# LYRIC TEXT MINING IN MUSIC MOOD CLASSIFICATION

**Xiao Hu** **J. Stephen Downie** **Andreas F. Ehmann**

**International Music Information Retrieval Systems Evaluation Laboratory**
**University of Illinois at Urbana-Champaign**

xiaohu@illinois.edu jdownie@illinois.edu aehmann@illinois.edu

## ABSTRACT

This research examines the role lyric text can play in improving audio music mood classification. A new method is proposed to build a large ground truth set of 5,585 songs and 18 mood categories based on social tags so as to reflect a realistic, user-centered perspective. A relatively complete set of lyric features and representation models were investigated. The best performing lyric feature set was also compared to a leading audio-based system. In combining lyric and audio sources, hybrid feature sets built with three different feature selection methods were also examined. The results show patterns at odds with findings in previous studies: audio features do not always outperform lyrics features, and combining lyrics and audio features can improve performance in many mood categories, but not all of them.

## 1. INTRODUCTION

There is a growing interest in developing and evaluating Music Information Retrieval (MIR) systems that can provide automated access to the mood dimension of music. Twenty-two systems have been evaluated between 2007 and 2008[1] in the Audio Mood Classification (AMC) task of the Music Information Retrieval Evaluation eXchange (MIREX). However, during these evaluations, several important issues have emerged and resolving these issues will greatly facilitate further progress on this topic.

### 1.1 Difficulties Creating Ground Truth Data

Due to the inherent subjectivity of music perception, there are no generally accepted standard mood categories. Music psychologists have created many different mood models but these have been criticized for missing the social context of music listening [1]. Some MIR researchers have exploited professionally assigned mood labels (e.g. AMG, MoodLogic[2]) [2,3], but none of these taxonomies has gained general acceptance. Professionally created labels have been criticized for not capturing the users' perspectives on mood.



To date, the AMC dataset is the only ground truth set that has been used to evaluate mood classification systems developed by multiple labs. However, this dataset contains only 600 30 sec. song clips. In fact, reported experiments are seldom evaluated against datasets of more than 1,000 music pieces. The subjective nature of music makes it very difficult to achieve cross assessor agreements on music mood labels. A post-hoc analysis of the 2007 AMC task revealed discrepancies among human judgments on about 30% of the audio excerpts [4]. To overcome the limitation, one could recruit more assessors to assess more candidate tracks. Unfortunately this would require too much human labor to be realistic for most projects. Thus, it is clear that a scalable and efficient method is sorely needed for building ground truth sets for music mood classification experimentation and evaluation.

### 1.2 Need for Multimodal Mood Classification

The seminal work of Aucouturier and Pachet [5] revealed a "glass ceiling" in spectral-based MIR, due to the fact that many high-level (e.g., semantic) music features simply are not discernable using spectral-only techniques. Thus, researchers started to supplement audio with lyrics and have reported improvements in such tasks as genre classification and artist identification [6,[7]. However, very few studies have combined audio and text for music mood classification [8], and their limitations (see below) call for more studies to investigate whether and how lyrics might help improve classification performance.

### 1.3 Related Work

Hu et al. [11] derived a set of three primitive mood categories using social tags on last.fm. They collected social tags of single adjective words on a publicly available audio dataset, USPOP [12], and manually selected 19 mood related terms of the highest popularity which then reduced to three latent mood categories using multidimensional scaling. This set was not adopted by others because three categories were seen as a domain oversimplification.

Yang and Lee [8] performed early work on supplementing audio mood classification with lyric text analysis. They combined a lyric bag-of-words (BOW) approach with 182 psychological features proposed in the General Inquirer [13] to disambiguate categories that audio-based classifiers found confusing and the overall classification accuracy was improved by 2.1%. However, their dataset was too small (145 songs) to draw any reliable conclusions. Laurier et al. [9] also combined audio

[1] http://www.music-ir.org/mirex/2007/index.php/AMC

[2] http://en.wikipedia.org/wiki/MoodLogic

and BOW lyric features. They conducted binary classification experiments on 1,000 songs in four categories and experimental results showed that audio + lyrics combined features improved classification accuracies in all four categories. Yang et al. [10] evaluated both unigram and bigram BOW lyric features as well as three methods for fusing lyric and audio sources on 1,240 songs in four categories. In these studies, the set of four mood categories was most likely oversimplified, the datasets were relatively small and the lyric text features, namely BOW in tf-idf representation, were very limited.

