that were often adapted from classical music psychology models, especially Russell’s model of the *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 data sets were relatively small, which limited the generalizability of the findings to a variety of music. It is also noteworthy that performance reported in these studies may not be directly comparable because they were evaluated with different data sets.

Research Questions

To fill the aforementioned gaps, this study answers the following research questions within the proposed framework:

1. Are there significant differences between the performance 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 advantages associated with 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: data set construction, feature generation and selection, classification and multimodal combination, and evaluation and analysis. This section describes each component in detail.

Data Set Construction

This component is used 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 data set mainly depend on which music audio files the researchers can gain access. As an effort to overcome this limitation, some research initiatives have started sharing extracted audio features of a 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 data sets. 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. Caution must be used in deriving ground truths from social tags, as social tags can be noisy and idiosyncratic (Saari & Eerola, 2014). The method described in Hu, Downie, and Ehmann (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 lyrics.wikia.com. Although Figure 1 only shows two information sources (that are used in the experiment in this study), the framework can include more 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 on features extracted from information sources. Although features

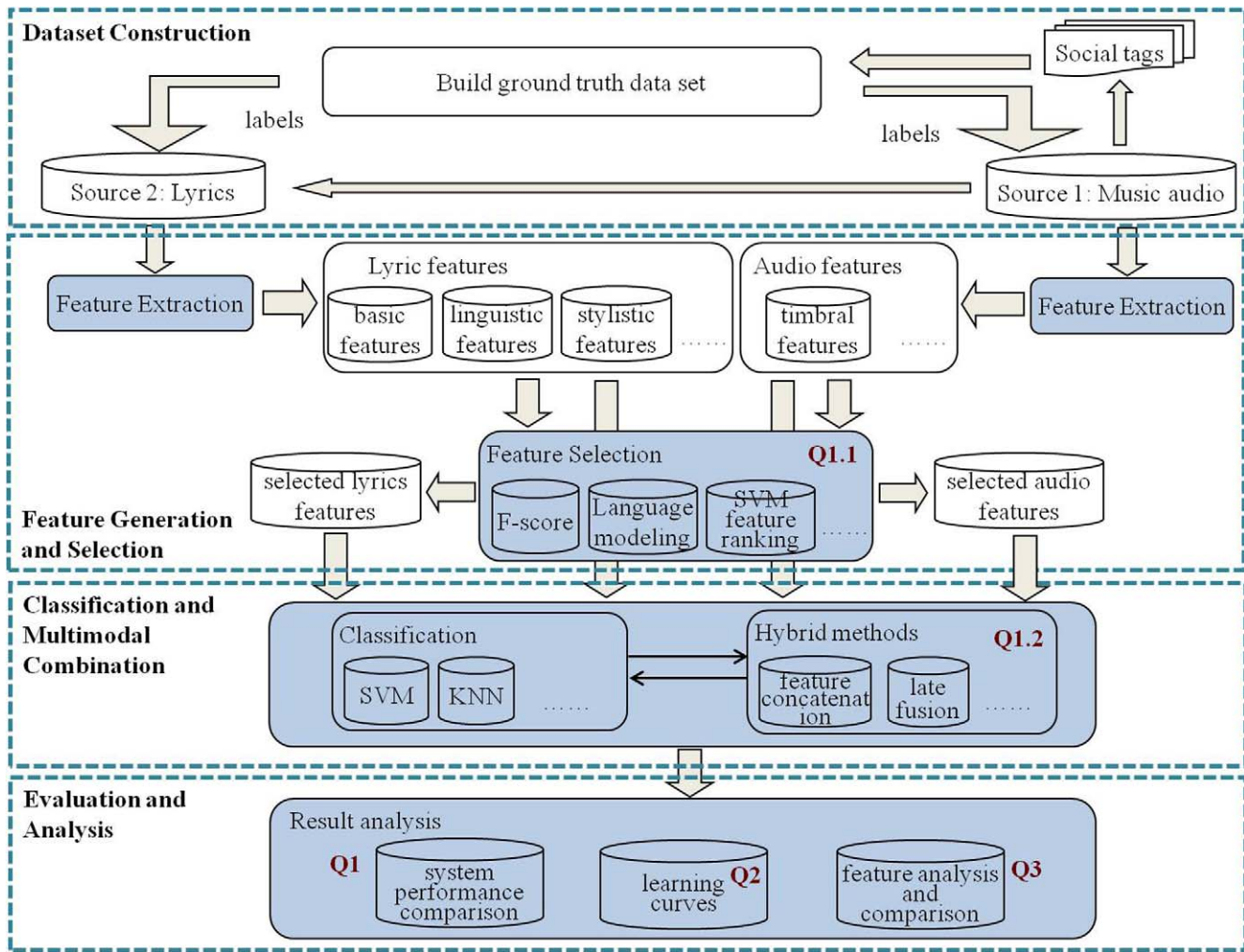


FIG. 1. Proposed framework for multimodal music mood classification (research questions in this study are marked Q1 to Q3). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

extracted from all available information sources are important for multimodal classification, in this study we pay more attention to lyric features because most existing studies have focused on analyzing audio features (Baume, Fazekas, Barthet, Marston, & Sandler, 2014; Saari, Eerola, & Lartillot, 2011; Song, Dixon, & Pearce, 2012). 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 (a) General Inquirer (GI), a psycholinguistic lexicon mapping English words to psychological categories (Stone, 1966), (b) WordNet (Fellbaum, 1998), (c) WordNet-Affect, an extension of WordNet in the affect domain (Strapparava & Valitutti, 2004), and (d) 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 and Downie (2010).

Studies in text categorization have shown that feature selection can improve the 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 SVM algorithm.

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 , $k = 1, \dots, m$, 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 data sets, 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.

Chi-square (χ^2)

In statistics the chi-square test is used 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 of 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.

SVM Feature Weighting

