MUSIC MOOD DETECTION BASED ON AUDIO AND LYRICS WITH DEEP NEURAL NET

Rémi Delbouys **Romain Hennequin** **Francesco Piccoli**
Jimena Royo-Letelier **Manuel Moussallam**
Deezer, 12 rue d'Athènes, 75009 Paris, France
research@deezer.com

ABSTRACT

We consider the task of multimodal music mood prediction based on the audio signal and the lyrics of a track. We reproduce the implementation of traditional feature engineering based approaches and propose a new model based on deep learning. We compare the performance of both approaches on a database containing 18,000 tracks with associated valence and arousal values and show that our approach outperforms classical models on the arousal detection task, and that both approaches perform equally on the valence prediction task. We also compare the *a posteriori* fusion with fusion of modalities optimized simultaneously with each unimodal model, and observe a significant improvement of valence prediction. We release part of our database for comparison purposes.

1. INTRODUCTION

Music Information Retrieval (MIR) has been an ever growing field of research in recent years, driven by the need to automatically process massive collections of music tracks, an important task to, for example, streaming companies. In particular, automatic music mood detection has been an active field of research in MIR for the past twenty years. It consists of automatically determining the emotion felt when listening to a track. [1] In this work, we focus on the task of multimodal mood detection based on the audio signal and the lyrics of the track. We apply deep learning techniques to the problem and compare our approach to classical feature engineering-based ones on a database of 18,000 songs labeled with a continuous arousal/valence representation. This database is built on the Million Songs Dataset (MSD) [2] and the Deezer catalog. To our knowledge this constitutes one of the biggest datasets for multimodal mood detection ever proposed.

1.1 Related work

Music mood studies appeared in the first half of the 20th century, with the work of Hevner [7]. In this work, the author defines groups of emotions and studies classical music works to unveil correlations between emotions and characteristics of the music. A first indication that music and lyrics should be jointly considered when analyzing musical mood came from a psychological study exposing independent processing of these modalities by the human brain [3]. For the past 15 years, different approaches have been developed with a wide range of datasets and features. An important fraction of them was put together by Kim et al. in [15]. Li and Ogihara [18] used signal processing features related to timbre, pitch and rhythm. Tzanetakis et al. [28] and Peeters [22] also used classical audio features, such as Mel-Frequency Cepstral Coefficients (MFCCs), as input to a Support Vector Machine (SVM). Lyrics-based mood detection was most often based on feature engineering. For example, Yang and Lee [31] resorted to a psycholinguistic lexicon related to emotion. Argamon et al. [1] extracted stylistic features from text in an author detection task. Multimodal approaches were also studied several times. Laurier et al. [16] compared prediction level and feature level fusion, referred to as late and early fusion respectively. In [26], Su et al. developed a sentence level fusion. An important part of the work based on feature engineering was compiled into more complete studies, among which the one from Hu and Downie [9] is one of the most exhaustive, and compares many of the previously introduced features.

Influenced by advances in deep learning, notably in speech recognition or machine translation, new models began to emerge, based on fewer feature engineering. Regarding audio-based methods, the Music Information Retrieval Evaluation eXchange (MIREX) competition [5] has monitored the evolution of the state of the art. In this framework, Lidy et al. [19] have shown the promise of audio-based deep learning. Recently, Jeon et al. [14] presented the first multimodal deep learning approach using a bimodal convolutional recurrent network with a binary mood representation. However, they neither compared their work to classical approaches, nor evaluated the advantage of their mid-level fusion against simple late fusion of unimodal models. In [12], Huang et al. resorted to deep

[1] We use the words emotion and mood interchangeably, as done in the literature (see [15]).

Boltzmann machines to unveil early correlations between audio and lyrics, but their method was limited by the incompleteness of their dataset, which made impossible the use of temporally local layers, e.g. recurrent or convolutional ones. To our knowledge, there is no clear answer as to whether feature engineering yields better results than more end-to-end systems for the multimodal task, probably because of the lack of easily accessible large size datasets.

