WHEN LYRICS OUTPERFORM AUDIO FOR MUSIC MOOD CLASSIFICATION: A FEATURE ANALYSIS

Xiao Hu **J. Stephen Downie**
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
xiaohu@illinois.edu jdownie@illinois.edu

ABSTRACT

This paper builds upon and extends previous work on multi-modal mood classification (i.e., combining audio and lyrics) by analyzing in-depth those feature types that have shown to provide statistically significant improvements in the classification of individual mood categories. The dataset used in this study comprises 5,296 songs (with lyrics and audio for each) divided into 18 mood categories derived from user-generated tags taken from last.fm. These 18 categories show remarkable consistency with the popular Russell's mood model. In seven categories, lyric features significantly outperformed audio spectral features. In one category only, audio outperformed all lyric features types. A fine grained analysis of the significant lyric feature types indicates a strong and obvious semantic association between extracted terms and the categories. No such obvious semantic linkages were evident in the case where audio spectral features proved superior.

1. INTRODUCTION

User studies in Music Information Retrieval (MIR) have found that music mood is a desirable access point to music repositories and collections (e.g., [1]). In recent years, automatic methods have been explored to classify music by mood. Most studies exploit the audio content of songs, but some studies have been using song lyrics in music mood classification as well [2-4].

Music mood classification studies using both audio and lyrics consistently find that combining lyric and audio features improves classification performance (See Section 2.3). However, there are contradictory findings on whether audio or lyrics are more useful in predicting music mood, or which source is better for individual mood classes. In this paper, we continue our previous work on multi-modal mood classification [4] and go one step further to investigate these research questions: 1) Which source is more useful in music classification: audio or lyrics? 2) For which moods is audio more useful and for which moods are lyrics more useful? and, 3) How do lyric features associate with different mood categories? Answers to these questions can help shed light on a profoundly important music perception question: How does the interaction of sound and text establish a music mood?

This paper is organized as follows: Section 2 reviews related work on music mood classification. Section 3 introduces our experimental dataset and the mood categories used in this study. Section 4 describes the lyric and audio features examined. Section 5 discusses our findings in light of our research questions. Section 6 presents our conclusions and suggests future work.

2. RELATED WORK

2.1 Music Mood Classification Using Audio Features

Most existing work on automatic music mood classification is exclusively based on audio features among which spectral and rhythmic features are the most popular (e.g., [5-7]). Since 2007, the Audio Mood Classification (AMC) task has been run each year at the Music Information Retrieval Evaluation eXchange (MIREX) [8], the community-based framework for the formal evaluation of MIR techniques. Among the various audio-based approaches tested at MIREX, spectral features and Support Vector Machine (SVM) classifiers were widely used and found quite effective [9].

2.2 Music Mood Classification Using Lyric Features

Studies on music mood classification solely based on lyrics have appeared in recent years (e.g., [10,11]). Most used bag-of-words (BOW) features in various unigram, bigram, trigram representations. Combinations of unigram, bigram and trigram tokens performed better than individual n-grams, indicating higher-order BOW features captured more of the semantics useful for mood classification. Features used in [11] were novel in that they were extracted based on a psycholinguistic resource, an affective lexicon translated from the Affective Norm of English Words (ANEW) [12].

2.3 Multi-modal Music Mood Classification Using Both Audio and Lyric Features

Yang and Lee [13] is often regarded as one of the earliest studies on combining lyrics and audio in music mood classification. They used both lyric BOW features and the 182 psychological features proposed in the General Inquirer [14] to disambiguate categories that audio-based classifiers found confusing. Besides showing improved classification accuracy, they also presented the most salient psychological features for each of the considered mood categories. Laurier et al. [2] also combined audio and lyric BOW features and showed that the combined features improved classification accuracies in all four of their categories. Yang et al. [3] evaluated both unigram and bigram BOW lyric features as well as three methods for fusing lyric and audio sources and concluded that leveraging lyrics could improve classification accuracy over audio-only classifiers.

