Mood Classification Using Lyrics and Audio: A Case-Study in Greek Music

Spyros Brilis, Evagelia Gkatzou, Antonis Koursoumis, Karolos Talvis, Katia L. Kermanidis, and Ioannis Karydis

Dept. of Informatics, Ionian University, 49100, Kerkyra, Greece
{p08bri,p08gkat,p08kour,p08talv,kerman,karydis}@ionio.gr

Abstract. This paper presents a case-study of the effectiveness of a trained system into classifying Greek songs according to their audio characteristics or/and their lyrics into moods. We examine how the usage of different algorithms, featureset combinations and pre-processing parameters affect the precision and recall percentages of the classification process for each mood model characteristic. Experimental results indicate that the current selection of features offers accuracy results, the superiority of lyrics content over generic audio features as well as potential caveats with current research in Greek language stemming pre-processing methods.

Keywords: music mood classification, lyrics, audio, Greek music.

1 Introduction

In the last decade, the paradigm of music distribution has made a shift from physical to online under the auspices of digitally encoded, high quality and portability musical content [1]. Worldwide music lovers have since accumulated large musical collections that require efficient management in order to allow for "natural and diversified access points to music" [2].

Music, being an artistic expression, is a multidimensional phenomenon. This fact is strongly considered in the Music Information Retrieval (MIR) field in order to enhance knowledge retrieval or item searching in large musical collections. To that cause, elicited emotions or mood taxonomy are regarded as important factors. This can be partially attributed to the considerable contextually semantic information hidden within emotional expressions describing mood, as such type of information has been argued to be the key element in any human process concerning music [3].

Despite the highly subjective nature of the perception of mood left to a listener by a musical piece [4], the organisation produced by assigning mood labels to a piece can be of significant importance to a plethora of MIR tasks such as auto-tagging, recommendation and playlist-generation, among others. In particular, the task of automated playlist generation in both web and stand-alone applications, has recently received growing attention by users, developers and researchers, as listeners tend to listen to a sequence of related musical pieces than a single song [5]. Mood classification can not only alleviate the burden

L. Iliadis et al. (Eds.): AIAI 2012 Workshops, IFIP AICT 382, pp. 421–430, 2012.

of creating such playlists based on emotional expression input but can also help users identify musical pieces of their collection that are not part of the commonly played songs and thus, in a sense, forgotten [6].

1.1 Contribution and Paper Organisation

In this work, we propose the utilisation of both musical content and the song's lyrics in order to extract linguistic and acoustic signal features that will be subsequently utilised for the classification of the song into mood categories.

The key contributions of this work can be summarised as follows:

- We have solely concentrated on Modern Greek musical data and accumulated a corpus of 943 songs, for which lyrics have been collected as well as manual mood annotation has been assigned.
- We present a novel methodology to extract a language model for each mood category, i.e. a list of most common/representative words in lyrics per category, that can be of further use to research.
- We have conducted extensive experimentation, with various audio & linguistic features, which results in high performance.

The rest of the paper is organised as follows. Section 2 describes related work while Section 3 presents the proposed features for mood classification. Next, Section 4 details the setup of the experimentation carried out, the results obtained as well as a short discussion on the experimental results. Finally, the paper concludes in Section 5.

2 Related Work

Research in mood detection and classification in musical pieces has received extensive attention during most of the last decade, while since 2007, the Music Information Retrieval Evaluation eXchange (MIREX) evaluation campaign [7] additionally hosts the "Audio Music Mood Classification" task. In this section, we present some of the key assumptions in mood modeling as well as related works in songs' mood classification.

2.1 Mood Taxonomies

In order to be able to categorise songs according to their mood, a mood modeling and mapping process is required and thus, in this work, as in a number of works in the same domain [6,8], the model of Thayer [9] is adopted. In this model, there exist 2 dimensions, valence and arousal, that divide a 2-dimensional emotive plane into 4 parts by having positive/high and negative/low values respectively. In this context, arousal and valence are linked to energy and tension, respectively.