In this paper, we describe a novel method of building a large-scale ground truth dataset with 5,585 songs in 18 mood categories. We then report experiments on a relatively complete set of lyric text features, including function words, POS features and the effect of stemming. Finally, we examine the impact of lyric features on music mood classification by comparing 1) lyric features; 2) audio features; 3) hybrid (lyric + audio features without feature selection; and, 4) hybrid features generated by three feature selection methods.

## 2. BUILDING A GROUND TRUTH SET

### 2.1 Data Collection

We began with an in-house collection of about 21,000 audio tracks. Social tags on these songs were then collected from last.fm. 12,066 of the pieces had at least one last.fm tag. Simultaneously, song lyrics were gathered from online lyrics databases. Lyricwiki.org was the major resource because of its broad coverage and standardized format. To ensure data quality, our crawlers used song title, artist and album information to identify the correct lyrics. In total, 8,839 songs had both tags and lyrics. A language identification program[3] was then run against the lyrics, and 55 songs were identified and manually confirmed as non-English, leaving lyrics for 8,784 songs. Table 1 presents the composition of the collection.

| Collection | Avg. length (sec.) | Unique | Have tags | Have English Lyrics |
|---|---|---|---|---|
| USPOP | 253.6 | 8,271 | 7,301 | 6,948 |
| USCRAP | 243.5 | 2,553 | 456 | 237 |
| American music | 183.2 | 5,049 | 2,209 | 790 |
| Metal music | 311.8 | 105 | 105 | 104 |
| Beatles | 163.8 | 163 | 162 | 161 |
| Magnatune | 253.9 | 4,204 | 1,261 | 19 |
| Assorted pop | 233.8 | 600 | 572 | 525 |
| Total | (Avg.) 234.8 | 20,945 | 12,066 | 8,784 |

**Table 1.** Descriptions and statistics of the collection.

### 2.2 Identifying Mood Categories

Social tag data are noisy. We employed a linguistic resource, WordNet-Affect [14], to filter out junk tags and tags with little or no affective meanings. WordNet-Affect is an extension of WordNet where affective labels are assigned to concepts representing emotions, moods, or emotional responses. There were 1,586 unique words in the latest version of WordNet-Affect and 348 of them exactly matched the 61,849 unique tags collected from last.fm. However, these 348 words were not all mood related in the music domain. We turned to human expertise to clean up these words. Two human experts were consulted for this project. Both are MIR researchers with a music background and native English speakers. They first identified and removed tags with music meanings that did not involve an affective aspect (e.g., "trance" and "beat"). Second, judgmental tags such as "bad", "poor", "good" and "great" were removed. Third, some words have ambiguous meanings and there was not enough information to determine the intentions of the users when they applied the tags. For example, does "love" mean the song is about love or the user loves the song? To ensure the quality of the labels, these ambiguous words were removed. 186 words remained and 4,197 songs were tagged with at least one of the words.

Not all the 186 words represent distinguishable meanings. In fact, many of them are synonyms and should be grouped together [3]. WordNet is a natural resource for synonym identification, because it organizes words into *synsets*. Words in a synset are synonyms from the linguistic point of view. WordNet-Affect goes one step further by linking each non-noun synset (verb, adjective and adverb) with the noun synset from which it is derived. For instance, the synset of "sorrowful" is marked as derived from the synset of "sorrow". Hence, for the 186 words, those belonging to and being derived from the same synset in WordNet-Affect were grouped together. As a result, the tags were merged into 49 groups.

Several tag groups were further merged if they were deemed musically similar by the experts. For instance, the group of ("cheer up", "cheerful") was merged with ("jolly", "rejoice"); ("melancholic", "melancholy") was merged with ("sad", "sadness"). This resulted in 34 tag groups, each representing a mood category for this dataset. Using the linguistic resources allowed this process to proceed quickly and minimized the workload of the human experts.