Our previous work [4] evaluated a wide range of lyric features from n-grams to features based on psycholinguistic resources such as WordNet-Affect [15], General Inquirer and ANEW, as well as their combinations. After identifying the best lyric feature types, audio-based, lyric-based as well as multi-modal classification systems were compared. The results showed the multi-modal system performed the best while the lyric-based system outperformed the audio-based system. However, our reported performances were accuracies ***averaged*** across all of our 18 mood categories. In this study, we go deeper to investigate the performance differences of the aforementioned feature types on ***individual*** mood categories. More precisely, this paper examines, in some depth, those feature types that provide statistically significant performance improvements in identifying individual mood categories.

2.4 Feature Analysis in Text Sentiment Classification

Except for [13], most existing studies on music mood classification did not analyze or compare which specific feature values were the most useful. However, feature analysis has been widely used in text sentiment classification. For example, a study on blogs, [16] identified discriminative words in blog postings between two categories, "happy" and "sad" using Naïve Bayesian classifiers and word frequency thresholds. [17] uncovered important features in classifying customer reviews with regard to ratings, object types, and object genres, using frequent pattern mining and naïve Bayesian ranking. Yu [18] presents a systematic study of sentiment features in Dickenson's poems and American novels. Besides identifying the most salient sentiment features, it also concluded that different classification models tend to identify different important features. These previous works inspired the feature ranking methods examined in this study.

3. DATASET AND MOOD CATEGORIES

3.1 Experimental Dataset

As mentioned before, this study is a continuation of a previous study [4], and thus the same dataset is used. There are 18 mood categories represented in our dataset, and each of the categories comprises 1 to 25 mood-related social tags downloaded from last.fm. A mood category consists of tags that are synonyms identified by WordNet-Affect and verified by two human experts who are both native English speakers and respected MIR researchers. The song pool was limited to those audio tracks at the intersection of being available to the authors, having English lyrics available on the Internet, and having social tags available on last.fm. For each of these songs, if it was tagged with any of the tags associated with a mood category, it was counted as a positive example of that category. In this way, one single song could belong to multiple mood categories. This is in fact more realistic than a single-label setting since a music piece may carry multiple moods such as "happy and calm" or "aggressive and depressed".

A binary classification approach was adopted for each of the mood categories. Negative examples of a category were songs that were not tagged with any of the tags associated with this category but were heavily tagged with many other tags. Table 1 presents the mood categories and the number of positive songs in each category. We balanced equally the positive and negative set sizes for each category. This dataset contains 5,296 unique songs in total. This number is much smaller than the total number of examples in all categories (which is 12,980) because categories often share samples.

Category	No. of songs	Category	No. of songs	Category	No. of songs
calm	1,680	angry	254	anxious	80
sad	1,178	mournful	183	confident	61
glad	749	dreamy	146	hopeful	45
romantic	619	cheerful	142	earnest	40
gleeful	543	brooding	116	cynical	38
gloomy	471	aggressive	115	exciting	30

Table 1. Mood categories and number of positive examples

3.2 Mood Categories

Music mood categories have been a much debated topic in both MIR and music psychology. Most previous studies summarized in Section 2 used two to six mood categories which were derived from psychological models. Among the many emotion models in psychology, Russell's model [19] seems the most popular in MIR research (e.g., [2, 5]).

Russell's model is a *dimensional* model where emotions are positioned in a continuous multidimensional space. There are two dimensions in Russell's model: *valence* (negative-positive) and *arousal* (inactive-active). As shown in Figure 1, this model places 28 emotion-denoting adjectives on a circle in a bipolar space subsuming these two dimensions.

