Music Emotion Recognition from Lyrics: A Comparative Study

Ricardo Malheiro, Renato Panda, Paulo Gomes, Rui Pedro Paiva

CISUC – Centre for Informatics and Systems of the University of Coimbra, Portugal
{rsmal, panda, pgomes, ruipedro}@dei.uc.pt

Abstract. We present a study on music emotion recognition from lyrics. We start from a dataset of 764 samples (audio+lyrics) and perform feature extraction using several natural language processing techniques. Our goal is to build classifiers for the different datasets, comparing different algorithms and using feature selection. The best results (44.2% F-measure) were attained with SVMs. We also perform a bi-modal analysis that combines the best feature sets of audio and lyrics.The combination of the best audio and lyrics features achieved better results than the best feature set from audio only (63.9% F-Measure against 62.4% F-Measure).

Keywords. Music emotion recognition, lyrics, multi-modal fusion, natural language processing, machine learning

1 Introduction

In the first days of music emotion recognition (MER), most classification systems were based on audio content analysis (e.g., [1]). Recently, researchers started addressing the problem of emotion detection in music lyrics (e.g., [2]). Namely, bi-modal systems, combining audio and lyrics, are being researched. Several bi-modal studies have shown improved classification performances (e.g., [3] [4]).

Our main goal is to investigate the performance of music emotion recognition from lyrics. We also aim to assess whether a bi-modal approach will improve the results obtained with a one-dimension approach based on standard audio features only, like the one we followed in the past (e.g., [5]). The results obtained in this study, along with the results in [6], will be our baseline for future work. We have created a bi-modal dataset that, for the same musical piece, comprises both audio signals and lyrical information. We study the importance of each, as well as their combined effect. The created dataset follows the same organization as the one used in the MIREX[1] mood classification task, i.e., 5 emotion clusters. We have used supervised learning algorithms combined with feature selection strategies. The best results were achieved with an SVM classifier: 63.9% F-Measure in a bi-modal dataset composed by 12 features (11 from audio and 1 from the lyrics). The second best result was

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

attained in an audio-only dataset: 62.4% F-Measure. We believe this paper offers a number of relevant contributions to the MIR/MER research community: a new bi-modal dataset for MER (764 audio and lyrics) and a bi-modal methodology for MER, combining audio and lyrics. The dataset can be downloaded from http://mir.dei.uc.pt/resources/MIREX-like_mood.zip.

2 Related Work

Most studies in music emotion classification are based on datasets collected by the authors on the Internet. These datasets are usually pre-classified with emotions, through tags, taken for instance from sites like AllMusic[2] or Last.FM [7]. In the case of feature extraction of lyrics, the most used features are statistical features like Bag-Of-Words (BOW) [8]. Other kind of features, like linguistic and text stylistic features [3], are also used. In those studies, features are represented by several measures like for example tf-idf or boolean representation [9]. In most studies, audio features outperform lyric features and the combination of both usually yields better results [7]. Our work applies the same approach used in [9] for lyrical features to our newly proposed dataset. As for audio features, standard as well as melodic audio features are extracted, as described in [6].

3 Methods

We used a dataset of 903 audio excerpts organized into five clusters, similarly to the MIREX campaign. This dataset and user annotated clusters were gathered from the Allmusic database. Next, we developed tools to automatically search for lyrics files of the same songs using the Google API. In this process, three sites were used for lyrical information (lyrics.com, ChartLyrics and MaxiLyrics). After removal of some deficient files, the interception of the 903 original audio clips with the lyrics resulted in a dataset containing 764 lyrics and audio excerpts. We have used 2 types of features: features based on existing frameworks like Jlyrics[3], Synesketch[4] and ConceptNet[5] (FF) and BOW features. We considered BOW features with several transformations: stemming, stopwords removal, with none or with both of the previous operations. For each operation, we compared two types of representations for the features: Boolean and tf-idf. For each one of the previous combinations, we calculate unigrams, bigrams and trigrams, creating a total of 24 feature sets. The best feature sets with unigrams, bigrams and trigrams are combined as follows: unigrams+bigrams (combination of unigrams and bigrams) (UB) and unigrams+bigrams+trigrams (UBT). We have also evaluated UB and UBT combined to the best features extracted from FF. At the end, we evaluated the feature sets

[2] http://www.allmusic.com/
[3] http://jmir.sourceforge.net/jLyrics.html
[4] http://synesketch.krcadinac.com/blog/
[5] http://web.media.mit.edu/~hugo/conceptnet/

UB+FF+Audio and UBT+FF+Audio, where (Audio is the best set of audio features, as reported in [6]). Various tests were run with the following supervised learning algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), C4.5 and Naïve Bayes (NB). In addition to classification, feature selection and ranking with the ReliefF algorithm [10] were also performed in order to reduce the number of features and improve the results. For both feature selection and classification, results were validated with repeated stratified 10-fold cross validation (with 20 repetitions), reporting the average obtained accuracy.