High arousal values correspond to moods such as "angry" and "exciting", while high valence values to moods such as "happy" and "relaxing". On the contrary, negative values of arousal contain moods like "sad" and "serene" while negative values of valence moods such as "bored" and "anxious". Accordingly, each axis is divided into 4 separate parts, each having equal parts in both positive and negative values.

2.2 Mood Classification Using Lyrics

The linguistic features extracted from the lyrics text for applications like mood classification usually include bag-of- words collections [10,11,12], i.e. the text is treated as a collection of unordered words, accompanied by their frequency. Aiming at a metric that is more discriminative between the various text types, the tfidf score takes into account not only the frequency of a term in a given song, but its overall frequency in the collection [6].

The bag-of-words model on lyrics leads to moderate performance [11], unless abundant amount of data is available [6]. To overcome this difficulty, approaches have experimented with language modeling techniques, i.e. the identification of statistical properties of the text of each mood category. Laurier et al. [11] mine the 200 most frequent terms for each mood category and the 200 most frequent terms for its negative counterpart (e.g. "angry" - "not angry"), in an attempt to identify the discriminative terms between the two categories. The most discriminative terms constitute the lyrics features used for the learning experiments. Results are significantly better than the ones achieved by the bag-of-words model.

2.3 Mood Classification Using Audio and Lyrics

Approaches that build on both audio and lyrics content, in order to detect mood, support the assumption that the complementary character of audio and lyrics is based on the common songwriter's effort to produce interrelated audio characteristics and word selection in the lyrics of a song [11,2,8].

Accordingly, this approach is adopted by numerous works [13,14,11,2,8]. Yang and Lee [13], in one of the earlier works in the field, proposed the combination of lyrics and a number of audio features in order to maximise the classification accuracy and minimise the mean error. Nevertheless, the significantly small data corpus (145 songs with lyrics) made the work exploratory to make safe conclusions. In their work, Yang et al. [14], extracted a number of low-level acoustic features from a 30 second part of the song that in addition to the lyrics features produced by 3 different approaches (Uni-gram, Probabilistic Latent Semantic Analysis & Bi-gram) are combined by 3 fusion methods. Therein, songs are classified in four categories following the Russell model [15] to conclude that the use of textual features offers a significant accuracy amelioration of the methods examined. Similarly, Laurier et al.[11] conclude that the combination of audio and lyrics features offer an improvement in the overall classification performance for 4 categories based on Russell's model. Therein, audio features included timbral, rhythmic, tonal and temporal descriptors while from lyrics after mining the most frequent terms, the most discriminative terms constituted the lyrics features, outperforming the bag-of-words model. Hu and Downie [2], presented a differentiated approach as to the assignment of mood labels by exploiting social tags attributed to songs, defining, thus, 18 mood categories. Accordingly, their dataset is significantly larger than previous works (5296 songs). Finally, McVicar et al. [8] explore factors of both audio and lyrics that simultaneously affect the mood of a song.

3 Feature Extraction

Content-based MIR approaches assume that documents are represented by features extracted from the musical documents. As MIR processes depend heavily on the quality of the representation (extracted content features), the performance of a classification process is, to a great extent, defined by the quality of the extracted features. In the analysis to follow, the notion of content is extended from audio to lyrics as well.

3.1 Audio Features

For the extraction of audio features the jAudio application [16] that produces a set of, generic for the purposes of MIR, features was utilised. The audio feature set consists of both one-dimensional (e.g., Zero Crossings) and multi- dimensional feature vectors (e.g., MFCC's).

For the purposes of experimentation in this work, the following features have been retained: Spectral Centroid, Spectral Rolloff Point, Spectral Flux, Compactness, Spectral Variability, Root Mean Square, Fraction Of Low Energy Windows, Zero Crossings, Strongest Beat, Beat Sum, Strength Of Strongest Beat, 13 MFCC coefficients, 9 LPC coefficients and 5 Method of Moments coefficients.

