Music Subject Classification Based on Lyrics and User Interpretations

Kahyun Choi
School of Information Sciences
University of Illinois
ckahyu2@illinois.edu

Jin Ha Lee
Information School
University of Washington
jinhalee@uw.edu

Xiao Hu
Faculty of Education
University of Hong Kong
xiaoxhu@hku.hk

J. Stephen Downie
School of Information Sciences
University of Illinois
jdownie@illinois.edu

ABSTRACT

That music seekers consider song subject metadata to be helpful in their searching/browsing experience has been noted in prior published research. In an effort to develop a subject-based tagging system, we explored the creation of automatically generated song subject classifications. Our classifications were derived from two different sources of song-related text: 1) lyrics; and 2) user interpretations of lyrics collected from songmeanings.com. While both sources contain subject-related information, we found that user-generated interpretations always outperformed lyrics in terms of classification accuracy. This suggests that user interpretations are more useful in the subject classification task than lyrics because the semantically ambiguous poetic nature of lyrics tends to confuse classifiers. An examination of top-ranked terms and confusion matrices supported our contention that users' interpretations work better for detecting the meaning of songs than what is conveyed through lyrics.

Keywords

Music Subject Classification, Music Digital Library, Interpretations of Lyrics, Subject Metadata

INTRODUCTION

Subject metadata, in addition to basic bibliographic metadata, serve an important role for users searching or browsing information in music digital libraries. Music listeners often ponder about the meaning of their favorite songs and spend much time discussing what the songs are about with other music listeners online and offline. Prior user studies such as Lee and Downie (2004) suggest that people consider the subject as potentially useful metadata; in an online survey of 427 music listeners, 33.4% and 17.9% of the respondents indicated that they would use the theme/main subject and storyline of music for searching/browsing music, if they were available. The proportion of respondents who thought theme/main subject was useful was in fact higher than metadata like popularity, mood, time period, or instrument. Another study by Bainbridge et al. (2003) which analyzed 502 music-related postings made in Google Answers also identified 'lyric story' as one of the ten categories of different metadata people used in describing their music information needs. These studies suggest that annotating subject metadata has the potential to be useful for navigating music collections in digital libraries.

In the music information retrieval and music digital libraries domains, determining and representing what a song is about--in other words, its subject--has always been a challenge (Byrd & Crawford, 2002; Kim & Belkin, 2002; Lee & Downie, 2004; McLane, 1996). The massive amount of digital music and the high cost of human subject annotation call for the development of automatic music subject classification methods. To date, lyrics and tags have been used as sources for feature vectors for music classification (Bischoff et al., 2009; Hu et al., 2005; Kleedorfer et al., 2008). However, although lyrics do provide some basis for identifying the subject of music, they tend to share similar characteristics with poetry such as common use of figurative language and metaphor (Singhi & Brown, 2014). As a result, it can be difficult to identify the subject of music pieces by analyzing terms in lyrics only.

Beyond lyrics and tags, user-generated information such as music reviews or users' interpretation of songs have received relatively little attention. Such information may be a good resource for identifying the subject of the songs because users' interpretations can often reveal deeper meanings of songs and/or what the artists intended to convey in addition to words literally represented by lyrics (Zavalina et al., 2008). Furthermore, users' interpretations may contain more information about the song than lyrics because the latter typically have limited length.

In this paper, we explore the potential value of users' interpretation for identifying the subject of music. In order to evaluate its effectiveness, we compare music subject

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

classification results based on users' interpretation to a previously tested textual source, lyrics. We aim to answer the following research questions:

Q1. Between interpretations and lyrics, which source is more useful for subject classification?

Q2. How different are the best features of interpretations and lyrics?

Q3. Does combining the user interpretations and lyrics improve the classification results?

Q4. Which ensemble method is more effective when combining the two sources?

BACKGROUND AND RELATED WORK

This section first defines the concept of music subject. Then, we cover relevant prior work on text-based music classification systems in general. Finally, we introduce our previous work on music subject classification and summarize the contributions of this paper.

Defining Subject

The term "subject" in the context of music is defined as "a theme (or group of themes) on which a composition is based", according to Grove Music Online (Walker, 2016). In prior literature, several terms, in addition to "subject," have been used interchangeably to refer to what the song is about: "topic," "theme," and "aboutness." The usage of these terms varies across different studies. For example, Bischoff et al. (2009) uses "themes" that are taken from allmusic.com, referring to particular events/occasions where the music can be used, or mood of the music (e.g., "party music," "wedding songs," "mellow music"). Kleedorfer et al. (2008) uses the term "topic" to refer to various topic spaces derived by using topic modeling algorithm on song lyrics. "Theme" in Lee et al. (2004) is used to refer to what a song is about, similar to what we are calling "subject" in this study. In this paper, we will be following an operational definition of the term "subject" as "the topic or recurring theme in the song".

Prior Studies on Automatic Classification of Music based on Text Sources

This section reviews relevant works dealing with the automatic annotation of music, emphasizing works concerning extrinsic metadata such as mood, genre, usage, and subject, rather than music features (e.g. tempo, onset, key, and melody). We specifically focus on studies using various text sources including lyrics, tags, and reviews as their input instead of features from audio signal processing.

Bischoff et al. (2009) proposed a music classification system that predicts opinion, usage, genre, and style information based on social tags and lyrics. In their work, the classifier using a combination of tags and lyrics outperformed those using only tags or lyrics in theme and mood classification tasks. However, using lyrics in addition to tags reduced the performance in genre and style classification tasks. This indicates that the usefulness of text sources depends on the type of metadata predicted from them. Hence in our study, we also investigate the usefulness of multiple text sources and their combinations for predicting the subject information.