1.2 Mood representation

A variety of mood representations have been used in the literature. They either consist of monolabel tagging with either simple tags (e.g. in [9]), clusters of tags (e.g. in the MIREX competition) or continuous representation. In this work, we resort to the latter option. Russell [24] defined a 2-dimensional continuous space of embedding for emotions. A point in this space represents the valence (from negative to positive mood) and arousal (from calm to energetic mood) of an emotion. This representation was used multiple times in the literature [12, 27, 29], and presents the advantage of being satisfyingly exhaustive. It is worth noting that this representation has been validated by embedding emotions in a 2-dimensional space based on their co-occurrences in a database [10]. Since we choose this representation we formulate mood estimation as a 2-dimensional regression problem based on a track's lyrics and/or audio.

1.3 Contributions of this work

We study end-to-end lyrics-based approaches to music mood detection and compare their performance with classical lyrics-based methods performance, and give insights on the performing architectures and networks types. We show that lyrics-based networks show promising results both in valence and arousal prediction.

We describe our bimodal deep learning model and evaluate the performance of a mid-level fusion, compared to unimodal approaches and to late fusion of unimodal predictions. We show that arousal is highly correlated to the audio source, whereas valence requires both modalities to be predicted significantly better. We also see that the latter task can be notably improved by resorting to mid-level fusion.

Finally, we compare our model to traditional feature engineering methods and show that deep-learning-based approaches outperform classical models, when it comes to multimodal arousal detection, and we show that both systems are equally performing on valence prediction. For future comparison purposes, we also release part of our database consisting of valence/arousal labels and corresponding song identifiers.

2. CLASSICAL FEATURE ENGINEERING-BASED APPROACHES

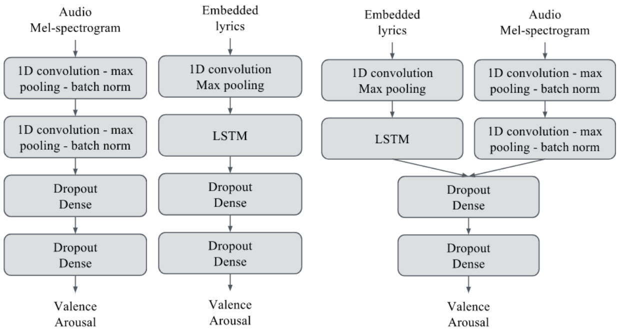
We compare our model to classical approaches based on feature engineering. These methods were iteratively deepened over the years : for audio-based models, a succession of works [18, 22, 28] indicated the top performing audio features for mood detection tasks ; for lyrics-based approaches, a series of studies [1, 10, 31] investigated a wide variety of text-based features. Finally, fusion methods were also studied multiple times [9, 16, 29]. Hu and Downie compiled and deepened these works in a series of papers [8–10], which is the most accomplished feature-engineering-based approach of the subject. We reimplement this work and compare its performance to ours. This model consists in the choice of the optimal weighted average of the predictions of two unimodal models: an SVM on top of MFCCs, spectral flux, rolloff and centroid, for audio; and an SVM on top of basic, linguistic and stylistic features (n-grams, lexicon-based features, etc.) for lyrics.

3. DEEP LEARNING-BASED APPROACH

We first explore unimodal deep learning models and then combine them into a multimodal network. In each case, the model simultaneously predicts valence and arousal. Inputs are subdivided in several segments for training, so that each input has the same length. Output is the average of the predictions computed by the model on several segments of the input. For the bimodal models, subdivision of audio and lyrics requires synchronization of the modalities.

3.1 Audio only

We use a mel-spectrogram as input, which are 2-dimensional. We choose a convolutional neural network (ConvNet) [17], the architecture is shown in Fig. 1 (a). It is composed of two consecutive 1-dimensional convolution layers (convolutions along the temporal dimension) with 32 and 16 feature maps of size 8, stride 1, and max pooling of size 4 and stride 4. We resort to batch normalization [13] after each convolutional layer. We use two fully connected layers as output to the network, the intermediate layer being of size 64.