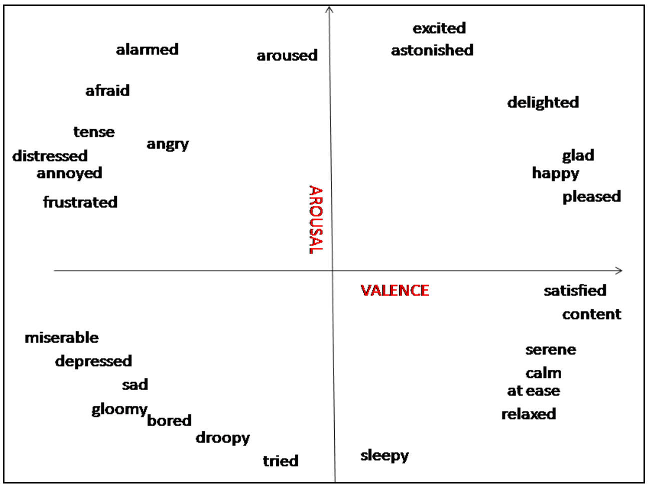

Figure 1. Russell's model with two dimensions

From Figure 1, we can see that Russell's space demonstrates relative distances or similarities between moods. For instance, "sad" and "happy", "calm" and "angry" are at opposite places while "happy" and "glad" are close to each other.

The relative distance between the 18 mood categories in our dataset can also be calculated by co-occurrence of songs in the positive examples. That is, if two categories share many positive songs, they should be similar. Figure 2 illustrates the relative distances of the 18 categories plotted in a 2-dimensional space using Multidimensional Scaling where each category is represented by a bubble in a size proportional to the number of positive songs in this category.

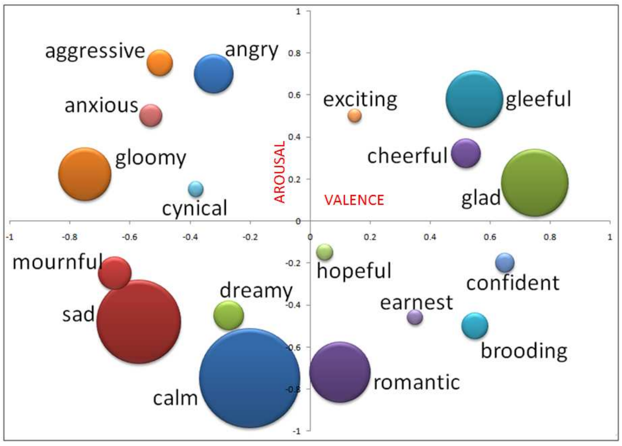

Figure 2. Distances between the 18 mood categories in the experimental dataset

The patterns shown in Figure 2 are similar to those found in Figure 1: 1) Categories placed together are intuitively similar; 2) Categories at opposite positions represent contrasting moods; 3) The horizontal and vertical dimensions correspond to *valence* and *arousal* respectively. Taken together, these similarities indicate that our 18 mood categories fit well with Russell's mood model which is the most commonly used model in MIR mood classification research.

4. LYRIC AND AUDIO FEATURES

In [4], we systematically evaluated a range of lyric feature types on the task of music mood classification, including: 1) basic text features that are commonly used in text categorization tasks; 2) linguistic features based on psycholinguistic resources; and, 3) text stylistic features. In this study, we analyze the most salient features in each of these feature types. This section briefly introduces these feature types. For more detail, please consult [4].

4.1 Features based on N-grams of Content Words

"Content words" (CW) refer to all words appearing in lyrics except function words (also called "stop words"). Words were not stemmed as our earlier work showed stemming did not yield better results. The CW feature set used was a combination of unigrams, bigrams and trigrams of content words since this combination performed better than each of the n-gram types individually [4]. For each n-gram, features that occurred less than five times in the training dataset were discarded. Also, for bigrams and trigrams, function words were not eliminated because content words are usually connected via function words as in "I love you" where "I" and "you" are function words. There were totally 84,155 CW n-gram features.

4.2 Features based on General Inquirer

General Inquirer (GI) is a psycholinguistic lexicon containing 8,315 unique English words and 182 psychological categories [14]. Each of the 8,315 words in the lexicon is manually labeled with one or more of the 182 psychological categories to which the word belongs. For example, the word "happiness" is associated with the categories "Emotion", "Pleasure", "Positive", "Psychological well being", etc. GI's 182 psychological features were a feature type evaluated in [4], and denoted as "GI".

Each of the 8,315 words in General Inquirer conveys certain psychological meanings and thus were evaluated in [4]. In this feature set (denoted as "GI-lex"), feature vectors were built using only these 8,315 words.