The work by Mahedero et al. (2005) is the most similar to ours in that they used lyrics for music subject classification, although our work includes more subject categories and a larger number of songs. Furthermore, we use not only lyrics but also previously unexplored user-generated data (i.e., interpretations). Hu and Downie (2010) compared various lyric features for music mood classification and combined texts and audio to improve the performance. In contrast, we compare two different text sources and combinations of them for music subject classification. Online music reviews are another type of potentially useful information source for music metadata enhancement. Hu et al. (2005) presented a genre prediction system based on reviews from epinions.com. However, we decided to use users' song interpretations from songmeanings.com rather than general reviews, as the former are more focused on subject matters whereas the latter may include opinions on the products (e.g., sound quality or CD cover images).

The Million Song Dataset (Bertin-Mahieux et al., 2012) contains social tags, another type of textual source, collected from Last.fm. Although user-generated tags have been actively used to enrich metadata in digital libraries in prior research, it was also found out that only a small number of music tags are related to what a song is about (Bischoff et al., 2008). Therefore, in this study, we decided to focus on user interpretations and lyrics rather than social tags.

Our Prior Work on Music Subject Classification

This is a follow-up study of our previous preliminary work reported in (Choi et al., 2014), where a simple music subject classification system was built using k-Nearest Neighbors (kNN) classifiers taking lyrics and their interpretations as the input. This work expands our previous effort in a number of ways. First, we compare four classifiers (SVMs with linear and RBF kernels, Naive Bayes and kNN) to identify the best classifier. As for the preprocessing step, for further analysis, we use a lemmatization technique instead of Porter stemming to obtain more accurate and readable lemmas based on morphological analysis (Manning et al., 2008). Moreover, we provide more detailed evaluation of the classification results, specifically the accuracies of each subject category as well as the average accuracies. In addition, we compare two ensemble methods, concatenation and late fusion when combining lyrics and interpretations. In this work, we also review the top 20 features of each category from each source to compare and contrast the nature of lyrics and interpretations. Finally, we examine four confusion matrices from each source to determine the degree of confusion between each pair of categories.

METHOD

Our research was conducted, following the steps which are summarized in Figure 1. First, we collected data including

lyrics, interpretations, and subject labels. Then, we preprocessed the data to extract features. We applied a proper ensemble method when both lyric and interpretation sources are used together. From those features, we built classification models using various approaches, and then finally evaluated the performance of different combinations of choices.

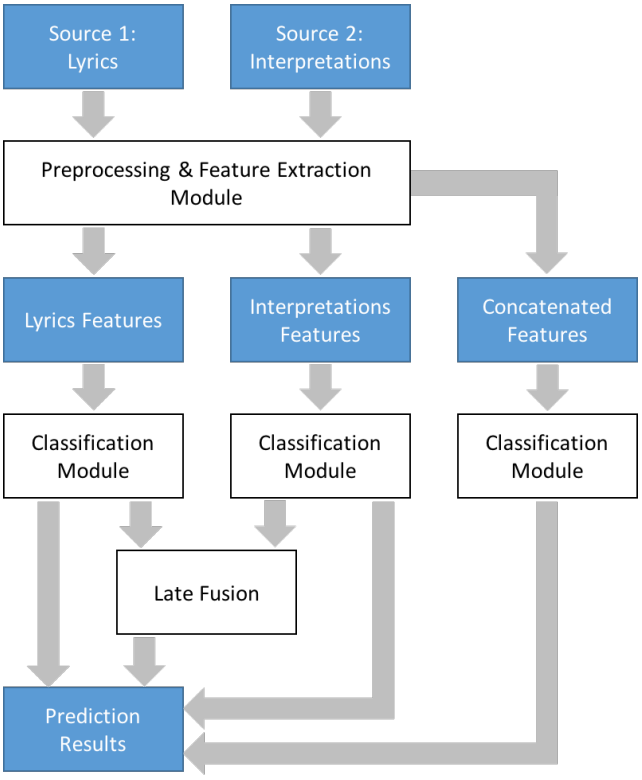

Figure 1. Research Framework

Dataset

Data for our experiments were obtained from two sources: songmeanings.com and songfacts.com. Songmeanings.com (hereinafter, Songmeanings) is an online community of music fans and enthusiasts who contribute lyrics and discuss their interpretations of song lyrics. From this website, users' interpretations as well as lyrics were collected for our experiment. Interpretations are user comments for each song provided in an open text form, similarly to those found on Amazon.com and imdb.com, except that the Songmeanings comments are not titled. Comments are rated by other users of the site which is used as a default sorting mechanism.

Songmeanings provides an API through which we can retrieve comments and their ratings by querying artist names and song titles. As for the list of songs, we relied on another website, Songfacts.com (hereinafter, Songfacts). Songfacts is "a searchable database of song information where you can find out the stories behind the songs, get the lyrics, and watch the videos." Although it provides 16 browsing options such as "inspired by", "used in", "about", etc., we focus on the "about" option, which corresponds to the *topic* of the song lyrics. In the "about" category, there were 123 subcategories, some of which are shown in Table 1. According to the description on the 'about us' section of the website, the subject of songs was determined based on interviews of the songwriters (when possible), as well as other sources including "books, magazines, newspaper articles, reference materials, and publicity releases." From this website, we were able to obtain the ground truth data (i.e., what we consider to be the most "accurate" subject of song) for our classification experiments.

To select the songs for our study, we first found the songs that appeared in both Songmeanings and Songfacts, but limiting the search to songs with at least one comment in Songmeanings. This identification was done with the title and the artist name, which are the only information that is always present in both websites. To make the identification process more reliable, we only included the exact string matches. During this search, we ensured that all songs in our dataset had attached interpretations. Because the match is rigorous and that this identification process keeps only the common songs that appear in both websites (e.g., in Songfacts, there are 256 songs about war, and of those songs, 127 have been discussed in Songmeanings), some subject categories only contained a few songs and their user comments.