For the classification experiments, each category should have enough samples to build classification models. Thus, categories with fewer than 20 songs were dropped resulting in 18 mood categories containing 135 tags. These categories and their member tags were then validated for reasonableness by a number of native English speakers. Table 2 lists the categories, a subset of their member tags and number of songs in each category (after the filtering step described below)[4].

### 2.3 Selecting the Songs

A song was not selected for a category if its title or artist contained the same terms within that category. For example, all but six songs tagged with "disturbed" were songs by the artist "Disturbed." In this case, the taggers may simply have used the tag to restate the artist instead of describing the mood of the song. In order to ensure enough data for lyric-based experiments, we only selected those songs with lyrics whose word count was greater than 100 (after unfolding repetitions as explained

[3] http://search.cpan.org/search%3fmodule=Lingua::Ident

[4] Due to space limit, the complete tag list can be found at http://www.music-ir.org/archive/figs/18moodcat.htm

in Section 3.2). After these filtering criteria were applied, we were left with 2,829 unique songs.

Multi-label classification is relatively new in MIR, but in the mood dimension, it is more realistic than single-label classification: A music piece may be "happy and calm" or "aggressive and depressed," etc. This is evident in our dataset as we have many songs that are members of more than one mood category. Table 3 shows the distribution of songs belonging to multiple categories. We adopted a binary classification approach for each of the 18 mood categories, and the 2,829 songs formed the positive example set.

| Categories | # of tags | #of songs |
|---|---|---|
| calm, comfort, quiet, serene, mellow, chill out,… | 25 | 1,394 |
| sad, sadness, unhappy, melancholic, melancholy | 8 | 916 |
| happy, happiness, happy songs, happy music, … | 6 | 472 |
| romantic, romantic music | 2 | 447 |
| upbeat, gleeful, high spirits, zest, enthusiastic, … | 8 | 321 |
| depressed, blue, dark, depressive, dreary, | 11 | 288 |
| anger, angry, choleric, fury, outraged, rage, … | 7 | 156 |
| grief, heartbreak, mournful, sorrow, sorry, … | 14 | 112 |
| dreamy | 1 | 85 |
| cheerful, cheer up, festive, jolly, jovial, merry, … | 13 | 76 |
| brooding, contemplative, meditative, reflective, … | 8 | 69 |
| aggression, aggressive | 2 | 53 |
| confident, encouraging, encouragement, optimism | 5 | 43 |
| angst, anxiety, anxious, jumpy, nervous, angsty | 6 | 36 |
| earnest, heartfelt | 2 | 34 |
| desire, hope, hopeful, mood: hopeful | 4 | 28 |
| pessimism, cynical, pessimistic, weltschmerz,… | 5 | 27 |
| excitement, exciting, exhilarating, thrill, ardor,… | 8 | 20 |
| TOTAL | 135 | 4,578 |

**Table 2.** Mood categories and song distributions.

| # of categories | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| # of songs | 1,625 | 788 | 305 | 91 | 17 | 2 |

**Table 3.** Distribution of songs with multiple labels.

In a binary classification task, each category needs negative samples as well. To create our negative sample set for a given category, we chose songs that were not tagged with any of the terms found within that category but are heavily tagged with many other terms. Since there were plenty of negative samples for each category, we randomly selected songs tagged with at least 15 other terms including mood terms in other categories. Hence, some negative samples of one category are positive samples of another category. In order to make samples of various categories as diverse as possible, we set a constraint that no negative samples were members of more than one category. Similar to positive samples, all negative samples have at least 100 words in their unfolded lyric transcripts. We balanced equally the positive and negative set sizes for each category. Our final dataset comprised 5,585 unique songs.

## 3. EXPERIMENTS

### 3.1 Evaluation Measures and Classifiers

This study uses classification accuracy as the performance measure. For each category, accuracy was averaged over a 10-fold cross validation. For each feature set, the accuracies across categories were averaged in a macro manner, giving equal importance to all categories regardless of the size of the categories. To determine if performances differed significantly, we chose the non-parametric Friedman's ANOVA test because the accuracy data are rarely normally distributed.

Support Vector Machines (SVM) were chosen as our classifier because of their strong performances in text categorization and MIR tasks. We used the LIBSVM [15] implementation of SVM and chose a linear kernel as trial runs with polynomial kernels did not yield better results. Parameters were tuned using the grid search tool in LIBSVM, and the default parameters performed best for most cases. Thus, the default parameters were used for all the experiments.