Audio Mel-spectrogram
1D convolution - max pooling - batch norm
1D convolution - max pooling - batch norm
Dropout Dense
Dropout Dense
Valence Arousal
Embedded lyrics
1D convolution Max pooling
LSTM
Dropout Dense
Dropout Dense
Valence Arousal
Embedded lyrics
1D convolution Max pooling
LSTM
Audio Mel-spectrogram
1D convolution - max pooling - batch norm
1D convolution - max pooling - batch norm
Dropout Dense
Dropout Dense
Valence Arousal

(a) Audio (b) Lyrics (c) Bimodal

Figure 1. Architecture of unimodal and bimodal models

3.2 Lyrics only

We use a word embedding as input to the network, i.e. each word is embedded in a continuous space and the vectors corresponding to each word are stacked, the input being consequently 2-dimensional. We choose to resort to

Model name	Description
CBOW	Continuous bag-of-words: random forest on top of means of input words embedding
GRU	Single Gated Recurrent Unit (GRU) [4], size 40, dense layers of size 64 and 2, preceded by dropout layers of parameter 0.5
LSTM	Single Long Short-Term Memory (LSTM) [6], size 80, dense layers of size 64 and 2, preceded by dropout layers of parameter 0.5
biLSTM	Single LSTM, size 40, dense layers of size 64 and 2, preceded by dropout layers of parameter 0.5
2LSTMs	Two LSTM layers, of size 40, dense layers of size 64 and 2, preceded by dropout layers of parameter 0.5
ConvNet+LSTM	Convolutional layer with 16 features maps of size (2,2), stride 1, max-pooling of size 2, stride 2, an LSTM layer of size 40 and dense layers of size 32 and 2, preceded by dropout layers of parameter 0.5
2ConvNets+2LSTMs	Two convolutional layers with 16 features maps of size (2,2), stride 1, max-pooling of size 2, stride 2, two LSTM layers of size 40 and dense layers of size 32 and 2, preceded by dropout layers of parameter 0.5

Table 1. Description of lyrics-based models.

a word2vec [21] embedding trained on 1.6 million lyrics, as first results seemed to indicate that this specialized embedding performs better than embedding pretrained on an unspecialized, albeit bigger, dataset. We compare several architectures, with recurrent and convolutional layers. One of them is shown in Fig. 1 (b). We also compare this approach with a simple continuous bag-of-words method that acts as a feature-free baseline. The models that were tested are described in Table 1.

3.3 Fusion

For the fusion model, we reuse the unimodal architecture from which we remove the fully connected layers and concatenate the outputs of each network. On top of this concatenation, we use two fully connected layers with an intermediate vector length of size 100. This architecture is presented in Fig. 1(c). This allows for detection of more complex correlations between modalities. We choose to compare this with a simple late fusion, which is a weighted average of the outputs of the unimodal models, the weight being grid-searched. The mid-level fusion model is referred to as `middleDL` and the late fusion model as `lateDL`.

4. EXPERIMENT

4.1 Dataset

The MSD [2] is a large dataset commonly used for MIR tasks. The tracks are associated with tags from LastFM[2], some of which are related to mood. We apply the procedure described by Hu and Downie in [11] to select the tags that are akin to a mood description. We then make use of the dataset published by Warriner et al. [30] which associates 14,000 English words with their embedding in Russell's valence/arousal space. We use it for embedding previously selected tags into the valence/arousal space. When several tags are associated with the same track, we retain the mean of the embedding values. Finally, we normalize the database by centering and reducing valence and arousal. It would undoubtedly be more accurate to have tracks directly labeled with valence/arousal values by humans, but no database with sufficient volume exists. An advantage of this procedure is its applicability to different mood representations, and thus to different existing databases.

The raw audio signal and lyrics are not provided in the MSD. Only features are available, namely MFCCs for audio, word-counts for lyrics. For this reason, we use a mapping between the MSD and the Deezer catalog using the song metadata (song title, artist name, album title) and have then access to raw audio signals and original lyrics for a part of the songs. As a result, we collected a dataset of 18,644 annotated tracks. We note that lyrics and audio are not synchronized. Automatic synchronization being outside of the scope of this work, we resort to a simple heuristic for audio-lyrics alignment. It consists of aligning both modalities proportionally based on their respective length, i.e. for a certain audio segment, we extract words from the lyrics that are at the corresponding location relatively to the length of the lyrics. We release the labels, along with Deezer song identifiers, MSD identifiers, artist and track name[3]. More data can be retrieved using the Deezer API[4]. Unfortunately, we cannot release the lyrics and music, due to rights restrictions.