A trained decision function in a linear SVM contains a weight for each feature that 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 the dimensionality of feature spaces (Chapelle, Vapnik, Bousquet, & Mukherjee, 2002; Guyon, Weston, Barnhill, & Vapnik, 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), who used fuzzy clustering, most existing studies on music classification used standard supervised learning models such as KNN, GMM, and SVM. Although the proposed framework can include multiple classification models (which can then be compared using a common feature set), many studies have found that SVMs are effective in both music mood classification (Barthet et al., 2013; Hu et al., 2008; Kim et al., 2010; Yang & Chen, 2012) as well as text categorization (e.g., Yu, 2008). Therefore, this study uses SVM in the experiment without focusing on an evaluation of classifiers.

Because 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 with 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:

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

where α is the weight given to the posterior probability estimated by the lyric-based classifier, with a 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, and 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 performance, 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 performance 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

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	romantic	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

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 (Bischoff et al., 2009b; He et al., 2008; Hu et al., 2009a; Laurier et al., 2008). 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 performance but also that of feature spaces of multiple sources. Classification performance tell us whether a certain experiment setup works, whereas 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 below.

Data Set

The data set 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 data set 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 presents 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.

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, 2008). MARSYAS has taken part in the AMC task from 2007 to 2012, with consistently top-ranked performance. 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), and 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 that gives equal importance to all categories. Within each category, accuracy was calculated with a 10-fold cross-validation. A nonparametric Kruskal–Wallis test was applied to compare performance, as the accuracy data may not conform to normal distribution (Downie, 2008). When comparing performance 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 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 performance on categories G2 (“sad”) and G16 (“earnest”). The

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%

results indicate no significant difference among the three classifiers ($p = .58$ for G, $p = .06$ for G16). As the linear kernel achieved higher accuracies and is computationally efficient, it is used with the default parameter throughout the experiments.

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 the experimental results in the order described in the proposed framework.

