Lyrics, Music, and Emotions

Rada Mihalcea
University of North Texas
rada@cs.unt.edu

Carlo Strapparava
FBK-irst
strappa@fbk.eu

Abstract

In this paper, we explore the classification of emotions in songs, using the music and the lyrics representation of the songs. We introduce a novel corpus of music and lyrics, consisting of 100 songs annotated for emotions. We show that textual and musical features can both be successfully used for emotion recognition in songs. Moreover, through comparative experiments, we show that the joint use of lyrics and music brings significant improvements over each of the individual textual and musical classifiers, with error rate reductions of up to 31%.

1 Introduction

Language and music are peculiar characteristics of human beings. The capability of producing and enjoying language and music appears in every human society, regardless of the richness of its culture (Nettl, 2000).

Importantly, language and music complement each other in many different ways. For instance, looking at music and language in terms of features, we can observe that music organizes pitch and rhythm in ways that language does not, and it lacks the specificity of language in terms of semantic meaning. On the other hand, language is built from categories that are absent in music (e.g., nouns and verbs), whereas music seems to have a deeper power over our emotions than does ordinary speech.

Composers, musicians, and researchers in poetry and literature alike have been long fascinated by the combination of language and music, even since the time of the earliest written records of music encountered in musical settings for poetry. Despite this interest, and despite the long history of the interaction between music and lyrics, there is only little work that explicitly focuses on the connection between music and lyrics.

In this paper, we focus on the connection between the musical and linguistic representations in popular songs, and their role in the expression of *affect*. We introduce a novel corpus of lyrics and music, annotated for emotions at line level, and explore the automatic recognition of emotions using both textual and musical features. Through comparative experiments, we show that emotion recognition can be performed using either textual or musical features, and that the joint use of lyrics and music can improve significantly over classifiers that use only one dimension at a time. We believe our results demonstrate the promise of using joint music-lyric models for song processing.

2 Related Work

The literature on music analysis is noticeably large, and there are several studies concerning the music's power over emotions (Juslin and Sloboda, 2001), thinking (Rauscher et al., 1993), or physical effort (Karageorghis and Priest, 2008).

In particular, there has been significant research in music and psychology focusing on the idea of a parallel between affective cues in music and speech (Sundberg, 1982; Scherer, 1995). For instance, (Scherer, 2004) investigated the types of emotions that can be induced by music, their mechanisms, and how they can be empirically measured. (Juslin and

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 590–599, Jeju Island, Korea, 12–14 July 2012.

Laukka, 2003) conducted a comprehensive review of vocal expressions and music performance, finding substantial overlap in the cues used to convey basic emotions in speech and music.

The work most closely related to ours is the combination of audio and lyrics for emotion classification in songs, as thoroughly surveyed in (Kim et al., 2010). Although several methods have been proposed, including a combination of textual features and beats per minute and MPEG descriptors (Yang and Lee, 2004); individual audio and text classifiers for arousal and valence, followed by a combination through meta-learning (Yang et al., 2008); and the use of crowdsourcing labeling from Last.fm to collect large datasets of songs annotated for emotions (Laurier et al., 2008; Hu et al., 2009), all this previous work was done at song level, and most of it focused on valence-arousal classifications. None of the previous methods considered the fine-grained classification of emotions at line level, as we do, and none of them considered the six Ekman emotions used in our work.

Other related work consists of the development of tools for music accessing, filtering, classification, and retrieval, focusing primarily on music in digital format such as MIDI. For instance, the task of music retrieval and music recommendation has received a lot of attention from both the arts and the computer science communities (see for instance (Orio, 2006) for an introduction to this task). There are also several works on MIDI analysis. Among them, particularly relevant to our research is the work by (Das et al., 2000), who described an analysis of predominant up-down motion types within music, through extraction of the kinematic variables of music velocity and acceleration from MIDI data streams. (Cataltepe et al., 2007) addressed music genre classification (e.g., classic, jazz, pop) using MIDI and audio features, while (Wang et al., 2004) automatically aligned acoustic musical signals with their corresponding textual lyrics. MIDI files are typically organized into one or more parallel "tracks" for independent recording and editing. A reliable system to identify the MIDI track containing the *melody*[1] is very relevant for music information retrieval, and there are several approaches that have been proposed to address this issue (Rizo et al., 2006; Velusamy et al., 2007).