3.2 Lyrics Features

Lyrics text, especially in the dataset used in the present approach that includes all genre categories (e.g. ethnic and folklore songs as well), is highly problematic. The song lyrics undergo a series of pre-processing steps that include:

- removal of punctuation marks, symbols, exclamations, apostrophes & commas.
- dealing with truncated words(e.g. "μου είπες" written as "μου ΄πες", in which case "είπες" and "΄πες" are the same word but are treated as different ones), separate words that are erroneously joined together (e.g. "εγώ πήγα" appearing as "εγώπήγα" and therefore, being treated as one word) weird, archaic popular and poetic word forms, Greek words and/or their first letter written using English alphabet (e.g. "αγάπη" written as "agapi" and "Ζωή" written as "Ζωή" but with the Z being an English capital letter).
- removal of functional and stop words, i.e. words that carry no or very little meaning (e.g. articles, pronouns, prepositions), and therefore do not contribute to the discrimination process.
- stemming (Modern Greek is a highly inflectional language; the identification of the base form of declinable words is of great importance. Stemming was performed using the tool described by Saroukos [17].

Bag-of-Words. For the bag-of-words model, the lyrics of each song are represented as a set of the 20 most frequent words (stems) in the song. Each word is accompanied by its frequency in the song and its tfidf score. The total number of linguistic features is 60.

Language Modeling. Unlike the work by Laurier et al. [11], the extracted language model aims at discriminating between the different mood categories, and not between the positive version of each category and its negative counterpart. Furthermore, no language model distances are taken into account (i.e. not just words discriminating one category from the others, as in Laurier et al. [11]), but the absolute language models (words describing one category, disregarding whether they appear in other categories also; the only precondition is that they don't appear in all the categories), so as to avoid the sparse data problem. For this purpose, the fifty most frequent words in the lyrics of a given category are computed, leading to a total of 200 words for the four categories (less in the case of duplicates). Each of these terms constitutes a linguistic learning feature, and its value is the tfidf metric of the given term in the given song. The complete feature set is not shown here due to page limitation.

It was interesting to observe that, while the top ranked words were shared among the categories, below a certain rank position (25/50) the discriminative power of the terms started to show.

Various feature combinations were experimented with. In the bag-of-words approach, the lyrics feature sets were using: (a) of all 60 features, (b) only the word forms, (c) only tfidf scores and finally, (d) all the previous but with the stems instead of the original word forms. Using the language model, experiments were run using tfidf scores only as well as using tfidf scores with term frequencies (i.e. the number of times the term appears in a song), with and without stemming. All experiments were run using 10-fold cross validation.

4 Performance Evaluation

In this section we experimentally compare the accuracy of the aforementioned audio and lyrics features for each axis of the selected arousal-valence mood modeling using a number of algorithms.

4.1 Experimental Setup

The dataset utilised in this work consist of 943 Greek songs from various genres that include lyrics collected from several sources. The annotation of the dataset with the labels of the mood model selected was made by manual appointment.

Experiments were run using the Weka Machine Learning Workbench [18]. Several learning algorithms were experimented with for investigative purposes. The classifiers for the performed experiments are the ones below. The Naive Bayes classifier, a probabilistic algorithm which applies the Bayes' theorem with strong (naive) independence assumptions. The J48, an algorithm which generates an unpruned or pruned C4.5 decision tree. The IBk, a K-nearest neighbors classifier, using 5-13 neighbors. The Random Forest, a method based on bagging models built using the Random Tree method, in which classification trees are grown on a random subset of descriptors, using 50-80 trees. Support Vector Machines (SVMs) [19] were also experimented with, due to their ability to deal

efficiently with few data and high dimensional spaces, both valid properties of the data used in the current approach. Experiments were run using first degree polynomial kernel functions and Platt's Sequential Minimal Optimization (SMO) algorithm for training [20].