4.3 Features based on ANEW and WordNet

Affective Norms for English Words (ANEW) is another specialized English lexicon [12]. It contains 1,034 unique English words with scores in three dimensions: *valence* (a scale from unpleasant to pleasant), *arousal* (a scale from calm to excited), and *dominance* (a scale from submissive to dominated). As these 1,034 words are too few to cover all the songs in our dataset, we expanded the ANEW word list using WordNet [20] such that synonyms of the 1,034 words were included. This gave us 6,732 words in the expanded ANEW. We then further expanded this set of affect-related words by including the 1,586 words in WordNet-Affect [15], an extension of WordNet containing emotion related words. Therefore, this set of 7,756 affect-related words formed a feature type denoted as "Affe-lex".

4.4 Text Stylistic Features

The text stylistic features evaluated in [4] included such text statistics as number of unique words, number of unique lines, ratio of repeated lines, number of words per minute, as well as special punctuation marks (e.g., "!") and interjection words (e.g., "hey"). There were 25 text stylistic features in total.

4.5 Audio Features

In [4] we used the audio features selected by the MARSYAS submission [21] to MIREX because it was the leading audio-based classification system evaluated under both the 2007 and 2008 Audio Mood Classification (AMC) task. MARSYAS used 63 spectral features: means and variances of Spectral Centroid, Rolloff, Flux, Mel-Frequency Cepstral Coefficients (MFCC), etc. Although there are audio features beyond spectral ones, spectral features were found the most useful and most commonly adopted for music mood classification [9]. We leave it as our future work to analyze a broader range of audio features.

5. RESULTS AND DISCUSSIONS

5.1 Feature Performances

Table 2 shows the accuracies of each aforementioned feature set on individual mood categories. Each of the accuracy values was averaged across a 10-fold cross validation. For each lyric feature set, the categories where its accuracies are significantly higher than that of the audio feature set are marked as bold (at $p < 0.05$). Similarly, for the audio feature set, bold accuracies are those significantly higher than all lyric features (at $p < 0.05$).

Category	CW	GI	GI-lex	Affe-lex	Stylistic	Audio
calm	0.5905	0.5851	0.5804	0.5708	0.5039	**0.6574**
sad	0.6655	0.6218	0.6010	0.5836	0.5153	0.6749
glad	0.5627	0.5547	0.5600	0.5508	0.5380	0.5882
romantic	**0.6866**	0.6228	**0.6721**	0.6333	0.5153	0.6188
gleeful	0.5864	0.5763	0.5405	0.5443	0.5670	0.6253
gloomy	0.6157	0.5710	0.6124	0.5859	0.5468	0.6178
angry	**0.7047**	0.6362	0.6497	**0.6849**	0.4924	0.5905
mournful	0.6670	0.6344	0.5871	0.6615	0.5001	0.6278
dreamy	0.6143	0.5686	0.6264	0.6269	0.5645	0.6681
cheerful	**0.6226**	0.5633	0.5707	0.5171	0.5105	0.5133
brooding	0.5261	0.5295	0.5739	0.5383	0.5045	0.6019
aggressive	**0.7966**	**0.7178**	**0.7549**	0.6746	0.5345	0.6417
anxious	**0.6125**	0.5375	0.5750	0.5875	0.4875	0.4875
confident	0.3917	0.4429	0.4774	0.5548	0.5083	0.5417
hopeful	**0.5700**	0.4975	**0.6025**	**0.6350**	**0.5375**	0.4000
earnest	0.6125	0.6500	0.5500	0.6000	0.6375	0.5750
cynical	0.7000	0.6792	0.6375	0.6667	0.5250	0.6292
exciting	0.5833	0.5500	**0.5833**	0.4667	**0.5333**	0.3667
AVERAGE	0.6172	0.5855	0.5975	0.5935	0.5290	0.5792

Table 2.Accuracies of feature types for individual categories

From the averaged accuracies in Table 2, we can see that whether lyrics are more useful than audio, or vice versa depends on which feature sets are used. For example, if using CW n-grams as features, lyrics are more useful than audio spectral features in terms of overall classification performance averaged across all categories. However, the answer is reversed if text stylistics is used as lyric features (i.e., audio works better).