4 Experimental Results

Several experiments were performed to evaluate the importance of the various subsets of features and the effect of their combination in emotion classification. In these experiments we performed feature selection to identify the best features in each dataset. In Table 1, we present the best results achieved for the evaluated classifiers in each feature set: UB, UBT, FF and Audio.

Table 1. – Best F-Measure results per dataset and classifier.

Name of the dataset - number of features in the dataset	SVM	C4.5	NB	KNN
UB – 1393 features	40.9%	32%	39.1%	31.1%
UBT – 1897 features	42.2%	32.3%	41.1%	31.8%
FF – 32 features	33.7%	25.5%	26.1%	27.2%
Audio – 11 features	62.4%	**59.1%**	56.5%	**58.2%**
UB + FF - 1425 features	43.2%	27.6%	36.2%	32.2%
UBT + FF – 2005 features	44.2%	31.2%	39.2%	32.7%
UB + FF + Audio – 1436 features	**63.9%**	54.5%	**56.8%**	49%
UBT + FF + Audio – 2016 features	**63.9%**	55.2%	56.7%	49.1%

The best results were always reached with SVM classifiers. Concerning to lyrical features, content-based features (BOW) achieved better results than FF features (predominantely based on the structure of the lyric). These results reinforce the importance of content-based features, as we can see in other studies like [9]. The results in datasets containing unigrams, bigrams and trigrams are always better than the ones attained in datasets with unigrams and bigrams.

The results achieved with the combination of features from audio and lyrics are slightly better than the reference (audio). These results support our initial hypothesis that the combination of audio+lyric features helps to improve the performance attained by each one of them separately. The best results (63.9% F-Measure) were obtained in a feature set of 12 features (after feature selection) (11 from audio and 1 from lyrics). This feature from lyrics is a unigram (the token achieved after stemming – babi). The next 3 more important features from lyrics were also unigrams: gonna, love, night. We can see the description of the best 11 features from audio in [6].

5 Conclusions and Future Work

We investigate the importance of combining both audio and lyric features to improve the results in a typical music emotion recognition task. We applied some of the state of art techniques based on natural language processing to reach our goals. The results with lyric features are worse than the ones with audio features, in agreement with other similar works referenced in this paper. The results obtained suggest that bi-modal approaches help surpassing the current glass ceiling in emotion classification when we use only audio features. In the future we intend to explore more natural language processing algorithms and techniques to obtain more effective lyric features.

Acknowledgments.
This work was supported by the MOODetector project (PTDC/EIA-EIA/102185/2008), financed by the Fundação para Ciência e a Tecnologia (FCT) and Programa Operacional Temático Factores de Competitividade (COMPETE) - Portugal.

6 References

1. Lu, L., Liu, D., Zhang, H.: Automatic Mood Detection and Tracking of Music Audio Signals. IEEE Transactions on Audio, Speech, and Language Processing, 14(1): 5-18. (2006)
2. Hu, Y., Chen, X., Yang, D.: Lyric-based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method. In: Proceedings of the 10th International Conference on Music Information Retrieval. (2009)
3. Hu, X., Downie, J.S.: Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. In: Proceedings of Joint Conference on Digital Libraries. (2010)
4. Laurier, C., Grivolla, J., Herrera, P.: Multimodal Music Mood Classification using Audio and Lyrics. In: Proceedings of the International Conference on Machine Learning and Applications. ICMLA'08. Seventh International Conference, pp. 688-693, IEEE.(2008)
5. Panda, R., Paiva, R. P.: Music Emotion Classification: Dataset Acquisition and Comparative Analysis. 15th International Conference on Digital Audio Effects – DAFx '12, York, UK.(2012)
6. Rocha, B., Panda, R., Paiva, R. P.: Music Emotion Recognition: The Importance of Melodic Features. 5th International Workshop on Machine Learning and Music, Prague, Czech Republic. (2013)
7. Hu, X., Downie, J.: When Lyrics Outperform Audio for Music Mood Classification: a Feature Analysis. In:International Society for Music Information Retrieval Conference, pages 1-6. (2010)
8. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34 (1), 1-47. (2001)
9. Hu, X.: Improving Music Mood Classification Using Lyrics, Audio and Social Tags. PhD Thesis, University of Illinois at Urbana-Champaign. (2010)
10. Robnik-Šikonja, M., Kononenko, I.: Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning, vol. 53, no. 1–2, pp. 23–69. (2003)