4.2 Experimental Results

The experimentation is divided into 3 parts; in the first part, Figures 1, 2, 3 and 4, the capability of the algorithms to identify the arousal dimension of the mood modeling is examined for (a) the tf combined with tfidf and solely tfidf representations in language model for the lyrics features, (b) using or not the stemming pre-processing and (c) data representations using solely audio or lyrics features and both. The precision and recall values presented in the sequel are weighted average values.

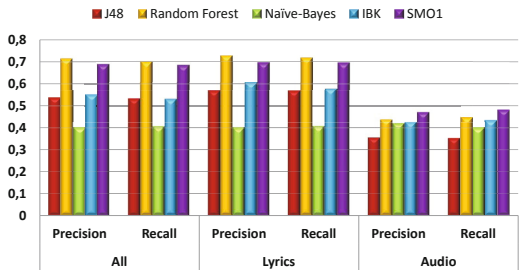

Fig. 1. Precision & recall % on arousal for all algorithms, using tf+tfidf in language model & stemming, for audio only, lyrics only and their combination

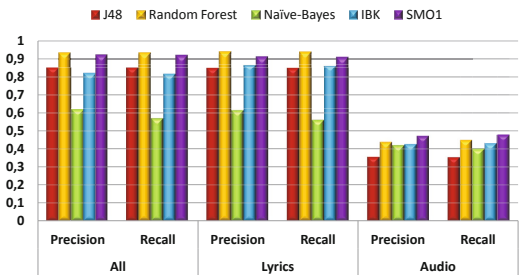

Fig. 2. Precision & recall % on arousal for all algorithms, using tf+tfidf in language model without stemming, for audio only, lyrics only and their combination

In the second part, Figures 5, 6, 7 and 8, the capability of the algorithms to identify the valence dimension of the mood modeling is examined for as in the previous part of the experimentation.

Finally, the third part of our experimental results, Figure 9, is targeted on the bag-of-words model, without the application of stemming for the lyrics features. In this case, the experiment refers only on results obtained from the bag-of-words model. Due to space limitations, only results obtained from the best

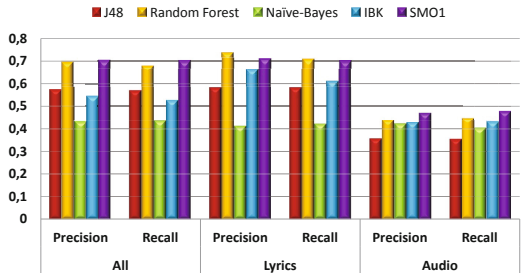

Fig. 3. Precision & recall % on arousal for all algorithms, using only tfidf in language model & stemming, for audio only, lyrics only and their combination

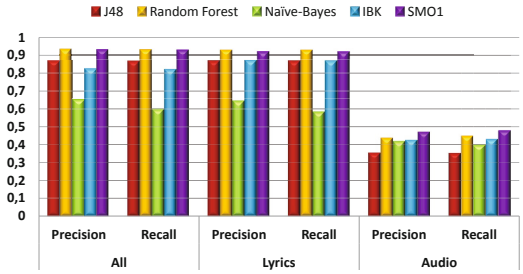

Fig. 4. Precision & recall % on arousal for all algorithms, using only tfidf in language model without stemming, for audio only, lyrics only and their combination

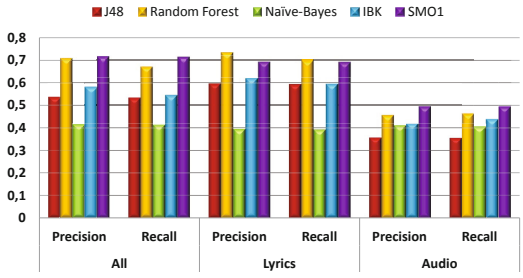

Fig. 5. Precision & recall % on valence for all algorithms, using tf+tfidf in language model & stemming, for audio only, lyrics only and their combination

(unstemmed) dataset are depicted, and the resulting precision and recall values are shown in Figure 9 for both valence and arousal using, only, the lyric features.