### 3.2 Lyric Preprocessing

Lyric text has unique structures and characteristics requiring special preprocessing techniques. First, most lyrics consist of such sections as *intro*, *interlude*, *verse*, *prechorus*, *chorus* and *outro*, many with annotations on these segments. Second, repetitions of words and sections are extremely common. However, very few available lyric texts were found as verbatim transcripts. Instead, repetitions were annotated as instructions like *[repeat chorus 2x]*, *(x5)*, etc. Third, many lyrics contain notes about the song (e.g., *"written by"*), instrumentation (e.g., *"(SOLO PIANO),"* and/or the performing artists. In building a preprocessing program that took these characteristics into consideration, we manually identified about 50 repetition patterns and 25 annotation patterns. The program converted repetition instructions into the actual repeated segments for the indicated number of times while recognizing and removing other annotations.

### 3.3 Lyrics Features

Lyrics are a very rich resource and many types of textual features can be extracted from them. This work compares some of the feature types most commonly used in related text classification tasks.

#### *3.3.1 Bag-of-Words (BOW)*

Bag-of-words (BOW) are collections of unordered words. Each word is assigned a value that can represent, among others, the frequency of the word, tf-idf weight, normalized frequency or a Boolean value indicating presence or absence. Among these variations, tf-idf weighting is the most widely used in text analysis and MIR, but some studies in text sentiment analysis also reported other representations outperformed tf-idf weighting [16]. These four representations were compared in our experiments.

Selecting the set of words to comprise the BOW set is an important consideration. Stemming is a process of merging words with the same morphological roots, and

has shown mixed effects in text classification. Thus, we experimented with both options. We used the Snowball stemmer[5] supplemented with irregular nouns and verbs[6] as this stemmer cannot handle irregular words. Function words (see below) were removed for both the stemming and not stemming cases.

#### *3.3.2 Part-of-Speech (POS)*

Part-of-Speech (POS) is a popular feature type in text sentiment analysis [17] and text style analysis [18]. Other MIR studies on lyrics have also used POS features [6,19]. We used the Stanford POS tagger[7] which tags each word with one of 36 unique POS tags.

#### *3.3.3 Function Words*

Function words (e.g. *the*, *a*, etc.) carry little meaning. However, function words have been shown to be effective in text style analysis [18]. To evaluate the usefulness of function words in mood classification, the same list of 435 function words found in [18] were used as an independent feature set.

### 3.4 Audio Processing and Features

Studies in other MIR tasks have generally found lyrics alone are not as informative as audio [6,7]. To find out whether this is true in music mood classification, our best performing lyrics feature set was compared to Marsyas[8], the best performing audio system evaluated in the MIREX 2007 AMC task. Marsyas uses 63 spectral features: means and variances of Spectral Centroid, Rolloff, Flux, Mel-Frequency Cepstral Coefficients, etc. It also uses LIBSVM with a linear kernel for classification. Every audio track in the dataset was converted to 44.1KHz stereo .wav files and fed into Marsyas. The extracted spectral features were subsequently processed by SVM classifiers.

### 3.5 Hybrid Features and Feature Selection

Previous MIR studies suggest that combining lyric and audio features improves classification performance. Thus, we concatenated our best performing lyrics features and the spectral features to see whether and how much the hybrid features could improve classification accuracies.

In text categorization with BOW features, the dimensionality of document vectors is usually high. Thus, feature selection is often used for the sake of good generalizability and efficient computation. In this study, we compared three methods in selecting the most salient lyric features:

1. Select features with high F-scores. F-score measures the discrimination power of a feature between two sets [20]. The higher a feature's F-score is, the more likely it is to be discriminative. F-score is a generic feature reduction technique independent of classification task and method.

2. Select features using language model differences (LMD) proposed in [9], where the top 100 terms with largest LMD were combined with audio features and showed improved classification accuracies. We wish to find out if this method works in this study with more categories.

3. Select features based on the SVM itself. A trained decision function in a linear SVM contains weights for each feature indicating the relevance of the feature to the classifier. [16] has shown that trimming off features with lower weights improved SVM performance in literature sentimentalism classification. This study investigates if it works for music mood classification.