The accuracies marked in bold in Table 2 demonstrate that lyrics and audio have their respective advantages in ***different*** mood categories. Audio spectral features significantly outperformed *all* lyric feature types in only one mood category: "calm". However, lyric features achieved significantly better performance than audio in seven divergent categories: "romantic", "angry", "cheerful", "aggressive", "anxious", "hopeful" and "exciting".

In the following subsections, we will rank (by order of influence), and then examine, the most salient features of those lyric feature types that outperformed audio features in the seven aforementioned mood categories. Support Vector Machines (SVM) were adopted as the classification model in [4] where a variety of kernels were tested and a linear kernel was finally chosen. In a linear SVM, each feature was assigned a weight indicating its influence in the classification model, and thus the features in this study were ranked by the assigned weights in the same SVM models trained in experiments in [4].

5.2 Top Features in Content Word N-Grams

There are six categories where CW n-gram features significantly outperformed audio features. Table 3 lists the top-ranked content word features in these categories. Note how "love" seems an eternal topic of music regardless of the mood category! Highly ranked content words seem to have intuitively meaningful connections to the categories, such as "with you" in "romantic" songs, "happy" in "cheerful" songs, and "dreams" in "hopeful" songs. The categories, "angry", "aggressive" and "anxious" share quite a few top-ranked terms highlighting their emotional similarities. It is interesting to note that these last three categories sit in the same top-left quadrant in Figure 2.

romantic	cheerful	hopeful	angry	aggressive	anxious
with you	i love	you ll	baby	fuck	hey
on me	night	strong	i am	dead	to you
with your	ve got	i get	shit	i am	change
crazy	happy	loving	scream	girl	left
come on	for you	dreams	to you	man	fuck
i said	new	i ll	run	kill	i know
burn	care	if you	shut	baby	dead
hate	for me	to be	i can	love	and if
kiss	living	god	control	hurt	wait
let me	rest	lonely	don t know	but you	waiting
hold	and now	friend	dead	fear	need
to die	all around	dream	love	don t	i don t
why you	heaven	in the eye	hell	pain	i m
i ll	met	coming	fighting	lost	listen
tonight	she says	want	hurt you	i ve never	again and
i want	you ve got	wonder	kill	hate	but you
love	more than	waiting	if you want	have you	my heart
give me	the sun	i love	oh baby	love you	hurt
cry	you like	you best	you re my	yeah yeah	night

Table 3. Top-ranked content word features for moods where content words significantly outperformed audio

5.3 Top-Ranked Features Based on General Inquirer

"Aggressive" is the only category where the GI set of 182 psychological features outperformed audio features with a statistically significant difference. Table 4 lists the top GI features for this category.

GI Feature	Example Words
Words connoting the physical aspects of well being, including its absence	blood, dead, drunk, pain
Words referring to the perceptual process of recognizing or identifying something by means of the senses	dazzle, fantasy, hear, look, make, tell, view
Action words	hit, kick, drag, upset
Words indicating time	noon, night, midnight
Words referring to all human collectivities	people, gang, party
Words related to a loss in a state of well being,	burn, die, hurt, mad

Table 4. Top GI features for "aggressive" mood category

It is somewhat surprising that the psychological feature indicating "hostile attitude or aggressiveness" (e.g., "devil", "hate", "kill") was ranked at 134 among the 182 features. Although such individual words ranked high as content word features, the GI features were aggregations of certain kinds of words. The mapping between words and psychological categories provided by GI can be very helpful in looking beyond word forms and into word meanings.

By looking at rankings on specific words in General Inquirer, we can have a clearer understanding about which GI words were important. Table 5 presents top GI word features in the four categories where "GI-lex" features significantly outperformed audio features.

romantic	aggressive	hopeful	exciting
paradise	baby	i'm	come
existence	fuck	been	now
hit	let	would	see
hate	am	what	up
sympathy	hurt	do	will
jealous	girl	in	tear
kill	be	lonely	bounce
young	another	saw	to
destiny	need	like	him
found	kill	strong	better
anywhere	can	there	shake
soul	but	run	everything
swear	just	will	us
divine	because	found	gonna
across	man	when	her
clue	one	come	free
rascal	dead	lose	me
tale	alone	think	more
crazy	why	mine	keep

Table 5. Top-ranked GI-lex features for categories where GI-lex significantly outperformed audio

5.4 Top Features Based on ANEW and WordNet

According to Table 2, "Affe-lex" features worked significantly better than audio features on categories "angry" and "hopeful". Table 6 presents top-ranked features.