Another related study concerned with the interaction of lyrics and music using an annotated corpus is found in (O'Hara, 2011), who presented preliminary research that checks whether the expressive meaning of a particular harmony or harmonic sequence could be deduced from the lyrics it accompanies, by using harmonically annotated chords from the Usenet group alt.guitar.tab.

Finally, in natural language processing, there are a few studies that mainly exploited the lyrics component of the songs, while generally ignoring the musical component. For instance, (Mahedero et al., 2005) dealt with language identification, structure extraction, and thematic categorization for lyrics. (Xia et al., 2008) addressed the task of sentiment classification in lyrics, recognizing positive and negative moods in a large dataset of Chinese pop songs, while (Yang and Lee, 2009) approached the problem of emotion identification in lyrics, classifying songs from allmusic.com using a set of 23 emotions.

3 A Corpus of Music and Lyrics Annotated for Emotions

To enable our exploration of emotions in songs, we compiled a corpus of 100 popular songs (e.g., *Dancing Queen* by ABBA, *Hotel California* by Eagles, *Let it Be* by The Beatles). Popular songs exert a lot of power on people, both at an individual level as well as on groups, mainly because of the message and emotions they convey. Songs can lift our moods, make us dance, or move us to tears. Songs are able to embody deep feelings, usually through a combined effect of both music and lyrics.

The corpus is built starting with the MIDI tracks of each song, by extracting the parallel alignment of melody and lyrics. Given the non-homogeneous quality of the MIDI files available on the Web, we asked a professional MIDI provider for high quality MIDI files produced for singers and musicians. The MIDI files, which were purchased from the provider, contain also lyrics that are synchronized with the notes. In these MIDI files, the melody channel is unequivocally decided by the provider, making it easier to extract the music and the corresponding lyrics.

[1] A melody can be defined as a "cantabile" sequence of notes, usually the sequence that a listener can remember after hearing a song.

MIDI format. MIDI is an industry-standard protocol that enables electronic musical instruments, computers, and other electronic equipment to communicate and synchronize with each other. Unlike analog devices, MIDI does not transmit an audio signal: it sends event messages about musical notation, pitch, and intensity, control signals for parameters such as volume, vibrato, and panning, and cues and clock signals to set the tempo. As an electronic protocol, it is notable for its widespread adoption throughout the music industry.

MIDI files are typically created using computer-based sequencing software that organizes MIDI messages into one or more parallel "tracks" for independent recording, editing, and playback. In most sequencers, each track is assigned to a specific MIDI channel, which can be then associated to specific instrument patches. MIDI files can also contain lyrics, which can be displayed in synchrony with the music.

Starting with the MIDI tracks of a song, we extract and explicitly encode the following features. At the song level, the key of the song (e.g., G major, C minor). At the line level, we represent the raising, which is the musical interval (in half-steps) between the first note in the line and the most important note (i.e., the note in the line with the longest duration). Finally, at the note level, we encode the time code of the note with respect to the beginning of the song; the note aligned with the corresponding syllable; the degree of the note with relation to the key of the song; and the duration of the note.

Table 1 shows statistics on the corpus. An example from the corpus, consisting of the first two lines from the Beatles' song *A hard day's night*, is illustrated in Figure 3.

SONGS	100
SONGS IN "MAJOR" KEY	59
SONGS IN "MINOR" KEY	41
LINES	4,976
ALIGNED SYLLABLES / NOTES	34,045

Table 1: Some statistics of the corpus

Emotion Annotations with Mechanical Turk. In order to explore the classification of emotions in songs, we needed a gold standard consisting of manual emotion annotations of the songs. Following previous work on emotion annotation of text (Alm et al., 2005; Strapparava and Mihalcea, 2007), to annotate the emotions in songs we use the six basic emotions proposed by (Ekman, 1993): ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE. To collect the annotations, we use the Amazon Mechanical Turk service, which was previously found to produce reliable annotations with a quality comparable to those generated by experts (Snow et al., 2008).

The annotations are collected at line level, with a separate annotation for each of the six emotions. We collect numerical annotations using a scale between 0 and 10, with 0 corresponding to the absence of an emotion, and 10 corresponding to the highest intensity. Each HIT (i.e., annotation session) contains an entire song, with a number of lines ranging from 14 to 110, for an average of 50 lines per song.