## 4. RESULTS

### 4.1 Best Lyrics Features

Table 4 shows the average accuracies across all 18 categories for the considered lyrics features and representations.

| Representation | Boolean | term frequency (tf) | normalized tf | tf-idf weighting |
|---|---|---|---|---|
| BOW-Stemming | 0.5748 | 0.5819 | 0.5796 | **0.6043** |
| BOW-Not Stemming | 0.5817 | 0.5829 | 0.5840 | 0.5923 |
| POS | 0.5277 | 0.5768 | 0.5691 | 0.5571 |
| Function Words | 0.5653 | 0.5733 | 0.5692 | 0.5723 |

**Table 4.** Average accuracies for lyric features.

The best text feature type is BOW with stemming and tf-idf weighting (BSTI). The difference between stemming options is not significant at $p < 0.05$. The four representations of BOW features do not differ significantly in average performances.

For POS features, the Boolean representation is not as good as others. This is not unexpected because presumably, most lyrics would contain most POS types. In general, POS features and function words are not as good as BOW features. This confirms the heuristic that content words are more useful for mood classification.

### 4.2 Combining Audio and All Text Features

Three feature sets were compared: spectral features, BSTI, and direct concatenation of both. Their accuracies are shown as part of Table 5. Although their difference is not significant (at $p < 0.05$) on average, BSTI was significantly better than spectral features in these five categories: *romantic*, *grief*, *aggression*, *angst*, and *exciting*. This observation is different from findings in [9] where lyrics features alone did not outperformed audio features in any category.

The accuracies in individual categories are shown in Figure 1 where categories are ordered by decreasing number of samples.

[5] http://snowball.tartarus.org/

[6] The irregular verb list was obtained from http://www.englishpage.com/irregularverbs/irregularverbs.html, and the irregular noun list was obtained from http://www.esldesk.com/esl-quizzes/irregular-nouns/irregular-nouns.htm

[7] http://nlp.stanford.edu/software/tagger.shtml

[8] http://www.music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_tzanetakis.pdf

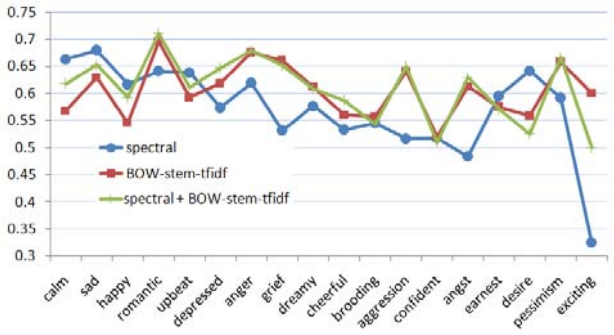


**Figure 1**. Accuracies of three systems in all categories.

As shown in Figure 1, system performances on different categories vary greatly, and no feature set performs best for all categories. It appears that spectral features are better for larger sized categories, while lyric features are better for middle sized categories. Data sparseness may be less an issue for text features because the samples were chosen to have a certain length of lyrics. From a semantic point of view, the categories where spectral features are significantly better than text features ("upbeat", "happy" and "calm") may have typical auditory characteristics that can be captured by audio spectral features. On the other hand, there may be certain lyrics words that connect well to the semantics of some categories like "grief", "romantic" and "anger." Thus, lyrics features possibly work better because of this connection. For example, the following stemmed words are ranked high for these categories by all of the three aforementioned feature selection methods:

*grief:* singl, scare, confus, heart, cri, sorry, lone, oooh,…
*romantic:* endless, love, promis, ador, whisper, lady,…
*anger:* fuck, man, dead, thumb, girl, bitch, kill,…

It is also clear from Figure 1 that the performances of the combined feature set closely follow the trend of the lyrics-only features. This is probably inevitable given the fact that there are several orders of magnitude more lyric features than spectral features in the combined set. This also demonstrates the necessity of feature selection.

We note that there is a general trend in terms of average accuracy decreasing with smaller sample sizes, sometimes even achieving lower than baseline (50%) performance. These cases also show the highest variance in terms of accuracies across folds. This is a somewhat expected result as the lengths of the feature vectors far outweigh the number of training instances. Therefore, it is difficult to make broad generalizations about these extremely sparsely represented mood categories.