Once the identification process was done, we selected the ten most populous categories out of the 123 subject categories, considering the balance of the distribution of songs across the categories. This was to avoid including all the categories with a small number of songs or user comments. This pruning may result in some limitation on the scale of this study. However, the setup of the study with the ten categories should be sufficient for achieving the main goal of comparing features from lyrics and their interpretations in terms of their expressive power in conveying the subject.

Of these ten categories, we removed the two classes that were deemed as conveying a mood (e.g., songs about heartache or loneliness) rather than a subject (e.g., songs about drugs or war). The exclusion is based on the goal of this work; we wanted to investigate the relatively unexplored area, music subject, rather than music mood, which is a fairly well explored concept in the field of music information retrieval (e.g., Kim et al., 2010; Hu & Downie, 2010; Hu et al., 2016). Although there is no widely accepted consensus about how to define music subjects, in this paper, we focus on the "topical" subject categories excluding the mood categories. Eventually we expect a comprehensive automatic music classification system that will be able to automatically assign metadata regardless of their type; however we leave this exploration to future work. We selected the following eight subject categories for our classification experiment (Table 1).

The second column in Table 1 shows how many songs were in Songfacts for each category, and the third column shows how many songs had user interpretations in Songmeanings for each category. In order to have a balanced dataset, we randomly chose 100 songs per category, thus resulting in the final dataset with 800 songs.

Subject Category	# of songs in Songfacts	# of songs in Songmeanings
Religion	316	181
Sex	286	169
Drugs	239	154
Parent	267	131
War	242	127
Places	232	127
ex-lover	203	115
Death	220	112

Table 1. Eight selected categories with the number of songs in Songfacts and Songmeanings.

Data Preprocessing and Feature Extraction

Data preprocessing and feature extraction were done in the following order:

- Tokenization: Interpretation and lyric sentences were broken down into words, respectively.
- Filtering: Only the words consisting of alphabetic letters were saved in order to filter punctuations, numbers, and user IDs.
- Lemmatization (Manning et al., 2014): Different grammatical variations of the same word roots were consolidated.
- Stopword elimination: In order to remove terms with especially high or low frequency of occurrences, we eliminated common stopwords in English and words that appeared fewer than three times.
- Name filtering: We filtered out proper nouns such as artist names using a named entity recognition technique (Manning et al., 2014) since proper nouns tend to be specific to the individual piece of music, rather than the subject category as a whole. We conducted two experiments with and without proper nouns. Both yielded almost the same classification accuracies, but representative features of each category in Table 5 and Table 6 were less noisy when proper nouns were eliminated. This is because in user interpretations, users often mention artists' names but they do not seem to serve as a proper feature for this subject classification job. Here, we report the cases where proper nouns were filtered out. However, we leave the question of how a certain artist is related to each subject category to be explored in future work.
- Term weightings: After deleting stopwords, the term frequencies in each song-specific document were normalized by dividing all the term counts by the maximum term frequency in the song. Term Frequency–Inverse Document Frequency (TFIDF) weighting was calculated for each word as it is one of the most effective term weighting schemes in information retrieval.

As a result, for the 800 songs in the dataset, we extracted 2,597 unique words in the lyric subset and 5,592 unique words in the interpretations subset as our features.

Ensemble Method

We also investigated whether combining lyrics and interpretations improves the classification performance or not. In order to integrate two different sources, two different ensemble methods were employed and compared.

Among many hybrid methods, we compared the straightforward concatenation method and the late fusion technique, since they were frequently used in the previous multimodal music classification research (Bischoff et al., 2009; Laurier et al., 2008; Mayer et al., 2008; Whitman & Smaragdis, 2002; Hu et al., 2016).

Concatenation is a feature integration method where features from the two sources are directly concatenated as one feature vector, which is then fed into a single classifier as an input. By applying this method, the TFIDF matrices of the two sources were combined into a larger matrix, whose number of terms is the sum of the terms of the two individual matrices.

On the other hand, late fusion is done after training two separate classifiers, each for one of the two sources. Once the classifiers predict the label with certain probabilities, the probabilities from the two classifiers are integrated by using a convex combination as follows:

$$p_{hybrid} = \alpha p_{interpretation} + (1-\alpha) p_{lyric},$$

where the mixing weight α defines how much each classifier's result contributes to the fused prediction. Although late fusion has a disadvantage of requiring more learning costs (as it needs to learn the parameter α), it focuses each classifier on one feature set, yielding better results in some benchmarks (Atrey et al., 2010; Snoek et al., 2005). In this work, we tuned the optimal α that maximized the training classification accuracy by using a 10-fold cross-validation.

Classification Setup

Using the interpretations and lyrics data, we tested and compared the performance of four popular multi-class classifiers: Support Vector Machine (SVM) with linear kernel, SVM with Radial Basis Function (RBF) kernel, k-Nearest Neighbor (kNN), and Naïve Bayes (NB) classifier. We performed a 10-fold cross validation to evaluate the performance of each classifier with different features. During the training processes, the parameters, C for both SVMs, γ for RBF SVM, and the number of neighbors (k) for kNN, were optimized using a grid search.

In addition, to prevent the classification results from being skewed by songs with a large number of interpretations

(e.g., if song X has 1,000 interpretations while others have significantly fewer interpretations, terms appearing in song X's interpretations may dominate the classification result), we set the maximum number of interpretations per song to be used to ten. We examined the actual number of interpretations for the songs in our dataset to determine this maximum cutoff point (Table 2). As a substantial number of songs have ten or less interpretations, we determined ten to be a reasonable cutoff point. In order to ensure the overall quality of the interpretations, we selected the top-rated interpretations.