We train the models on approximately 60% of the dataset, and validate their parameters with another 20%. Each model is then tested on the remaining 20%. We refer to these three sets as training, validation and test set, respectively. We split the dataset randomly, with the constraint that songs by the same artist must not appear in two different sets (since artist and moods may be correlated).

4.2 Implementation details

For audio, we use a mel-spectrogram as input to the network, with 40 mel-filters and 1024 sample-long Hann window with no overlapping, with a sampling frequency of 44.1kHz, computed with YAAFE [20]. We use data augmentation, that was investigated for audio and proven useful in [25], in order to grow our dataset. First, we decide to extract 30 second long segments from the original track. The input of the network is consequently of size 40*1292. We choose to sample seven extracts per track: we draw them uniformly from the song. We also use pitch shifting and lossy encoding, which are transformations with which emotion is invariant, and get three extra segments per original sample. In the end, we get a 28-fold increase in the size of the training set.

For lyrics, the input word embedding was computed with gensim's implementation of word2vec [23] and we used 100-dimensional vectors. We use data augmentation

[2] http://www.last.fm/

[3] https://github.com/deezer/deezer_mood_detection_dataset

[4] https://developers.deezer.com/api

mode	model	valence	arousal
audio	CA	0.118	0.197
	ConvNet	**0.179**	**0.235**
lyrics	CA	**0.140**	**0.032**
	CBOW	0.080	0.031
	LSTM	0.117	0.027
	GRU	0.106	0.017
	biLSTM	0.076	0.017
	2LSTMs	0.128	0.024
	ConvNet+LSTM	0.134	0.026
	2ConvNets+2LSTMs	0.127	0.022
bimodal	CA	**0.219**	0.216
	LateDL	0.194	**0.235**
	middleDL	**0.219**	**0.232**

Table 2. R^2 scores of the different tested approaches.

for lyrics as well by extracting seven 50-word segments from each track. Consequently, the input of each neural network is of size 100*50.

4.3 Results

We present the results and compare in particular deep learning approaches with classical ones. The results are presented in Tab. 2 and 3. In the latter, CA refers to classical models (described in Sect. 2).

Unimodal approaches. The results of each unimodal model are given in Table 2. For lyrics-based ones, we have tested several models without features engineering. The highest performing method, on both validation and test set, is based on both recurrent and convolutional layers. In the following, we choose this model as the one to be compared with classical models.

For both unimodal models, one can see a similar trend for classical and deep learning approaches: lyrics and audio achieve relatively similar performance on valence detection, whereas audio clearly outperforms lyrics when it comes to arousal prediction. This is unsurprising, as arousal is closely related to rhythm and energy, which are essentially induced by the audio signal. On the contrary, valence is explained by both lyrics and audio, indicating that the positivity of an emotion can be conveyed through the text as well as through the melody, the harmony, the rhythm, etc. Similar observations were made by Laurier et al. [16], where angry and calm songs were classified significantly better by audio than by lyrics, and happy and sad songs were equally well-classified by both modalities. This is consistent with our observations, as happy and sad emotions can be characterized by high and low valence, and angry and calm emotions by high and low arousal.

When looking more closely at the results, one can observe that deep learning approaches are much higher performing than classical ones when it comes to prediction based on audio. On the contrary, classical lyrics-based models are higher performing than our deep learning model, in particular when it comes to valence detection, which is the most informative task for the study on lyrics only (as stated above). The reason can be that classical systems resort to several emotion related lexicons designed by psychological studies. On the contrary, classical audio feature engineering for mood detection does not make use of such external resources curated by experts.