Best Lyric Features

As reported in Hu and Downie (2010), there were seven types of lyric features compared with one another, including (a) content words, (b) part-of-speech, (c) function words, (d) affective words, (e) psychological categories in GI, (f) scores derived from the ANEW, and (g) 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 parts-of-speech (Hu & Downie, 2010). In this study we 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 data set 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, the F-score and SVM feature selection methods did not perform as well as the original BEST-all feature set, whereas the chi-square method (χ^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 ($H = 1.45$, $df = 3$, $p = .70$), whereas on average 35% to 44% of 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 “BEST- χ^2 ”) are evaluated and compared.

Hybrid Systems

The performance of the audio-based classifier is 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

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 $p < .05$.

TABLE 4. Accuracies of late fusion across categories.

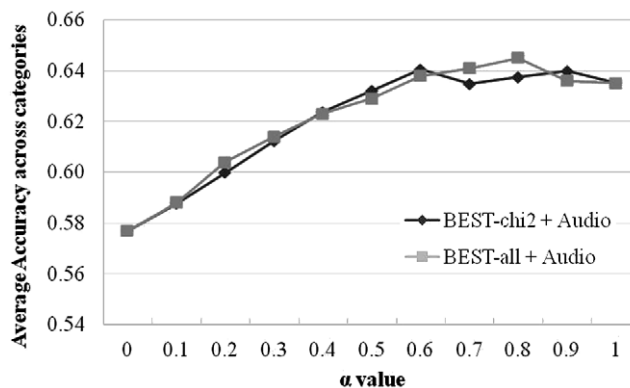
Mood category	Audio-only	BEST-chi ² + Audio		BEST-all + Audio		Lyric-only (BEST-all)
	Accuracy	Accuracy	$\bar{\alpha}$	Accuracy	$\bar{\alpha}$	Accuracy
calm	0.657	0.679^a	0.43	0.681^a	0.63	0.612
sad	0.676	0.716	0.53	0.719^{ab}	0.51	0.669
glad	0.590	0.637	0.52	0.640	0.47	0.613
romantic	0.617	0.699^b	0.57	0.721^b	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^b	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^b	0.87	0.593^b	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

^aThe accuracy is significantly higher than that of BEST-all at $p < .05$.

^bThe accuracy is significantly higher than that of Audio-only at $p < .05$.

(Equation 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 the audio-only system on most categories (Table 4). On average, the α value chosen in BEST-chi² + Audio ($\alpha = 0.67$) is larger than that in BEST-all + Audio system ($\alpha = 0.33$), indicating higher weights to the lyric classifiers were used for BEST-chi² + 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: (a) the lyrics in these categories contain words semantically representative of the category such as “kiss” in “romantic” song and “fight” in “angry” songs, and (b) 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.

To illustrate the trend of how α values affect the classification performance, 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 that these experiments were not aimed to evaluate the effectiveness of the late fusion method, but instead to compare α values that can reflect the relative importance of the audio-based and lyric-based classifiers. In both cases, performance improved quickly in a steady manner when α was increased from 0.1 to 0.6, indicating that even a 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

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

set, whereas the BEST-chi² + Audio system performed the best when α was 0.6, giving a small increase in weight to the audio-based classifier.

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-chi², but was helpful for BEST-all. Late fusion was a good method for both lyric feature sets, improving the 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 ($p < .05$). 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

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 the $p < .05$ level.