Subject Category	*X*=10	*X*=20	*X*=30	*X*=40	*X*=50
religion	65	50	39	33	29
sex	64	43	34	30	25
drugs	75	60	46	39	35
parent	67	44	34	27	22
war	66	46	36	25	20
places	52	27	19	13	9
ex-lover	62	51	37	32	25
death	52	40	33	23	20

Table 2. Number of songs with at least X interpretations

Additionally, we verified that increasing the maximum number of interpretations used in the classification experiment did not produce significantly better results by comparing the average accuracy across different maximum numbers of interpretations. This was consistent in all four classifiers and we report the average accuracies from linear SVM in Figure 2.

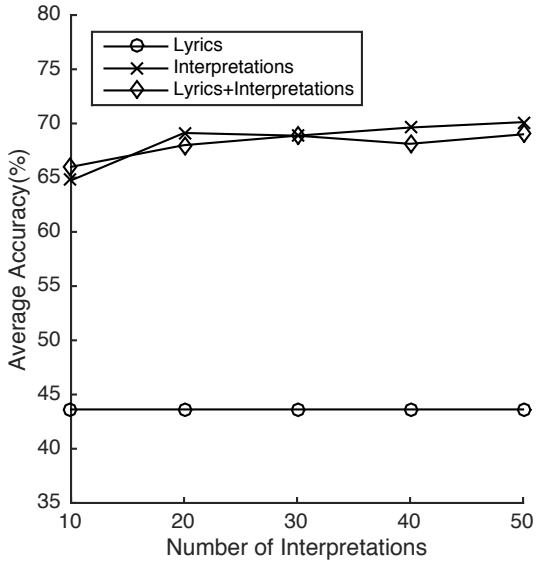

Figure 2. Average accuracy of different features and numbers of interpretations using linear SVM

Table 3 compares the average accuracy across all categories of the four classifiers using three different features; interpretations, lyrics, and lyrics plus interpretations. The maximum values are in bold. The features were represented using TFIDF weighting. Linear SVM had the best performance across all three feature sets, and kNN performed the worst. It is not surprising that SVM performed the best, as it has been shown to be effective in a number of previous studies on music and text classification (Marthet et al., 2013; Hu & Downie, 2005; Yang & Chen, 2012). In addition, the linear kernel of SVM often outperforms non-linear ones in text classification partially because of the high dimensionality of and redundancy among the features (Aggarwal & Zhai, 2012). Therefore, we decided to use the best performing classifier, linear SVM, for subsequent evaluation and comparisons. It is also worth noting that interpretations performed better than lyrics ($p < 0.05$), while the combination of interpretations and lyrics did not significantly increase the performance ($p = 0.615$). The next section will compare the three feature sets in more detail.

Classifiers	Lyrics	Interpretations	Concatenation
Linear SVM	43.6 %	64.8 %	**66.0** %
RBF SVM	41.8 %	**62.9** %	62.6 %
NB	42.1 %	**63.3** %	**63.3** %
kNN	32.4 %	**59.0** %	50.0 %

Table 3. Classification accuracy of the four classifiers

RESULTS

In this section, we compared different configurations of input sources in terms of their classification accuracy. On top of the straightforward use of either the lyrics or interpretations, we also investigated two different harmonization techniques of the two sources: concatenation and late fusion. In addition to the classification accuracy, we also report representative features per category, by enumerating the top 20 terms with the highest SVM coefficients. Finally, we show confusion matrices of different classification scenarios to analyze the similarities and relationships between the subject categories.

Interpretations vs. Lyrics vs. Combination of the Two Sources

Table 4 provides a detailed view on how these classifiers performed across multiple text inputs as well as subject categories. The maximum accuracies are highlighted in bold. We analyzed the table with various criteria as follows:

- The text features vs. a random guess: Both lyrics and interpretations, whether they are combined or not, showed much better performance than a random guess, 12.5%.
- Interpretations vs. lyrics: The classification results using interpretations statistically outperformed the cases with lyrics ($p < 0.05$), consistently across all the categories. We speculate that this is because lyrics often do not describe the subjects directly. Instead, they use rhetorical expressions, and thus the literal meanings of the words in lyrics may appear to be irrelevant for classifiers in determining the subject.

- Unimodal vs. multimodal: In Table 4, the fourth and fifth columns, concatenation and late fusion, represent the multimodal classification cases, while the first two are for the unimodal cases. First, we can see that the multimodal classification cases using lyrics and interpretations together improved the overall classification result in general, although it is not statistically significant (p = 0.78 for concatenation, p = 0.72 for late fusion). We can observe that one of the hybrid classifiers using both features produced the best results in all categories, except for the two categories, "religion" and "death", where the interpretation-only feature set was the best. This indicates that lyrics and interpretations compensate for each other to some degree. From the insignificant difference between the combined cases and the interpretations-only case, we can either speculate that (a) lyrics may not add a substantial amount of information to interpretations in both ensemble methods, or (b) the combined features (as in the concatenation case) may result in a disadvantage in the classification task due to the high dimensionality. This finding is similar to that in Bischoff et al. (2009), who compared tags, lyrics and the combinations of the two, and found that adding lyrics reduced the performance in genre and style classification.
- Concatenation vs. late fusion: Late fusion produced almost the same performance as concatenation, and the difference between the two was not statistically significant (p = 0.94). The optimal α values for each fold were found by a nested additional 10-fold cross validation on the fold-specific training set.
- Lyrics vs. interpretations in terms of contributions to the combined system: We conducted a separate experiment to do an in-depth investigation of how α values influence the classification accuracy. Unlike the previous late fusion experiment, we used fixed alpha values ranging from 0 to 1 with a step size of 0.1. Figure 3 shows the average accuracy across folds per alpha values. Like the previous feature harmonization experiments, the late fused classifier yielded the best performance of 66.1%, whose α value was 0.9. This suggests that there is some additional information in lyrics that interpretations are missing, although the amount of that information is very small. One possible explanation for this is that interpretations may already contain the information in lyrics most of the time.