The annotators were instructed to: (1) Score the emotions from the writer perspective, not their own perspective; (2) Read and interpret each line in context; i.e., they were asked to read and understand the entire song before producing any annotations; (3) Produce the six emotion annotations independent from each other, accounting for the fact that a line could contain none, one, or multiple emotions. In addition to the lyrics, the song was also available online, so they could listen to it in case they were not familiar with it. The annotators were also given three different examples to illustrate the annotation.

While the use of crowdsourcing for data annotation can result in a large number of annotations in a very short amount of time, it also has the drawback of potential spamming that can interfere with the quality of the annotations. To address this aspect, we used two different techniques to prevent spam. First, in each song we inserted a "checkpoint" at a random position in the song – a fake line that reads "Please enter 7 for each of the six emotions." Those annotators who did not follow this concrete instruction were deemed as spammers who produce annotations without reading the content of the song, and thus removed. Second, for each remaining annotator, we calculated the Pearson correlation between her emotion scores and the average emotion scores of all the other annotators. Those annotators with a correlation with the average of the other annotators below 0.4 were also removed, thus leaving only the reliable annotators in the pool.

<song filename=AHARDDAY.m2a>
<key time=0>G major</key>
<line pvers=1 raising=3 anger=1.5 disgust=0.7 sadness=2.5 surprise=0.8 >
<token time=5040 orig−note=B degree=3 duration=210>IT</token>
<token time=5050 orig−note=B degree=3 duration=210>′S </token>
<token time=5280 orig−note=C′ degree=4 duration=210>BEEN </token>
<token time=5520 orig−note=B degree=3 duration=210>A </token>
<token time=5760 orig−note=D′ degree=5 duration=810>HARD </token>
<token time=6720 orig−note=D′ degree=5 duration=570>DAY</token>
<token time=6730 orig−note=D′ degree=5 duration=570>′S </token>
<token time=7440 orig−note=D′ degree=5 duration=690>NIGHT</token>
</line>
<line pvers=2 raising=5 anger=3.5 disgust=2 sadness=1.2 surprise=0.2 >
<token time=8880 orig−note=C′ degree=4 duration=212>AND </token>
<token time=9120 orig−note=D′ degree=5 duration=210>I</token>
<token time=9130 orig−note=D′ degree=5 duration=210>′VE </token>
<token time=9360 orig−note=C′ degree=4 duration=210>BEEN </token>
<token time=9600 orig−note=D′ degree=5 duration=210>WOR</token>
<token time=9840 orig−note=F′ degree=7− duration=930>KING </token>
<token time=10800 orig−note=D′ degree=5 duration=210>LI</token>
<token time=11040 orig−note=C′ degree=4 duration=210>KE </token>
<token time=11050 orig−note=C′ degree=4 duration=210>A </token>
<token time=11280 orig−note=D′ degree=5 duration=330>D</token>
<token time=11640 orig−note=C′ degree=4 duration=90>O</token>
<token time=11760 orig−note=B degree=3 duration=330>G</token>
</line>

Figure 1: Two lines of a song in the corpus: *It-'s been a hard day-'s night, And I-'ve been wor-king li-ke a d-o-g*

For each song, we start by asking for ten annotations. After spam removal, we were left with about two-five annotations per song. The final annotations are produced by averaging the emotions scores produced by the reliable annotators. Figure 3 shows an example of the emotion scores produced for two lines. The overall correlation between the remaining reliable annotators was calculated as 0.73, which represents a strong correlation.

For each of the six emotions, Table 2 shows the number of lines that had that emotion present (i.e., the score of the emotion was different from 0), as well as the average score for that emotion over all 4,976 lines in the corpus. Perhaps not surprisingly, the emotions that are dominant in the corpus are JOY and SADNESS – which are the emotions that are often invoked by people as the reason behind a song.

Note that the emotions do not exclude each other: i.e., a line that is labeled as containing JOY may also contain a certain amount of SADNESS, which is the reason for the high percentage of songs containing both JOY and SADNESS. The emotional load for the overlapping emotions is however very different. For instance, the lines that have a JOY score of 5 or higher have an average SADNESS score of 0.34. Conversely, the lines with a SADNESS score of 5 or higher have a JOY score of 0.22.

Emotion	Number lines	Average
ANGER	2,516	0.95
DISGUST	2,461	0.71
FEAR	2,719	0.77
JOY	3,890	3.24
SADNESS	3,840	2.27
SURPRISE	2,982	0.83

Table 2: Emotions in the corpus of 100 songs: number of lines including a certain emotion, and average emotion score computed over all the 4,976 lines.