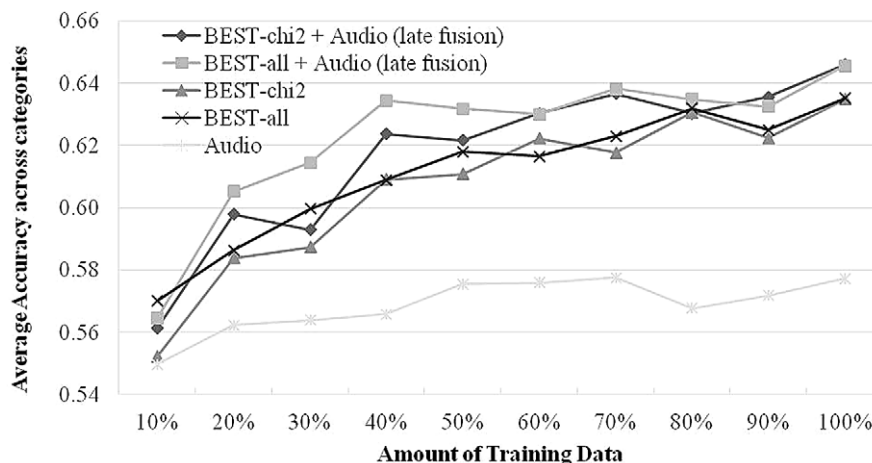


FIG. 3. Learning curves of hybrid and single-source systems.

shown that 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.

Effects on Training Data Size

Number of training examples. 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-chi² 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 performance 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-chi²) training examples, the hybrid systems achieved comparable performance to that of the 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.

It is also noteworthy that the BEST-all + Audio system seems to have an advantage over BEST-chi² + Audio in reaching better performance 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 performance of the audio-only and the late fusion hybrid systems on data sets 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 the 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.

The hybrid systems outperformed the 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 performance.

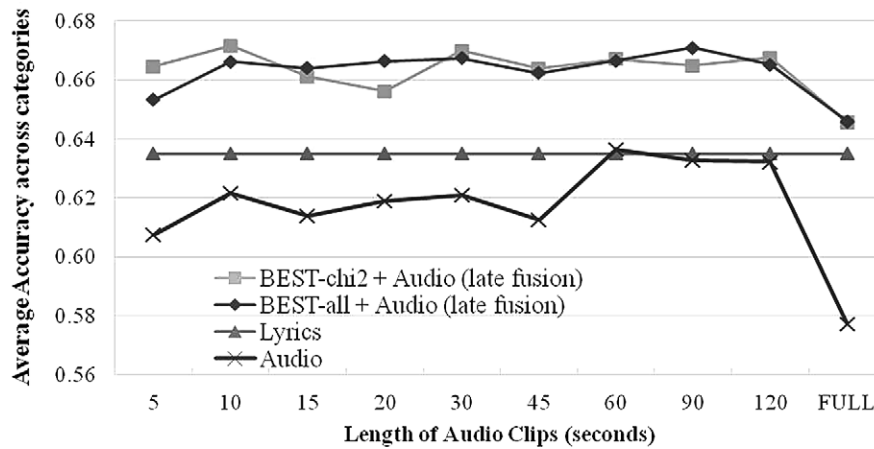


FIG. 4. System performance with varied audio lengths.

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

Category	Better	Worse	<i>p</i>	Category	Better	Worse	<i>p</i>
hopeful	BEST-all	Audio	.005	calm	Audio	BEST-chi ²	.001
hopeful	BEST-chi ²	Audio	.009	calm	Audio	BEST-all	.010
angry	BEST-all	Audio	.012	romantic	BEST-all	Audio	.001
angry	BEST-chi ²	Audio	.045	romantic	BEST-chi ²	Audio	.001
aggressive	BEST-all	Audio	.045	exciting	BEST-chi ²	Audio	.011

For the two hybrid systems, the performance within each system did not make any significant difference ($p < .05$). 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 performance 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.

For all three systems with audio input, the full track did not perform well, which suggests that shorter audio clips would work better with music mood classification. One possible reason could be that music mood can vary during 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 that 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 performance 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.”

To facilitate the comparison and contrast among the categories, we plotted the 18 mood categories in a 2D 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 data set. 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, 2014), with reference to the positions of the six overlapped terms in the well-adopted Russell 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 (Barthet et al., 2013; Kim et al., 2010; Yang & Chen, 2012) and shares commonality with many other mood models such as the equally influential Hevner’s adjective circles and the Geneva Emotional Music Scale (GEMS) (Gabrielsson & Lindström, 2001; Hevner, 1936; Vuoskoski & Eerola, 2010; Zentner, Grandjean, & Scherer, 2008). 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.