4.3 Discussion

The superiority of the language model approach is evident. Part of it is attributed to the nature of the numerical features involved in the corresponding dataset, and the lack of nominal word-based features that take many unique values, present in the bag-of-words dataset. But mostly, it is attributed to the discriminative power of the features. Language modeling results with term frequencies are misleadingly

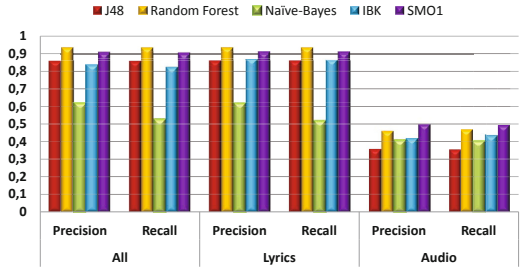

Fig. 6. Precision & recall % on valence for all algorithms, using tf+tfidf in language model without stemming, for audio only, lyrics only and their combination

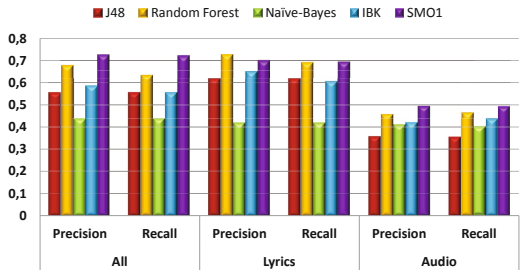

Fig. 7. Precision & recall % on valence for all algorithms, using only tfidf in language model & stemming, for audio only, lyrics only and their combination

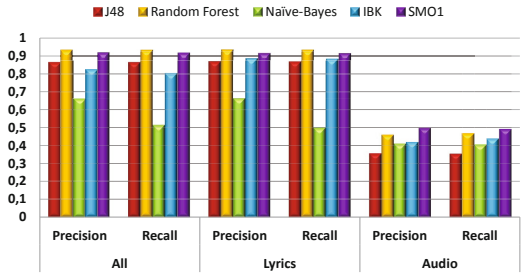

Fig. 8. Precision & recall % on valence arousal for all algorithms, using only tfidf in language model without stemming, for audio only, lyrics only and their combination

optimistic, due to the small range of integer term frequency values, making the learning process easier. Tfidf scores are more objective, as they take into account the distribution of each term in the entire collection. The discriminative power of the top 25 terms in each category is not as great as that of the following ranks, as the top 25 list is mainly comprised of words that are shared among many categories. Terms in ranks 25-50 in a certain category reach higher tfidf scores in instances of that category than in other categories, thus enabling discrimination.

The negative impact of stemming is quite surprising at first sight. Taking a closer look, stemming increases the term frequency range (stem frequency

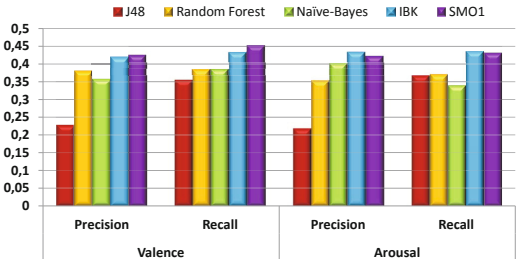

Fig. 9. Precision & recall % on valence & arousal for all algorithms, using the bag-of-words model without stemming, for lyrics only

increases accumulatively from the word forms that belong to it), making the learning process more difficult. Regarding the tool itself, its accuracy is bounded; several errors occur by assigning different lemmata erroneously to the same stem (e.g. "φορά"/turn and "φοράω"/wear are both stemmed as "φορ-"), and vice versa ("λέω"/say is stemmed as "λ-" and "είπα"/said as "είπ-"). Furthermore, the problematic nature of the lyrics text poses significant difficulties on the stemming tool. Truncated and concatenated words, quite frequent in the text, are impossible to stem correctly, while archaic, folklore, historical and popular word forms (popular referring to the Greek music type) make the problem worse.