4 Experiments and Evaluations

Through our experiments, we seek to determine the extent to which we can automatically determine the emotional load of each line in a song, for each of the six emotion dimensions.

We use two main classes of features: textual features, which build upon the textual representation of the lyrics; and musical features, which rely on the musical notation associated with the songs. We run three sets of experiments. The first one is intended to determine the usefulness of the textual features for

emotion classification. The second set specifically focuses on the musical features. Finally, the last set of experiments makes joint use of textual and musical features.

The experiments are run using linear regression,[2] and the results are evaluated by measuring the Pearson correlation between the classifier predictions and the gold standard. For each experiment, a ten-fold cross validation is run on the entire dataset.[3]

4.1 Textual Features

First, we attempt to identify the emotions in a line by relying exclusively on the features that can be derived from the lyrics of the song. We decided to focus on those features that were successfully used in the past for emotion classification (Strapparava and Mihalcea, 2008). Specifically, we use: (1) unigram features obtained from a bag-of-words representation, which are the features typically used by corpus-based methods; and (2) lexicon features, indicating the appartenance of a word to a semantic class defined in manually crafted lexicons, which are often used by knowledge-based methods.

Unigrams. We use a bag-of-words representation of the lyrics to derive unigram counts, which are then used as input features. First, we build a vocabulary consisting of all the words, including stopwords, occurring in the lyrics of the training set. We then remove those words that have a frequency below 10 (value determined empirically on a small development set). The remaining words represent the unigram features, which are then associated with a value corresponding to the frequency of the unigram inside each line. Note that we also attempted to use higher order n-grams (bigrams and trigrams), but evaluations on a small development dataset did not show any improvements over the unigram model, and thus all the experiments are run using unigrams.

Semantic Classes. We also derive and use coarse textual features, by using mappings between words and semantic classes. Specifically, we use the Linguistic Inquiry and Word Count (LIWC) and WordNet Affect (WA) to derive coarse textual features. LIWC was developed as a resource for psycholinguistic analysis (Pennebaker and Francis, 1999; Pennebaker and King, 1999). The 2001 version of LIWC includes about 2,200 words and word stems grouped into about 70 broad categories relevant to psychological processes (e.g., emotion, cognition). WA (Strapparava and Valitutti, 2004) is a resource that was created starting with WordNet, by annotating synsets with several emotions. It uses several resources for affective information, including the emotion classification of Ortony (Ortony et al., 1987). From WA, we extract the words corresponding to the six basic emotions used in our experiments. For each semantic class, we infer a feature indicating the number of words in a line belonging to that class.

Table 3 shows the Pearson correlations obtained for each of the six emotions, when using only unigrams, only semantic classes, or both.

Emotion	Unigrams	Semantic Classes	All Textual
ANGER	0.5525	0.3044	0.5658
DISGUST	0.4246	0.2394	0.4322
FEAR	0.3744	0.2443	0.4041
JOY	0.5636	0.3659	0.5769
SADNESS	0.5291	0.3006	0.5418
SURPRISE	0.3214	0.2153	0.3392
AVERAGE	0.4609	0.2783	0.4766

Table 3: Evaluations using textual features: unigrams, semantic classes, and all the textual features.

4.2 Musical Features.

In a second set of experiments, we explore the role played by the musical features. While the musical notation of a song offers several characteristics that could be potentially useful for our classification experiments (e.g., notes, measures, dynamics, tempo), in these initial experiments we decided to focus on two main features, namely the notes and the key.

Notes. A note is a sign used in the musical notation associated with a song, to represent the relative duration and pitch of a sound. In traditional music theory, the notes are represented using the first seven letters of the alphabet (C-D-E-F-G-A-B), al-

[2] We use the Weka machine learning toolkit.

[3] There is no clear way to determine a baseline for these experiments. A simple baseline that we calculated, which assumed by default an emotional score equal to the average of the scores on the training data, and measured the correlation between these default scores and the gold standard, consistently led to correlations close to 0 (0.0081-0.0221).

though other notations can also be used. Notes can be modified by "accidentals" – a sharp or a flat symbol that can change the note by half a tone. A written note can also have associated a value, which refers to its duration (e.g., whole note; eighth note). Similar to the unigram features, for each note, we record a feature indicating the frequency of that note inside a line.