Representative Features of Subject Categories

To identify the specific representative features in each category, we reviewed the top 20 words from each of the eight categories in interpretations (Table 5) and lyrics (Table 6). The ranking of the words is based on their contribution to the classification, which is represented by the corresponding coefficients trained in the SVM model. The terms that appear only in one of the sources are highlighted in bold.

Subjects	Lyrics	Interpre -tations	Concate -nation	Late Fusion
places	49.0 %	58.0 %	59.0 %	**61.0** %
sex	65.0 %	70.0 %	**75.0** %	73.0 %
ex-lover	36.0 %	67.0 %	67.0 %	**68.0** %
drugs	36.0 %	69.0 %	**71.0** %	70.0 %
war	65.0 %	76.0 %	**79.0** %	**79.0** %
parent	34.0 %	57.0 %	59.0 %	**60.0** %
religion	35.0 %	**70.0** %	67.0 %	**70.0** %
death	29.0 %	**51.0** %	**51.0** %	50.0 %
average	43.6 %	64.8 %	66.0 %	**66.4** %

Table 4. Classification accuracy across categories

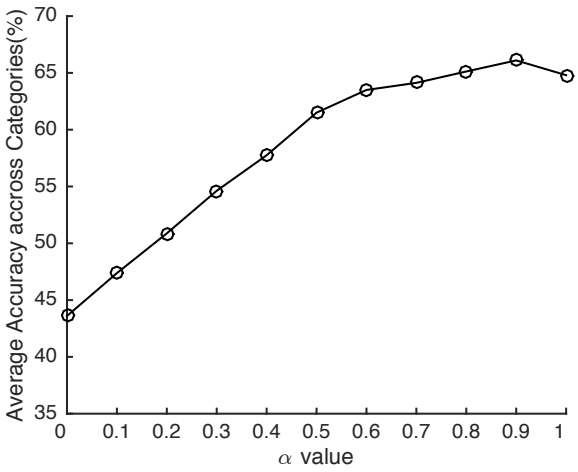

Figure 3. α values and the average accuracy in the entire dataset

In Table 5 (top terms in interpretations), the category "war" contains terms related to particular wars (e.g. the Iraq War and the Vietnam War). Additionally, we can see some nouns (e.g. "soldier", "country", and "troops") and verbs (e.g. "kill" and "fight") that are all related to war. Similarly, the category "religion" mostly contains religious terms such as "Jesus Christ," "Christian," "faith," "heaven," etc. Not surprisingly, several terms in the "places" category were names of particular cities or states (e.g., New York and California) as well as general terms referring to places (e.g., city, town, and hometown). Overall, some of the highly ranked terms in interpretations refer to abstract concepts such as "religion", "addiction", and "relationship," which are less common in the lyric terms. Moreover, interpretation terms tend to reflect a richer vocabulary, since the users might have elaborated on the subject with their own words (e.g. "Christianity", "Iraq", "afterlife", "optimistic", etc.).

places		sex		ex-lover		drugs		war		parent		religion		death	
city	0.13	sex	0.24	love	0.18	drug	0.4	war	0.51	father	0.26	god	0.36	**death**	0.16
york	0.11	girl	0.07	**relations**	0.14	heroin	0.16	soldier	0.17	**dad**	0.21	**religion**	0.19	die	0.15
town	0.08	**sexu**	0.07	**girl**	0.12	**addiction**	0.13	**iraq**	0.11	mother	0.21	jesus	0.13	life	0.08
californium	0.06	sexy	0.06	**guy**	0.09	**addict**	0.08	**nuclear**	0.09	mom	0.13	**christian**	0.1	dead	0.07
hometo	0.06	**danc**	0.05	**break**	0.09	cocaine	0.08	**vietnam**	0.07	**parent**	0.09	**faith**	0.09	**suicide**	0.06
place	0.05	**dirty**	0.05	**hurt**	0.06	marijuana	0.07	**country**	0.07	son	0.07	**religious**	0.09	**hold**	0.04
bay	0.04	**rock**	0.04	**situation**	0.05	**pot**	0.05	fight	0.06	**mum**	0.07	**belief**	0.07	**afterlife**	0.04
california	0.04	**fun**	0.04	**basically**	0.05	**high**	0.05	army	0.06	life	0.05	lord	0.06	**heaven**	0.04
evocativ	0.04	**ha**	0.04	**feel**	0.04	**dealer**	0.05	kill	0.06	**leave**	0.05	**christianity**	0.06	**friend**	0.04
land	0.03	**yeah**	0.04	**sense**	0.04	**smoking**	0.04	peace	0.06	**ocean**	0.05	**christ**	0.06	**sister**	0.04
detroit	0.03	**wom**	0.04	**boyfriend**	0.04	**weed**	0.04	battle	0.06	**die**	0.04	**prayer**	0.06	sleep	0.03
ny	0.03	**beat**	0.04	**girlfriend**	0.04	**speed**	0.04	bomb	0.05	**lose**	0.04	**sin**	0.05	**optimistic**	0.03
beach	0.03	**bitch**	0.04	stay	0.04	**smoke**	0.04	**govern**	0.05	**child**	0.04	**human**	0.05	**beautiful**	0.03
canajohari	0.03	**catch**	0.04	**feeling**	0.04	**medicine**	0.04	**troops**	0.04	**sad**	0.04	heaven	0.05	**save**	0.03
tower	0.03	hot	0.03	**time**	0.04	**escape**	0.03	**military**	0.04	**teach**	0.04	catholic	0.05	**body**	0.03
boston	0.03	**bass**	0.03	summer	0.04	**trip**	0.03	**rain**	0.04	daddy	0.04	evil	0.04	**pass**	0.03
angeles	0.03	blowjob	0.03	dream	0.03	**acid**	0.03	**oil**	0.04	**cancer**	0.03	**church**	0.04	**accord**	0.03
company	0.03	guy	0.03	**cheat**	0.03	**needle**	0.03	flag	0.03	**beautiful**	0.03	**satan**	0.04	**moment**	0.03
orleans	0.02	**lust**	0.03	**friend**	0.03	**monkey**	0.03	**win**	0.03	**chinese**	0.03	**bible**	0.04	**fear**	0.03
texas	0.02	**wan**	0.03	**bf**	0.03	train	0.03	**support**	0.03	**pain**	0.03	sinner	0.04	**hammer**	0.02