The classifier which comes up with the best result using the stemmed dataset is the Random Forest algorithm with 71,368% accuracy. To achieve that accuracy, we used 60 trees and the features used are tf+tfidf and the tracks are classified by the arousal axis. For the unstemmed data set, the best accuracy achieved, using the Random Forest algorithm, is 93,74%, using the same features as above and the same mood classification axis.

5 Conclusion

In this work, we present a case-study evaluation of five different classification algorithms for musical data (songs), based on mood. Songs are represented by audio and lyrics features and our experimentation is considering all three alternatives of using solely audio or lyrics features as well as using their combination. Evaluation of the performance of the examined approaches is done using manually annotated ground-truth labels assigned to each song.

Experimental results on a corpus of 943 Greek songs, reveal the superiority of the lyrics content and especially the language model approach, as well as a negative impact of the stemming pre-processing on lyrics that is attributed to the implementation used for the Greek language.

Future research directions in field include the expansion of the dataset in order to strengthen the obtained results, the cross-reference of the manual mood annotation by more users, the utilisation of a wider variety of audio features that are not generic in music information retrieval and the improvement of the morphology of the lyrics, aiming at minimisation of classification and pre-processing errors of the corpus.

References

1. Lam, C.K.M., Tan, B.C.Y.: The internet is changing the music industry. Commun. ACM 44(8), 62–68 (2001)
2. Hu, X., Downie, J.S.: Improving mood classification in music digital libraries by combining lyrics and audio. In: Proc. of Joint Conference on Digital Libraries, pp. 159–168 (2010)
3. Byrd, D.: Organization and searching of musical information, course syllabus (2008)
4. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions from audio. In: Proc. of International Society for Music Information Retrieval, pp. 465–470 (2010)
5. McFee, B., Lanckriet, G.R.G.: The natural language of playlists. In: Proc. of International Society for Music Information Retrieval, pp. 537–542 (2011)
6. van Zaanen, M., Kanters, P.: Automatic mood classification using tf*idf based on lyrics. In: Proc. of International Society for Music Information Retrieval, pp. 75–80 (2010)
7. Downie, S.J., West, K., Ehmann, A., Vincent, E.: The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): preliminary overview. In: Proc. of International Conference for Music Information Retrieval, pp. 320–323 (2005)
8. McVicar, M., Freeman, T., De Bie, T.: Mining the correlation between lyrical and audio features and the emergence of mood
9. Thayer, R.: The biopsychology of mood & arousal. Oxford University Press (1989)
10. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. In: Proc. of International Society for Music Information Retrieval (2009)
11. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: Proc. of International Conference on Machine Learning and Applications, pp. 688–693 (2008)
12. Mayer, R., Neumayer, A.R.: Rhyme and style features for musical genre classification by song lyrics. In: Proc. of International Conference on Machine Learning and Applications, pp. 337–342 (2008)
13. Yang, D., Lee, W.S.: Disambiguating music emotion using software agents. In: Proc. of International Conference on Music Information Retrieval (2004)
14. Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., Chen, H.H.: Toward Multi-modal Music Emotion Classification. In: Huang, Y.-M.R., Xu, C., Cheng, K.-S., Yang, J.-F.K., Swamy, M.N.S., Li, S., Ding, J.-W. (eds.) PCM 2008. LNCS, vol. 5353, pp. 70–79. Springer, Heidelberg (2008)
15. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)
16. McEnnis, D., McKay, C., Fujinaga, I.: jAudio: A feature extraction library. In: Proc. International Conference on Music Information Retrieval (2005)
17. Saroukos, S.: Enhancing a greek language stemmer - efficiency and accuracy improvements. Master's thesis, Dept. of Computer Sciences, University of Tampere, Finland (2008)
18. Holmes, G., Donkin, A., Witten, I.H.: Weka: A machine learning workbench. In: Proc. of Intelligent Information Systems, pp. 357–361 (1994)
19. Vapnik, V.: The nature of statistical learning theory. Springer, N. Y. (1995)
20. Platt, J.C.: Advances in kernel methods, pp. 185–208. MIT Press, Cambridge (1999)