Key. The key of a song refers to the harmony or "pitch class" used for a song, e.g., *C major*, or *F#*. Sometime the term *minor* or *major* can be appended to a key, to indicate a minor or a major scale. For instance, a song in "the key of C minor" means that the song is harmonically centered on the note C, and it makes use of the minor scale whose first note is C. The key system is the structural foundation of most of the Western music. We use a simple feature that reflects the key of the song. Note that with a few exceptions, when more than one key is used in a song, all the lines in a song will have the same key.

Table 4 shows the results obtained in these classification experiments, when using only the notes as features, only the key, or both.

Emotion	Notes	Key	All Musical
ANGER	0.2453	0.4083	0.4405
DISGUST	0.1485	0.2922	0.3199
FEAR	0.1361	0.2203	0.2450
JOY	0.1533	0.3835	0.4001
SADNESS	0.1738	0.3502	0.3762
SURPRISE	0.0983	0.2241	0.2412
AVERAGE	0.1592	0.3131	0.3371

Table 4: Evaluations using musical features: notes, key, and all the musical features.

4.3 Joint Textual and Musical Features.

To explore the usefulness of the joint lyrics and music representation, we also run a set of experiments that use all the textual and musical features. Table 5 shows the Pearson correlations obtained when using all the features. To facilitate the comparison, the table also includes the results obtained with the textual-only and musical-only features (reported in Tables 3 and 4).

Emotion	All Textual	All Musical	Textual & Musical
ANGER	0.5658	0.4405	0.6679
DISGUST	0.4322	0.3199	0.5068
FEAR	0.4041	0.2450	0.4384
JOY	0.5769	0.4001	0.6456
SADNESS	0.5418	0.3762	0.6193
SURPRISE	0.3392	0.2412	0.3855
AVERAGE	0.4766	0.3371	0.5439

Table 5: Evaluations using both textual and musical features.

5 Discussion

One clear conclusion can be drawn from these experiments: the textual and musical features are both useful for the classification of emotions in songs, and, more importantly, their joint use leads to the highest classification results. Specifically, the joint model gives an error rate reduction of 12.9% with respect to the classifier that uses only textual features, and 31.2% with respect to the classifier that uses only musical features. This supports the idea that lyrics and music represent orthogonal dimensions for the classification of emotions in songs.

Among the six emotions considered, the largest improvements are observed for JOY, SADNESS, and ANGER. This was somehow expected for the first two emotions, since they appear to be dominant in the corpus (see Table 2), but comes as a surprise for ANGER, which is less dominant. Further explorations are needed to determine the reason for this effect.

Looking at the features considered, textual features appear to be the most useful. Nonetheless, the addition of the musical features brings clear improvements, as shown in the last column from the same table.

Additionally, we made several further analyses of the results, as described below.

Feature ablation. To determine the role played by each of the feature groups we consider, we run an ablation study where we remove one feature group at a time from the complete set of features and measure the accuracy of the resulting classifier. Table 6 shows the feature ablation results. Note that feature ablation can also be done in the reverse direction, by

Emotion	All Features	All features, excluding				
		Unigrams	Semantic Classes	Notes	Key	Semantic Classes and Notes
ANGER	0.6679	0.4996	0.5525	0.6573	0.6068	0.6542
DISGUST	0.5068	0.3831	0.4246	0.5013	0.4439	0.4814
FEAR	0.4384	0.3130	0.3744	0.4313	0.4150	0.4114
JOY	0.6456	0.5141	0.5636	0.6432	0.5829	0.6274
SADNESS	0.6193	0.4586	0.5291	0.6176	0.5540	0.6029
SURPRISE	0.3855	0.3083	0.3214	0.3824	0.3421	0.3721
AVERAGE	0.5439	0.4127	0.4609	0.5388	0.4908	0.5249

Table 6: Ablation studies excluding one feature group at a time.

Emotion	Baseline	Textual	Musical	Textual and Musical
ANGER	89.27%	91.14%	89.63%	92.40%
DISGUST	93.85%	94.67%	93.85%	94.77%
FEAR	93.58%	93.87%	93.58%	93.87%
JOY	50.26%	70.92%	61.95%	75.64%
SADNESS	67.40%	75.84%	70.65%	79.42%
SURPRISE	94.83%	94.83%	94.83%	94.83%
AVERAGE	81.53%	86.87%	84.08%	88.49%

Table 7: Evaluations using a coarse-grained binary classification.

keeping only one group of features at a time; the results obtained with the individual feature groups are already reported in Tables 3 and 4.