### 4.3 Combining Audio and Selected Text Features

Using each of the three feature selection methods, we selected the top *n* BSTI features and combined them with the 63 spectral features. We first varied *n* from 63 to 500 (63, 100, 200,…, 500) for all categories. Since the number of features varies per category, we also varied *n* based on the number of features available in each category, from 10% to 90%. The results show that the best *n* varies across the three feature selection methods. Table 5 shows accuracies of the feature sets with the best average performances among each feature selection method.

| Category | Spec + F-score n=80% | Spec + LMD n=63 | Spec + SVM n=70% | Spec + BSTI | BSTI | Spectral |
|---|---|---|---|---|---|---|
| calm | *0.6112* | **0.6664** | *0.6054* | *0.6176* | *0.5674* | 0.6635 |
| sad | *0.6496* | **0.6976** | 0.6573 | 0.6524 | *0.6295* | 0.6796 |
| happy | 0.5965 | 0.6147 | 0.5784 | 0.5922 | *0.5455* | **0.6168** |
| romantic | *0.7014* | *0.7124* | ***0.7127*** | *0.7104* | *0.6959* | 0.6407 |
| upbeat | 0.6232 | *0.6075* | 0.6013 | 0.6107 | *0.5920* | **0.6389** |
| depressed | 0.6318 | 0.6448 | ***0.6613*** | 0.6475 | 0.6183 | 0.5741 |
| anger | 0.6692 | **0.6827** | 0.6721 | 0.6787 | 0.6754 | 0.6194 |
| grief | *0.6477* | *0.6386* | *0.6511* | *0.6511* | ***0.6610*** | 0.5314 |
| dreamy | **0.6396** | 0.6326 | 0.6354 | 0.6083 | 0.6118 | 0.5771 |
| cheerful | 0.5661 | 0.5732 | **0.5929** | 0.5866 | 0.5598 | 0.5330 |
| brooding | 0.5583 | 0.5071 | **0.5726** | 0.5440 | 0.5571 | 0.5452 |
| aggression | ***0.6683*** | 0.5667 | *0.6300* | *0.6500* | *0.6400* | 0.5167 |
| confident | 0.6417 | **0.7208** | 0.5050 | 0.5100 | 0.5200 | 0.5175 |
| angst | *0.4750* | *0.5875* | ***0.6292*** | ***0.6292*** | *0.6125* | 0.4833 |
| earnest | 0.5667 | **0.6250** | 0.5833 | 0.5708 | 0.5750 | 0.5958 |
| desire | 0.5083 | *0.4250* | 0.5417 | 0.5250 | 0.5583 | **0.6417** |
| pessimism | **0.6833** | 0.5333 | 0.6667 | 0.6667 | 0.6583 | 0.5917 |
| exciting | 0.5750 | 0.4250 | 0.5750 | 0.5000 | ***0.6000*** | 0.3250 |
| AVERAGE | 0.6118 | 0.6033 | **0.6151** | 0.6084 | 0.6043 | 0.5717 |

**Table 5.** Accuracies of feature sets for individual categories. (bold font denotes the best for that category, italic indicates significant difference from spectral features at $p < 0.05$.)

The results show that not all categories can be improved by combining lyric features with spectral features. Audio-only and lyric-only features outperform *all* combined feature sets in five of the 18 categories. Each of the combined feature sets outperforms lyric and audio features in at most nine categories. This is different from findings in previous studies [8,9] where combined features were best for all experimented categories.

In particular, the language model difference method with 63 lyric features (Spec + LMD n = 63) shows an interesting pattern: it improves accuracies in six of the 12 categories where lyric features outperform spectral features and three of the six categories where spectral features beat lyric features. This indicates that, with the same dimensionality, lyrics and audio have indeed a similar impact on combined features.