Category	Top Features (in order of influence)
angry	one, baby, surprise, care, death, alive, guilt, happiness, hurt, straight, thrill, cute, suicide, babe, frightened, motherfucker, down, misery, mad, wicked, fighting, crazy
hopeful	wonderful, sun, words loving, read, smile, better, heart, lonely, friend, free, hear, come, found, strong, letter, grow, safe, god, girl, memory, happy, think, dream

Table 6. Top Affe-lex features for categories where Affe-lex significantly outperformed audio

Again, these top-ranked features seem to have strong semantic connections to the categories, and they share common words with the top-ranked features listed in Tables 3 and 5. Although both Affe-lex and GI-lex are domain-oriented lexicons built from psycholinguistic resources, they contain different words, and thus each of them identified some novel features that are not shared by the other.

5.5 Top Text Stylistic Features

Text stylistic features performed the worst among all feature types considered in this study. In fact, the average accuracy of text stylistic features was significantly worse than each of the other feature types ($p < 0.05$). However, text stylistic features did outperform audio features in two categories: "hopeful" and "exciting". Table 7 shows the top-ranked stylistic features in these two categories.

Note how the top-ranked features in Table 7 are all text statistics without interjection words or punctuation marks. These kinds of text statistics capture very different characteristics of the lyrics from other word-based features, and thus combining these statistics and other features may yield better classification performance. Also noteworthy is that these two categories both have relatively low positive valence (but opposite arousal) as shown in Figure 2.

hopeful	exciting
Std of number of words per line	Average number of unique words per line
Average number of unique words per line	Average repeating word ratio per line
Average word length	Std of number of words per line
Ratio of repeating lines	Ratio of repeating words
Average number of words per line	Ratio of repeating lines
Ratio of repeating words	Average number of words per line
Number of unique lines	Number of blank lines

Table 7. Top-ranked text stylistic features for categories where text stylistics significantly outperformed audio

5.6 Top Lyric Features in "Calm"

"Calm", which sits in the bottom-left quadrant and has the lowest arousal of any category (Figure 2), is the only mood category where audio features were significantly better than all lyric feature types. It is useful to compare the top lyric features in this category to those in categories where lyric features outperformed audio features. Top-ranked words and stylistics from various lyric feature types in "calm" are shown in Table 8.

CW	GI-lex	Affe-lex	Stylistic
you all look	float	list	Standard derivation (std) of repeating word ratio per line
all look	eager	moral	
all look at	irish	saviour	Repeating word ratio
you all i	appreciate	satan	Average repeating word ratio per line
burning	kindness	collar	
that is	selfish	pup	Repeating line ratio
you d	convince	splash	Interjection: "Hey"
control	foolish	clams	Average number of unique words per line
boy	island	blooming	
that s	curious	nimble	Number of lines per minute
all i	thursday	disgusting	Blank line ratio
believe in	pie	introduce	Interjection: "ooh"
be free	melt	amazing	Average number of words per line
speak	couple	arrangement	
blind	team	mercifully	Interjection: "ah"
beautiful	doorway	soaked	Punctuation: "!"
the sea	lowly	abide	Interjection: "yo"

Table 8. Top lyric features in "calm" category

As Table 8 indicates, top-ranked lyric words from the CW, GI-lex and Affe-lex feature types do not present much in the way of obvious semantic connections with the category "calm" (e.g., "satan"!). However, some might argue that word repetition can have a calming effect, and if this is the case, then the text stylistics features do appear to be picking up on the notion of repetition as a mechanism for instilling calmness or serenity.