Summary of Experimental Results

The results answer the proposed research questions:

1. Multimodal systems combining lyric and audio features significantly outperformed the 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.

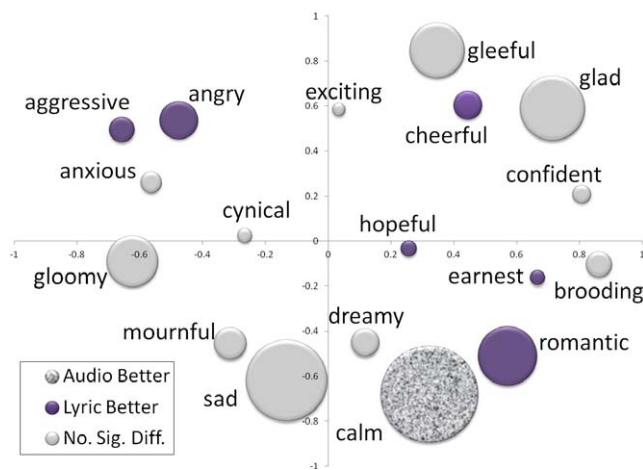


FIG. 5. The 18 mood categories plotted in a 2D space. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Conclusions and Future Work

This study proposed a framework for evaluating multimodal music mood classification systems that 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 data set of a significant scale. In addition, the study also analyzed the effects of a 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 for 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. This study contributes to

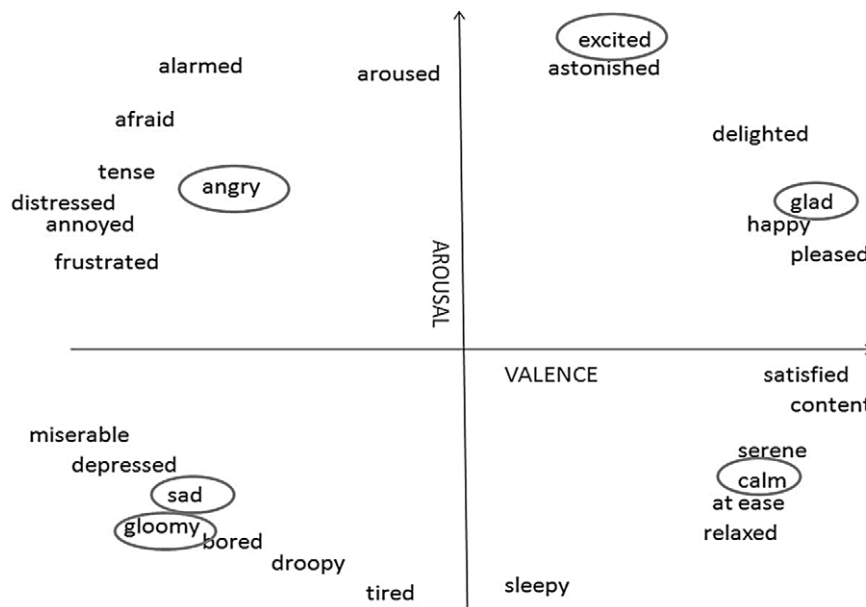


FIG. 6. Russell's 2D model of music mood (Russell, 1980, p. 1168). The terms matching those in Figure 5 are circled.

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 did not 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 of the correlations between multimodal features and mood categories will be conducted to find out why certain sources are more helpful for certain mood categories.

Acknowledgments

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

Endnotes

1. In the literature, music mood is also referred to 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 recognize the difference between the two and use “mood” throughout to refer to 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 the current stage of music information retrieval research is regarded as independent of the lyrics.