The ablation studies confirm the findings from our earlier experiments: while the unigrams and the keys are the most predictive features, the semantic classes and the notes are also contributing to the final classification even if to a lesser extent. To measure the effect of these groups of somehow weaker features (semantic classes and notes), we also perform an ablation experiment where we remove both these feature groups from the feature set. The results are reported in the last column of Table 6.

Coarse-grained classification. As an additional evaluation, we transform the task into a binary classification by using a threshold empirically set at 3. Thus, to generate the coarse binary annotations, if the score of an emotion is below 3, we record it as "negative" (i.e., the emotion is absent), whereas if the score is equal to or above 3, we record it as "positive" (i.e., the emotion is present).

For the classification, we use Support Vector Machines (SVM), which are binary classifiers that seek to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin (Vapnik, 1995). Applications of SVM classifiers to text categorization led to some of the best results reported in the literature (Joachims, 1998).

Table 7 shows the results obtained for each of the six emotions, and for the three major settings that we considered: textual features only, musical features only, and a classifier that jointly uses the textual and the musical features. As before, the classification accuracy for each experiment is reported as the average of the accuracies obtained during a ten-fold cross-validation on the corpus. The table also shows a baseline, computed as the average of the accuracies obtained when using the most frequent class observed on the training data for each fold.

As seen from the table, on average, the joint use of textual and musical features is also beneficial for this binary coarser-grained classification. Perhaps not surprisingly, the effect of the classifier is stronger for

Emotion	1,000 news headlines		4,976 song lines
	Best result SEMEVAL'07	(Strapparava and Mihalcea, 08)	Joint Text and Music
ANGER	0.3233	0.1978	0.6679
DISGUST	0.1855	0.1354	0.5068
FEAR	0.4492	0.2956	0.4384
JOY	0.2611	0.1381	0.6456
SADNESS	0.4098	0.1601	0.6193
SURPRISE	0.1671	0.1235	0.3855
AVERAGE	0.2993	0.1750	0.5439

Table 8: Results obtained in previous work on emotion classification.

those emotions that are dominant in the corpus, i.e., JOY and SADNESS (see Table 2). The improvement obtained with the classifiers is much smaller for the other emotions (or even absent, e.g., for SURPRISE), which is also explained by their high baseline of over 90%.

Comparison to previous work. There is no previous research that has considered the joint use of lyrics and songs representations for emotion classification at line level, and thus we cannot draw a direct comparison with other work on emotion classification in songs.

Nonetheless, as a point of reference, we consider the previous work done on emotion classification of texts. Table 8 shows the results obtained in previous work for the recognition of emotions in a corpus consisting of 1,000 news headlines (Strapparava and Mihalcea, 2007) annotated for the same six emotions. Specifically, the table shows the best overall correlation results obtained by the three emotion recognition systems in the SEMEVAL task on Affective Text (Strapparava and Mihalcea, 2007): (Chaumartin, 2007; Kozareva et al., 2007; Katz et al., 2007). The table also shows the best results obtained in follow up work carried out on the same dataset (Strapparava and Mihalcea, 2008).

Except for one emotion (FEAR), the correlation figures we obtain are significantly higher than those reported in previous work. As mentioned before, however, a direct comparison cannot be made, since the earlier work used a different, smaller dataset. Moreover, our corpus of songs is likely to be more emotionally loaded than the news titles used in previous work.

6 Conclusions

Popular songs express universally understood meanings and embody experiences and feelings shared by many, usually through a combined effect of both music and lyrics. In this paper, we introduced a novel corpus of music and lyrics, annotated for emotions at line level, and we used this corpus to explore the automatic recognition of emotions in songs. Through experiments carried out on the dataset of 100 songs, we showed that emotion recognition can be performed using either textual or musical features, and that the joint use of lyrics and music can improve significantly over classifiers that use only one dimension at a time.

The dataset introduced in this paper is available by request from the authors of the paper.

Acknowledgments

The authors are grateful to Rajitha Schellenberg for her help with collecting the emotion annotations. Carlo Strapparava was partially supported by a Google Research Award. Rada Mihalcea's work was in part supported by the National Science Foundation award #0917170. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

C. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing, pages 347–354, Vancouver, Canada.

Z. Cataltepe, Y. Yaslan, and A. Sonmez. 2007. Music genre classification using MIDI and audio features. *Journal on Advances in Signal Processing*.