In combining lyrics and audio features, feature selection often yields better results because many text features are either redundant or noisy. In terms of average accuracies, features selected by SVM models work slightly better for SVM classifiers than the other two feature selection methods. However, it is interesting to see that (Spec + LMD n = 63) outperforms lyric and audio features in nine categories which are the most among all combined feature sets. It also outperforms all others in five mood categories and achieves significantly better results than spectral features in three other mood categories. Similar patterns are observed for the F-score method with 63 lyric features. This suggests that in hybrid feature sets, lyric features can be and should be aggressively reduced.

## 5. CONCLUSIONS AND FUTURE WORK

This paper investigates the usefulness of text features in music mood classification on 18 mood categories derived from user tags. Compared to Part-of-Speech and function

words, Bag-of-Words are still the most useful feature type. However, there is no significant difference between the choice of stemming or not stemming, or among the four text representations (e.g. tf-idf, Boolean, etc) on average accuracies across all categories.

Our comparisons of lyric, audio and combined features discover patterns at odds with previous studies. In particular lyric features alone *can* outperform audio features in categories where samples are more sparse or when semantic meanings taken from lyrics tie well to the mood category. Also, combining lyrics and audio features improves performances on most, but not all, categories. Experiments on three different feature selection methods demonstrated that too many text features are indeed redundant or noisy and combining audio with the most salient text features may lead to higher accuracies for most mood categories.

Future work includes investigation of other text features, such as text statistics and affective words provided by domain lexicons. It would also be interesting to take a close look at individual categories and find out why lyrics features do or do not help. Moreover, more sophisticated feature and model combination techniques besides naïve feature vector concatenation are worth investigating.

## 6. ACKNOWLEDGEMENT

We thank the Andrew W. Mellon Foundation for their financial support.

## 7. REFERENCES

[1] P. N. Juslin and P. Laukka: "Expression, Perception, and Induction of Musical Emotions: A Review and A Questionnaire Study of Everyday Listening," *Journal of New Music Research*, Vol. 33, No. 3, pp. 217-238, 2004.

[2] M. Mandel, G. Poliner, and D. Ellis: "Support Vector Machine Active Learning for Music Retrieval," *Multimedia Systems*, Vol. 12, No. 1, pp. 3-13, 2006.

[3] X. Hu and J. S. Downie: "Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata," *Proceedings of the International Conference on Music Information Retrieval*, 2007.

[4] 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*, 2008.

[5] J-J. Aucouturier and F. Pachet: "Improving Timbre Similarity: How High Is the Sky?" *Journal of Negative. Results in Speech and Audio Sciences,* Vol.1, No.1, 2004.

[6] T. Li and M. Ogihara: "Semi-Supervised Learning from Different Information Sources," *Knowledge and Information Systems*, Vol.7, No.3, pp.289-309, 2004

[7] R. Neumayer and A. Rauber: "Integration of Text and Audio Features for Genre Classification in Music Information Retrieval," *Proceedings of the European Conference on Information Retrieval,* pp.724 – 727, 2007.

[8] D. Yang, and W. Lee: "Disambiguating music Emotion Using Software Agents," In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04),* 2004.

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

[10] 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", Proceedings of Pacific Rim Conference on Multimedia (PCM'08), 2008.

[11] X. Hu, M. Bay, and J. S. Downie: "Creating a Simplified Music Mood Classification Groundtruth Set," P*roceedings of the 8th International Conference on Music Information Retrieval,* 2007.

[12] D. Ellis, A. Berenzweig, and B. Whitman: "The USPOP2002 Pop Music Data Set," Retrieved from http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html, 2003.

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

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

[15] C. Chang and C. Lin: *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[16] B. Yu: "An Evaluation of Text Classification Methods for Literary Study," *Literary and Linguistic Computing* Vol. 23, No. 3, pp. 327-343, 2008.

[17] B. Pang and L. Lee: "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval,* Vol.2 No.1-2, pp. 1–135, 2008

[18] S. Argamon, M. Saric and S. S. Stein: "Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results," *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining,* pp. 475-480, 2003.

[19] R. Mayer, R. Neumayer, and A. Rauber: "Rhyme and Style Features for Musical Genre Categorisation by Song Lyrics," *Proceedings of the International Conference on Music Information Retrieval,* 2008.

[20] Y.-W. Chen and C.-J. Lin: "Combining SVMs with Various Feature Selection Strategies," In *Feature Extraction, Foundations and Applications*, Springer, 2006.