Late fusion analysis. As stated earlier, the late fusion consists of a simple optimal weighted average between the prediction of both unimodal models. We resort to a grid-search on the value of the weighting between 0 and 1. The result for the reimplementation of traditional approaches and for our model is presented in Table 3. One can observe a similar phenomenon for both classical models and ours. In both cases, the fusion of the modalities does not significantly improve arousal detection performance compared to audio-based models. It is as predicted, as we saw that audio-based models perform significantly better than lyrics-based ones. For deep learning models, using lyrics in addition to audio in a late fusion scheme leads to no improvement, so there is no gain added by using lyrics. When it comes to valence detection, both modalities are valuable: in both approaches, the top performing model is a relatively balanced average of unimodal predictions. Here also, these observations generalize to valence/arousal what was observed on the emotions happy, sad, angry and calm in [16]. Indeed, based on this study, not only are lyrics and audio equally performant for predicting happy and sad songs, but they are also complementary, so that fused models can achieve notably better accuracies. However, predicting angry and calm songs is not improved when using lyrics in addition to audio.

Bimodal approaches comparison. Bimodal method performances are reported in Table 2. Several interesting remarks can be made based on these results. First of all, one can notice that if one compares late fusion for both approaches, arousal detection is outperformed by deep learning systems, as the corresponding unimodal approach based on audio is more performant, and we have seen that lyrics-based arousal detection is in both cases performing poorly. On the contrary, late fusion for valence detection yields better results for classical systems. In this case, the lack of performance of lyrics-based methods relying on deep learning is not compensated for by a slightly improved audio-based performance.

However, when it comes to mid-level fusion presented in paragraph 3.3, there is a clear improvement for valence detection. It seems to indicate that there might be earlier correlations between both modalities, that our model is able to detect. Concerning arousal detection, the capacity of the network to unveil such correlations seems useless: we have seen that our lyrics-based model is not able to bring additional information to the audio-based model.

This performing fusion, along with more accurately predicted valence thanks to audio, is sufficient for achieving similar performance to classical approaches, without the use of any external data designed by experts. Interestingly, both models remain useful, as long as they learn complementary information. For valence detection, an optimized weighted average of the predictions of both models yields the performance presented in Table 4. We can see

	coefficient*	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Feature engineering approaches	valence	0.133	0.163	0.186	0.201	**0.211**	**0.211**	0.207	0.192	0.174	0.147	0.112
	arousal	0.034	0.081	0.121	0.152	0.178	0.199	0.211	0.217	**0.218**	0.212	0.201
Deep learning approaches	valence	0.118	0.136	0.152	0.165	0.175	0.182	0.186	**0.188**	0.187	0.183	0.177
	arousal	0.025	0.065	0.102	0.135	0.164	0.19	0.212	0.231	0.246	0.257	**0.265**

Table 3. R^2 scores of the late fusion of unimodal models for classical approaches and deep learning approaches, for different values of weighting. *This coefficient is the weight of the audio prediction. The weight of the lyrics prediction is its complementary to one.

modalities	BWC*	CA and DL mean	CA	DL
audio	0.7	**0.193**	0.118	0.179
lyrics	0.5	**0.177**	0.140	0.134
fused	0.5	**0.243**	0.219	0.219

Table 4. R^2 scores of the optimal weighted mean of classical and deep learning approaches for valence prediction for different modalities. *BWC: best weighting coefficient. This coefficient is the optimal weight of the deep learning-based prediction. CA and DL respectively refers to classical approaches and deep learning methods.

a significant gain obtained for a balanced average of both predictions, indicating that both models have different applications, in particular when it comes to lyrics-based valence detection.

5. CONCLUSION AND FUTURE WORK

We have shown that multimodal mood prediction can go without feature engineering, as deep learning-based models achieve better results than classical approaches on arousal detection, and both methods perform equally on valence detection. It seems that this gain of performance is the results of the capacity of our model to unveil and use mid-level correlations between audio and lyrics, particularly when it comes to predicting valence, as we have seen that for this task, both modalities are equally important.