F.R. Chaumartin. 2007. Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June.

M. Das, D. Howard, and S. Smith. 2000. The kinematic analysis of motion curves through MIDI data analysis. *Organised Sound*, 5(1):137–145.

P. Ekman. 1993. Facial expression of emotion. *American Psychologist*, 48:384–392.

X. Hu, J. S. Downie, and A. F. Ehmann. 2009. Lyric text mining in music mood classification. In *Proceedings of the International Society for Music Information Retrieval Conference*, Kobe, Japan.

T. Joachims. 1998. Text categorization with Support Vector Machines: learning with mny relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany.

P. Juslin and P. Laukka. 2003. Communication of emotion in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129:770–814.

P. N. Juslin and J. A. Sloboda, editors. 2001. *Music and Emotion: Theory and Research*. Oxford University Press.

C. Karageorghis and D. Priest. 2008. Music in sport and exercise : An update on research and application. *The Sport Journal*, 11(3).

P. Katz, M. Singleton, and R. Wicentowski. 2007. Swat-mp:the semeval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June.

Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull. 2010. Music emotion recognition: A state of the art review. In *International Symposium on Music Information Retrieval*.

Z. Kozareva, B. Navarro, S. Vazquez, and A. Montoyo. 2007. Ua-zbsa: A headline emotion classification through web information. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June.

C. Laurier, J. Grivolla, and P. Herrera. 2008. Multimodal music mood classification using audio and lyrics. In *Proceedings of the International Conference on Machine Learning and Applications*, Barcelona, Spain.

J. Mahedero, A. Martinez, and P. Cano. 2005. Natural language processing of lyrics. In *Proceedings of MM'05*, Singapore, November.

B. Nettl. 2000. An ethnomusicologist contemplates universals in musical sound and musical culture. In N. Wallin, B. Merker, and S. Brown, editors, *The origins of music*, pages 463—472. MIT Press, Cambridge, MA.

T. O'Hara. 2011. Inferring the meaning of chord sequences via lyrics. In *Proceedings of 2nd Workshop on Music Recommendation and Discovery (WOMRAD 2011)*, Chicago, IL, October.

N. Orio. 2006. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1):1–90, November.

A. Ortony, G. L. Clore, and M. A. Foss. 1987. The referential structure of the affective lexicon. *Cognitive Science*, (11).

J. Pennebaker and M. Francis. 1999. Linguistic inquiry and word count: LIWC. Erlbaum Publishers.

J. Pennebaker and L. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, (77).

F. Rauscher, G. Shaw, and K. Ky. 1993. Music and spatial task performance. *Nature*, 365.

D. Rizo, P. Ponce de Leon, C. Perez-Sancho, A. Pertusa, and J. Inesta. 2006. A pattern recognition approach for melody track selection in MIDI files. In *Proceedings of 7th International Symposyum on Music Information Retrieval (ISMIR-06)*, pages 61–66, Victoria, Canada, October.

K. Scherer. 1995. Expression of emotion in voice and music. *Journal of Voice*, 9:235–248.

K. Scherer. 2004. Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them ? *Journal of New Music Research*, 33:239–251.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii.

C. Strapparava and R. Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic.

C. Strapparava and R. Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the ACM Conference on Applied Computing ACM-SAC 2008*, Fortaleza, Brazile.

C. Strapparava and A. Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon.

J. Sundberg. 1982. Speech, song, and emotions. In M. Clynes, editor, *Music, Mind and Brain: The Neuropsychology of Music*. Plenum Press, New York.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

S. Velusamy, B. Thoshkahna, and K. Ramakrishnan. 2007. Novel melody line identification algorithm for polyphonic MIDI music. In *Proceedings of 13th International Multimedia Modeling Conference (MMM 2007)*, Singapore, January.

Y. Wang, M. Kan, T. Nwe, A. Shenoy, and J. Yin. 2004. LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of MM'04*, New York, October.

Y. Xia, L. Wang, K.F. Wong, and M. Xu. 2008. Lyric-based song sentiment classification with sentiment vector space model. In *Proceedings of the Association for Computational Linguistics*, Columbus, Ohio.

D. Yang and W. Lee. 2004. Disambiguating music emotion using software agents. In *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain.

D. Yang and W. Lee. 2009. Music emotion identification from lyrics. In *Proceedings of 11th IEEE Symposium on Multimedia*.

Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. Chen. 2008. Toward multi-modal music emotion classification. In *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*.