References

- Aggarwal, C.C., & Zhai, C. (2012). A survey of text classification algorithms. In C.C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 163–222). New York: Springer.
- 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://ayasha.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: ISMIR.
- 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 (Eds.), *From sounds to music and emotions* (pp. 228–252). Berlin, Heidelberg: Springer.
- Baume, C., Fazekas, G., Barthet, M., Marston, D., & Sandler, M. (2014). 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. In Proceedings of the 12th International Society for Music Information Retrieval Conference (pp. 591–596). Miami, FL: University of Miami.
- Bischoff, K., Firan, C.S., Nejdil, W., & Paiu, R. (2009a). How do you feel about “Dancing Queen”? Deriving mood and theme annotations

- from user tags. In Proceedings of Joint Conference on Digital Libraries (JCDDL) (pp. 285–294). New York: ACM Press.
- Bischoff, K., Firan, C., Paiu, R., Nejdil, W., Laurier, C., & Sordo, M. (2009b). 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: ISMIR.
- Björn, S., Johannes, D., & Gerhard, R. (2010). Determination of nonprototypical valence and arousal in popular music: Features and performance. *Journal on Audio, Speech, and Music Processing*, 2010, 735854.
- Bradley, M.M., & Lang, P.J. (1999). *Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings*. Technical report C-1. Gainesville, FL: University of Florida.
- Brlis, 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 (Eds.), *Artificial intelligence applications and innovations* (pp. 421–430). Berlin, Heidelberg: Springer.
- Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters 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.), *Feature extraction, foundations and applications* (pp. 315–324). Berlin Heidelberg: 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: ISMIR.
- Dakshina, K., & Sridhar, R. (2014). LDA based emotion recognition from lyrics. In M. Kumar Kundu, D.P. Mohapatra, A. Konar, & A. Chakraborty (Eds.), *Advanced computing, networking and informatics* (pp. 187–194). Switzerland: 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.
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, UK: 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 & J.A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 223–248). New York: Oxford University Press.
- 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: ISMIR.
- 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: ISMIR.
- Hu, X., & Downie, J.S. (2010). Improving mood classification in music digital libraries by combining lyrics and audio. In Proceedings of the Joint Conference on Digital Libraries (pp. 159–168). Surfers Paradise, Australia: ACM.

- 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: ISMIR.
- Hu, X., Downie, J.S., & Ehmann, A.F. (2009a). Lyric text mining in music mood classification. In Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR) (pp. 411–416). Kobe, Japan: ISMIR.
- Hu, Y., Chen, X., & Yang, D. (2009b). 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: ISMIR.
- 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, the Netherlands: ISMIR.
- 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. In Proceedings of the 12th Annual ACM International Conference on Multimedia (Vol. 8, pp. 364–367). New York: ACM.
- 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: ISMIR.
- 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: ISMIR.
- McVicar, M., Freeman, T., & De Bie, T. (2011). Mining the correlation between lyrical and audio features and the emergence of mood. In Proceedings of the 12th Conference of International Society for Music Information Retrieval (pp. 783–788). Miami, FL: ISMIR.
- 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.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 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.
- Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178.
- Saari, P., & Eerola, T. (2014). Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, 26(10), 2548–2560.
- Saari, P., Eerola, T., & Lartillot, O. (2011). Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing* 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. In Proceedings of the 10th International Society for Music Information Retrieval Conference (pp. 451–456). Philadelphia, PA: ISMIR.
- Song, Y., Dixon, S., & Pearce, M. (2012). Evaluation of musical features for emotion classification. In Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR) (pp. 523–528). Porto, Portugal: ISMIR.
- Stone, P.J. (1966). *General Inquirer: A computer approach to content analysis*. Cambridge, MA: 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.
- 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.
- 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: ISMIR.
- Tzanetakis, G., & Lemstrom, K. (2008). Marsyas-0.2: A case study in implementing music information retrieval systems. In S. Shen & L. Cui (Eds.), *Information science reference. Intelligent music information systems: Tools and methodologies* (pp. 31–49). Hershey, PA: IGI Global.
- 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, the Netherlands: ISMIR.
- 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: ISMIR.
- 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 11th International Conference on Music Perception and Cognition (pp. 196–199). Seattle, WA: ICMPAC.
- Whitman, B., & Smaragdīs, 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.
- 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., & Chen, H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3), Article 40, 1–30.
- 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 (Washington, D.C.)*, 8(4), 494–521.