Table 5. Top 20 features from user interpretations with their SVM coefficients per category

places		sex		ex-lover		drugs		war		parent		religion		death	
city	1.04	girl	0.74	**kiss**	0.50	drug	0.62	war	1.21	father	0.68	god	1.16	life	0.54
town	0.86	**baby**	0.66	**wan**	0.49	cocaine	0.44	soldier	0.58	**mama**	0.60	**holy**	0.57	die	0.54
york	0.66	**love**	0.55	**hold**	0.45	**sweet**	0.34	peace	0.51	**papa**	0.60	jesus	0.56	dead	0.39
night	0.54	**animal**	0.46	stay	0.42	**ta**	0.33	death	0.47	mother	0.46	sin	0.44	**goodbye**	0.37
bay	0.42	**fuck**	0.45	**kinda**	0.40	**friend**	0.32	**march**	0.42	**man**	0.39	heaven	0.41	**waste**	0.36
california	0.40	sex	0.43	**understand**	0.39	**fall**	0.32	**gun**	0.41	daddy	0.39	long	0.35	**edge**	0.30
america	0.37	**lovin**	0.40	**guess**	0.36	**wake**	0.32	**die**	0.41	son	0.38	word	0.31	**curtain**	0.28
alien	0.29	**touch**	0.36	**add**	0.35	marijuana	0.32	fight	0.40	**afraid**	0.38	light	0.30	**dancing**	0.27
beach	0.29	**body**	0.35	**change**	0.34	**pill**	0.32	bomb	0.39	**lullaby**	0.37	lord	0.30	**river**	0.27
land	0.29	**gon**	0.34	**pretend**	0.33	**stick**	0.31	kill	0.37	**dream**	0.34	catholic	0.29	**deep**	0.25
shiny	0.28	sexy	0.33	love	0.31	**doctor**	0.30	flag	0.35	life	0.31	**heart**	0.27	**realize**	0.25
place	0.27	**dog**	0.32	**talk**	0.30	train	0.29	**mistake**	0.34	mom	0.31	**gloria**	0.27	**suffer**	0.24
barcelona	0.26	**ball**	0.32	dream	0.28	**cold**	0.29	battle	0.34	**easy**	0.30	evil	0.27	**lose**	0.24
shack	0.25	**grind**	0.31	**everytime**	0.28	**pay**	0.27	**weep**	0.32	**eye**	0.30	**jaya**	0.26	**air**	0.23
tokyo	0.25	**cake**	0.31	**bitch**	0.28	**ride**	0.25	**tear**	0.31	**wind**	0.29	**carry**	0.26	sleep	0.23
shadow	0.23	hot	0.30	summer	0.27	**pull**	0.25	**bullet**	0.31	**scare**	0.28	**bodhisattva**	0.25	**burst**	0.23
cadillac	0.23	**good**	0.30	**start**	0.27	**bagman**	0.25	**wall**	0.31	**hurt**	0.28	sinner	0.25	**ledge**	0.22
montego	0.23	**nasty**	0.26	**remedy**	0.26	**hit**	0.24	army	0.28	**aeon**	0.27	**higher**	0.25	**boil**	0.22
downtown	0.22	**feel**	0.26	**fit**	0.26	**wide**	0.24	**forget**	0.27	**happy**	0.26	**weaver**	0.25	**wrong**	0.21
haw	0.22	**denial**	0.26	**baby**	0.26	heroin	0.24	**galve**	0.27	**babylon**	0.26	**bring**	0.25	**head**	0.21

Table 6. Top 20 features from lyrics with their SVM coefficients per category

On the other hand, in Table 6 (top terms in lyrics), some terms seem less relevant to the categories compared to the terms in Table 5. For instance, in the "places" category, the terms, "night" and "shiny", do not seem to be particularly closely related to the subject, while there are no such terms in the same category in Table 5. We can also see that sometimes the most straightforward terms describing subject categories have lower ranks in lyrics or do not appear at all. For instance, "sex" in "sex" category ranked sixth in lyrics while it placed first in interpretations. Similarly, the term "death" does not appear in the "death" category feature set at all in Table 6. Based on the comparison, we believe that the semantic relevancy of top terms explains the performance gap between interpretations and lyrics.

Confusion among Subject Categories

We also examined the confusion matrices among the categories to find out which categories were often misclassified by each of the classifiers based on interpretations, lyrics, and combination of both. The columns of the matrix represent predicted classes and the rows represent our ground truth classes derived from Songfacts.