6. CONCLUSIONS AND FUTURE WORK

This paper builds upon and extends our previous work on multi-modal mood classification by examining in-depth those feature types that have shown statistically significant improvements in correctly classifying individual mood categories. While derived from user-generated tags found on last.fm, the 18 mood categories used in this study fit well with Russell's mood model which is commonly used in MIR mood classification research. From our 18 mood categories we uncovered seven divergent categories where certain lyric feature types significantly outperformed audio and only one category where audio outperformed all lyric-based features. For those seven categories where lyrics performed better than audio, the top-ranked words clearly show strong and obvious semantic connections to the categories. In two cases, simple text stylistics provided significant advantages over audio. In the one case where audio outperformed lyrics, no obvious semantic connections between terms and the category could be discerned.

We note as worthy of future study the observation that no lyric-based feature provided significant improvements in the bottom-left (negative valence, negative arousal) quadrant (Figure 2) while audio features were able to do so (i.e., "calm"). This work is limited to audio spectral features and thus we also plan on extending this work by considering other types of audio features such as rhythmic and harmonic features.

7. ACKNOWLEDGEMENT

We thank The Andrew Mellon Foundation for their financial support.

8. REFERENCES

[1] J. S. Downie and S. J. Cunningham: "Toward a Theory of Music Information Retrieval Queries: System Design Implications." In *Proceedings of the 1st International Conference on Music Information Retrieval (ISMIR'02).*

[2] C. Laurier, J. Grivolla and P. Herrera: "Multimodal Music Mood Classification Using Audio and Lyrics," In *Proceedings of the International Conference on Machine Learning and Applications*, 2008.

[3] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. H. Chen: "Toward multi-modal music emotion classification," In *Proceedings of Pacific Rim Conference on Multimedia (PCM'08)*.

[4] X. Hu and J. S. Downie: "Improving mood classification in music digital libraries by combining lyrics and audio," In *Proceedings of Joint Conference on Digital Libraries, (JCDL2010).*

[5] L. Lu, D. Liu, and H. Zhang: "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1): 5-18, 2006.

[6] T. Pohle, E. Pampalk, and G. Widmer: "Evaluation of Frequently Used Audio Features for Classification of Music into Perceptual Categories," In *Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing*, 2005.

[7] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas: "Multi-Label Classification of Music into Emotions," In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*.

[8] J. S. Downie: "The Music Information Retrieval Evaluation Exchange (2005-2007): A Window into Music Information Retrieval Research," *Acoustical Science and Technology* 29 (4): 247-255, 2008. Available at: http://dx.doi.org/10.1250/ast.29.247

[9] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. Ehmann: "The 2007 MIREX Audio Music Classification Task: Lessons Learned," *Proceedings of the International Conference on Music Information Retrieval (ISMIR'08)*.

[10] H. He, J. Jin, Y. Xiong, B. Chen, W. Sun, and L. Zhao: "Language Feature Mining for Music Emotion Classification via Supervised Learning From Lyrics," In *Proceedings of Advances in the 3rd International Symposium on Computation and Intelligence (ISICA'08).*

[11] Y. Hu, X. Chen, and D. Yang: "Lyric-Based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method," In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR'09)*.

[12] M. M. Bradley and P. J. Lang: "Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings," Technical report C-1. University of Florida, 1999.

[13] D. Yang, and W. Lee: "Disambiguating Music Emotion Using Software Agents," In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*.

[14] P. J. Stone: *General Inquirer: a Computer Approach to Content Analysis.* Cambridge: M.I.T. Press, 1966.

[15] C. Strapparava and A. Valitutti: "WordNet-Affect: an Affective Extension of WordNet," In *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1083-1086, 2004.

[16] R. Mihalcea and H. Liu: "A Corpus-based Approach to Finding Happiness," In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW'06).*

[17] X. Hu and J. S. Downie: "Stylistics in Customer Reviews of Cultural Objects," In *Proceedings of the 2nd SIGIR Stylistics for Text Retrieval Workshop*, pp.37-42. 2006.

[18] B. Yu: "An Evaluation of Text Classification Methods for Literary Study," *Literary and Linguistic Computing*, 23(3): 327-343, 2008.

[19] J. A. Russell: "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, 39: 1161-1178, 1980.

[20] C. Fellbaum: *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[21] G. Tzanetakis: "Marsyas Submissions to MIREX 2007", available at http://www.music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_tzanetakis.pdf