The gain of performance obtained when using this fusion instead of late fusion indicates that further work can be done for understanding correlations between both modalities, and there is no doubt that a database with synchronized lyrics and audio would be of great help to go further. Future work could also rely on a database with labels indicating the degree of ambiguity of the mood of a track, as we know that in some cases, there can be significant variability between listeners. Such databases would be particularly helpful to go further in understanding musical emotion. Temporally localized label in sufficient volume can also be of particular interest. Future work could also leverage unsupervised pretraining to deep learning models, as unlabeled data can be easier to find in high volume. We also leave it as a future work to pursue improvements of lyrics-based models, with deeper architectures or by optimizing word embeddings used as input. Studying and optimizing in detail ConvNets for music mood detection offers the opportunity to temporally localize zones responsible for the valence and arousal of a track, which could be of paramount importance to understand how music, lyrics and mood are correlated. Finally, by learning from feature engineering approaches, one could use external resources designed by psychological studies to improve significantly the prediction accuracy, as indicated by the complementarity of both approaches.

6. ACKNOWLEDGMENTS

The authors kindly thank Geoffroy Peeters and Gabriel Meseguer Brocal for their insights as well as Matt Mould for his proof-reading. The research leading to this work benefited from the WASABI project supported by the French National Research Agency (contract ANR-16-CE23-0017-01).

7. REFERENCES

[1] Shlomo Argamon, Marin Šarić, and Sterling S Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480. ACM, 2003.

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, 2011.

[3] Mireille Besson, Frederique Faita, Isabelle Peretz, A-M Bonnel, and Jean Requin. Singing in the brain: Independence of lyrics and tunes. *Psychological Science*, 9(6):494–498, 1998.

[4] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[5] J Stephen Downie. The music information retrieval evaluation exchange (mirex). *D-Lib Magazine*, 12(12):795–825, 2006.

[6] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. 1999.

[7] Kate Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268, 1936.

[8] Xiao Hu, Kahyun Choi, and J Stephen Downie. A framework for evaluating multimodal music mood

classification. *Journal of the Association for Information Science and Technology*, 2016.

[9] Xiao Hu and J Stephen Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 159–168. ACM, 2010.

[10] Xiao Hu and J Stephen Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *ISMIR*, pages 619–624, 2010.

[11] Xiao Hu, J Stephen Downie, and Andreas F Ehmann. Lyric text mining in music mood classification. *American music*, 183(5,049):2–209, 2009.

[12] Moyuan Huang, Wenge Rong, Tom Arjannikov, Nan Jiang, and Zhang Xiong. Bi-modal deep boltzmann machine based musical emotion classification. In *International Conference on Artificial Neural Networks*, pages 199–207. Springer, 2016.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[14] Byungsoo Jeon, Chanju Kim, Adrian Kim, Dongwon Kim, Jangyeon Park, and Jung-Woo Ha. Music emotion recognition via end-to-end multimodal neural networks.

[15] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *ISMIR*, pages 255–266, 2010.

[16] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688–693. IEEE, 2008.

[17] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE, 2010.

[18] Tao Li and Mitsunori Ogihara. Detecting emotion in music. In *ISMIR*, pages 239–240. Johns Hopkins University, 2003.

[19] Thomas Lidy and Alexander Schindler. Parallel convolutional neural networks for music genre and mood classification. *MIREX*, 2016.

[20] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *ISMIR*, pages 441–446, 2010.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[22] Geoffroy Peeters. A Generic Training and Classification System for MIREX08 Classification Tasks: Audio Music Mood, Audio Genre, Audio Artist and Audio Tag. In *MIREX*, Philadelphia, United States, September 2008.

[23] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[24] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[25] Jan Schluter and Sebastian Bock. Improved musical onset detection with convolutional neural networks. In *Acoustics, speech and signal processing (icassp), 2014 ieee international conference on*, pages 6979–6983. IEEE, 2014.

[26] Feng Su and Hao Xue. Graph-based multimodal music mood classification in discriminative latent space. In *International Conference on Multimedia Modeling*, pages 152–163. Springer, 2017.

[27] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5200–5204. IEEE, 2016.

[28] George Tzanetakis. Marsyas submissions to MIREX 2007. In *ISMIR*, 2007.

[29] Xing Wang, Xiaoou Chen, Deshun Yang, and Yuqian Wu. Music emotion classification of chinese songs based on lyrics using tf* idf and rhyme. In *ISMIR*, pages 765–770, 2011.

[30] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

[31] Dan Yang and Won-Sook Lee. Disambiguating music emotion using software agents. In *ISMIR*, volume 4, pages 218–223, 2004.