From Figure 4 (confusion matrix of the interpretation-based classifier), we can see the most confusing pair of categories was "parents" and "death"; 14% of songs about "parents" were misclassified as "death". Examining the top terms in the two categories (Table 5) reveals that terms such as "life" and "die" are all highly ranked in both categories, which explains the reason for confusion. Another highly confusing pair was "sex" and "ex-lover" (11% and 12% error rates). From Table 5, we can see that the term "girl" is highly ranked in both categories, which may have contributed to the confusion. Most categories were also often misclassified as the "places" category, which may be because users are explaining some aspects of the setting of the stories being told in the songs.

The confusion matrix based on lyrics, on the other hand, shows slightly more confusion as observed in Figure 5. The categories that were most frequently confused were "ex-lover" and "sex" (21% of "ex-lover" songs were misclassified as "sex"). "Parents", "religion", and "drugs" were also often misclassified as "death".

Both multimodal classification results, concatenation and late fusion, show similar confusion matrices, as can be seen in Figure 6 and 7. In both cases, the performance for the "sex" and "war" categories was substantially improved. This improvement seems to stem from the fact that the performance based on lyrics in those categories was also very high. Overall, the rest of the patterns seem closer to what is observed in the confusion matrix for interpretations rather than lyrics.

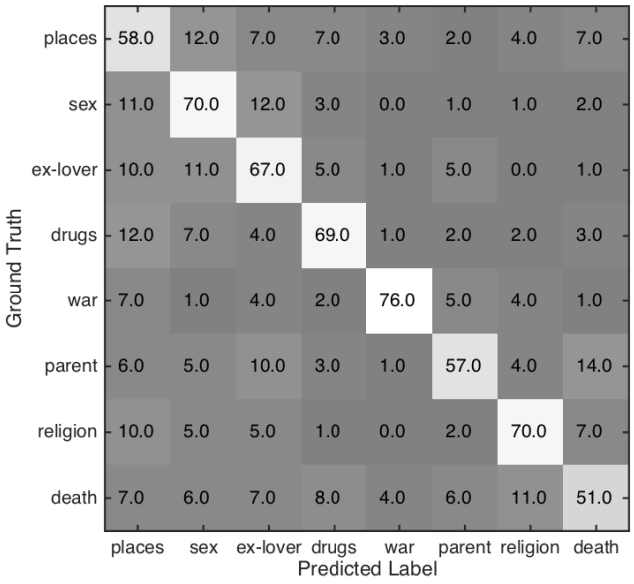

Figure 4. Confusion matrix from linear SVM classifier using interpretations

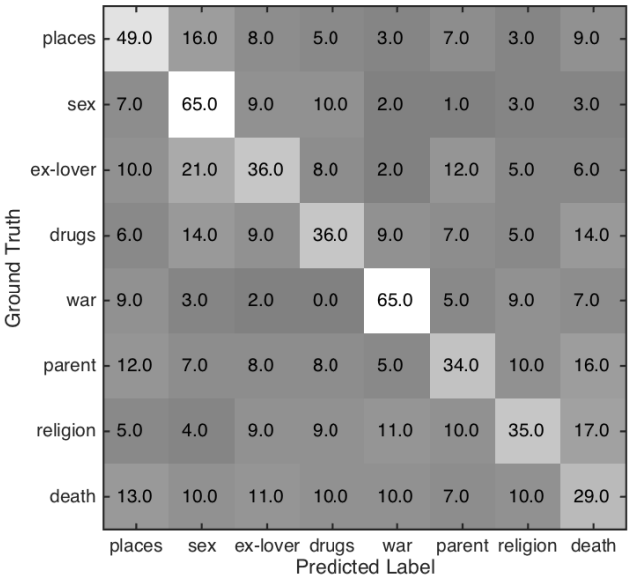

Figure 5. Confusion matrix from linear SVM classifier using lyrics

CONCLUSION AND FUTURE WORK

In this paper, we introduced a system that automatically classifies the subjects of music using lyrics and their user interpretations. The classification accuracies on some popular categories were over 70%. As the first attempt in exploiting user interpretations in song subject classification, this study shows great potential in this line of research.

We compared two different text sources for classifying songs by subject: lyrics and user-generated interpretations. Our experiment showed that, while both sources did contain

subject-related information to some degree, user-generated interpretations outperformed lyrics in the classification on the dataset collected from songmeanings.com and songfacts.com. This confirms our hypothesis that users' interpretations can help reveal the meaning of the song and the artist's intention better than what is conveyed through the lyric texts due to the poetic nature of song lyrics.

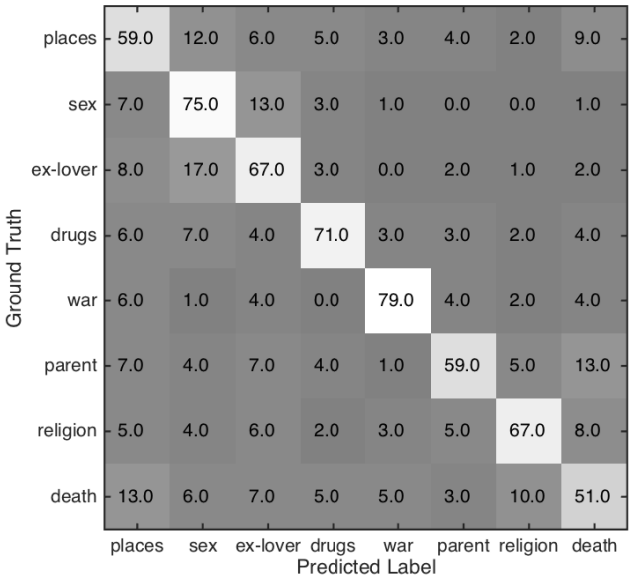

Figure 6. Confusion matrix from linear SVM classifier using combination of lyrics and interpretations

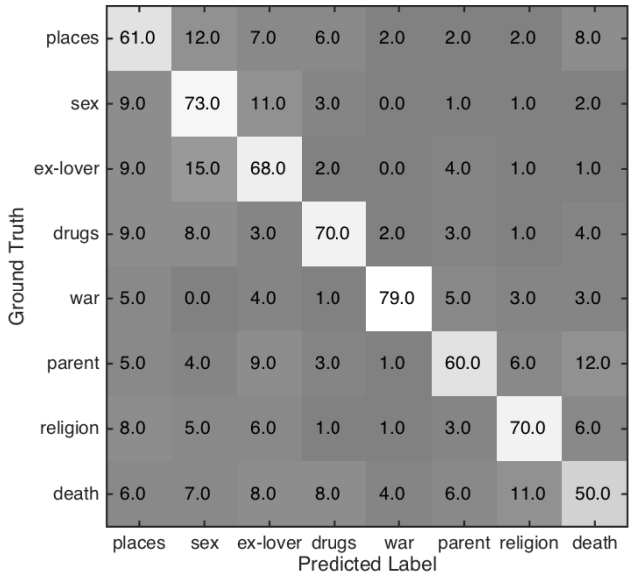

Figure 7. Confusion matrix from the late fused linear SVM classifiers trained from lyrics and interpretations, respectively

In addition, combining interpretations and lyrics slightly improved the classification performance, which indicates that the two sources do compensate for each other, but only a little. Comparison of top ranked features between interpretations and lyrics across categories also shows that the terms from interpretations tend to be more semantically relevant to the subject categories than those from lyrics. Also, the two different feature hybrid methods, concatenation and late fusion, did not make a substantial difference.

Based on our findings, we recommend that the developers prioritize interpretations over lyrics when automatically deriving subject metadata. Furthermore, we encourage the development of new opportunities for music listeners to generate interpretations of more songs.

Finally, the confusion matrices helped identify pairs of categories that were often misclassified as each other. Such findings indicate that aggregating confusing and semantically related categories (e.g., "sex" and "ex-lovers") may be a fruitful approach to improve classification results, though the level of granularity will reduce slightly. We plan to test this in our future work, and also plan to conduct experiments in a multi-label classification setting, as one song may be about multiple subjects. In addition, we plan to expand the setup to include a much larger unbalanced dataset with more categories, which is a more realistic condition. Finally, we plan to explore other feature types such as bigrams and trigrams in addition to unigrams.

ACKNOWLEDGMENTS

We appreciate Craig Willis, Garrick Sherman, and Jacob Jett for valuable comments.

REFERENCES

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

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems, 16*(6), 345-379.

Bainbridge, D., Cunningham, S. J., & Downie, J. S. (2003). How people describe their music information needs: A grounded theory analysis of music queries. In *Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR)* (pp. 221-222).

Barthet, M., Fazekas, G., & Sandler, M. (2012). Music emotion recognition: From content-to context-based models. In *International Symposium on Computer Music Modeling and Retrieval* (pp. 228-252). Springer Berlin Heidelberg.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Symposium on Music Information Retrieval (ISMIR)* (pp. 591-596).

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search?. *In Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 193-202). ACM.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2009). How do you feel about dancing queen?: Deriving mood & theme annotations from user tags. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 285-294). ACM.

Byrd, D., & Crawford, T. (2002). Problems of music information retrieval in the real world. *Information processing & management, 38*(2), 249-272.

Choi, K., Lee J. H., & Downie J. S. (2014). What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 453–454). ACM.

Hu, X., Downie, J. S., West, K., & Ehmann, A. F. (2005). Mining music reviews: Promising preliminary results. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR)* (pp. 536-539). London.

Hu, X., & Downie, J. S. (2010). Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 159-168). ACM.

Hu, X., Choi, K., & Downie, J. S. (2016). A framework for evaluating multimodal music mood classification. *Journal of the Association for Information Science and Technology*.

Kim, J. Y., & Belkin, N. J. (2002). Categories of music description and search terms and phrases used by non-music experts. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR)* (pp. 209-214).

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., & Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proceedings of the 11th International Symposium on Music Information Retrieval (ISMIR)* (pp. 255-266).

Kleedorfer, F., Knees, P., & Pohle, T. (2008). Oh Oh Oh Whoah! Towards automatic topic detection in song lyrics. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR)* (pp. 287-292).

Lee, J. H., & Downie, J. S. (2004). Survey of music information needs, uses, and seeking behaviours: Preliminary findings. In *Proceedings of the 5th International Symposium on Music Information Retrieval (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).

Mahedero, J. P., MartÍnez, Á., Cano, P., Koppenberger, M., & Gouyon, F. (2005). Natural language processing of lyrics. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 475-478). ACM.

Manning, C. D., Raghavan, P., & Schütze, H. (2008) *Introduction to Information Retrieval.* Cambridge University Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)* (pp. 55-60).

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

Meyer, L. B. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, 15(4), 412-424.

Singhi, A., & Brown, D. G. (2014). Are poetry and lyrics all that different?. In *Proceedings of the 15th International Symposium on Music Information Retrieval (ISMIR)* (pp. 471-476).

Snoek, C. G., Worring, M., & Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 399-402). ACM.

Walker P. M. (2016). Subject. Grove Music Online. Oxford University Press, Retrieved April 10, 2016 from http://www.oxfordmusiconline.com/subscriber/article/grove/music/27058.

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

Yang, Y. H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3), 40.

Zavalina, O., Palmer, C. L., Jackson, A. S., & Han, M. J. (2008). Assessing descriptive substance in free-text collection-level metadata. In *Proceedings of the 8th International Conference on Dublin Core and Metadata Applications*