International Journal on Artificial Intelligence Tools

COMBINING LANGUAGE MODELING AND LSA ON GREEK SONG "WORDS" FOR MOOD CLASSIFICATION

KATIA L. KERMANIDIS
IOANNIS KARYDIS
ANTONIS KOURSOUMIS
KAROLOS TALVIS
Department of Informatics, Ionian University
49100, Kerkyra, Greece
{*kerman, karydis, p08kour, p08talv*}*@ionio.gr*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

The present work presents a novel approach to song mood classification. Two language models, one absolute and one relative, are experimented with. Two distinct audio feature sets are compared against each other, and the significance of the inclusion of text stylistic features is established. Furthermore, Latent Semantic Analysis is innovatively combined with language modeling, and depicts the discriminative power of the latter. Finally, song "words" are defined in a broader sense that includes lyrics words as well as audio words, and LSA is applied to this augmented vocabulary with highly promising results. The methodology is applied to Greek songs, that are classified into one of four valence and into one of four arousal categories.

Keywords: music mood classification; lyrics; audio; stylistic features; LSA; 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 organization 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 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].

Mood recognition is addressed in the literature as a classification or regression problem, where an entire song or part of a song is tagged with a mood value. The latter may be binary (denoting the presence or absence of an emotion), categorical (denoting the type of emotion the song belongs to), numerical (a probability estimate or a Likert-type scale value is assigned to each song), one-dimensional (there is one single emotion class), multi-dimensional (emotion is viewed as a multi-dimensional entity defined by vectors such as valence, arousal and tension), or a time-series of vectors (where the emotional value of a song is tracked throughout its duration) [7].

The features reported to have impact on the mood a given song creates vary from audio content (acoustic) to linguistic (lyrics features representing the distribution of words in the song text and collection), stylistic (representing the distribution of punctuation, exclamation, word-, verse-, and song length etc. in the song and the collection) and user tag-based features. Several acoustic feature sets have been experimented with, including timbre-, rhythm-, intensity-, tempo- and pitch-related features (e.g. MFCCs, flatness, etc.). Regarding lyrics features, approaches have adopted bag-of-words models as well as language model differences and latent semantic models for their extraction.

Concerning the utilized learning techniques, these vary from Least Squares and Support Vectors to Gaussian Mixtures and Multiple Linear regression for predicting a numerical mood value, and from memory-based and decision tree, to Multinomial Naive Bayes and Support Vector Machines classifiers for categorizing songs to discrete moods.

1.1. *Contribution & Paper Organisation*

The present work describes the application of supervised learning to various Greek song feature sets for one-dimensional mood classification, taking into account audio, lyrics and stylistic features.

Unlike previous related work in the area, the presented approach

- focuses on Greek songs entirely, that belong to diverse musical genres, e.g. pop, rock, new wave, popular, traditional, folklore, etc.

- makes use of a relatively small song corpus (i.e. 943 songs), posing the challenge of knowledge-poor classification
- experiments with Random Forest classification for the specific task, achieving improved performance compared to the widely used Support Vector Machines, and benefitting from their descriptive power in qualitatively analyzing the various feature sets
- compares different audio feature sets, extracted from the same song data, on the task at hand
- redefines the concept of "song words", i.e. "words" are low-level, generic song elements pertaining to its lyrics, but also to its acoustics and stylistics, proposing thereby a new means to model the song semantics
- applies Latent Semantic Analysis (LSA) to songs and their "words" in order to reduce the semantic space dimensions and detect underlying semantic relations between song "words" and mood categories, as is the case in traditional LSA in Information Retrieval applications, and explores the impact of the number of latent dimensions on classification performance. Unlike previous related approaches that apply LSA to lyrics features only (e.g. word unigrams, bigrams etc.), LSA is applied herein to all song "words", i.e. lyrics as well as audio features, as both are carriers of the song's semantics. The features of both feature types are transferred into latent semantic dimensions
- combines a novel knowledge-poor language modeling approach for extracting lyrics features with LSA in order to further improve classification performance

The rest of the paper is organized 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 is concluded in Section 5.

2. Related Work

Research in mood detection and classification of musical pieces has received extensive attention during most of the last decade, while, since 2007, the Music Information Retrieval Evaluation eXchange (MIREX) evaluation campaign [10] additionally hosts the "Audio Music Mood Classification" task. In this section, we present some of the key approaches in mood classification and regression.

2.1. *Mood Taxonomies*

In order to be able to categorize songs according to their mood, emotion, or else, mood, needs to be modeled. Hevner [8] initially proposed mood groups of certain adjectives. Other mood models based on adjective groups have been proposed since

[9,10]. The most popular mood modeling approach, used extensively in automatic song mood classification [6,11], and serving as the basis of the present work as well, is the model of Thayer [12]. In this model, mood is defined according to 2 dimensions, valence and arousal, dividing thus the emotional 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 (energy-filled) mood states are described by adjectives like "excited", "psyched" etc., while low arousal (energy-lacking) mood corresponds to "sleepy" states. Positive valence indicates emotions like "pleased" and "happy", while examples of negative valence emotions are "nervous", and "sad". For a more fine-grained emotion representation, the division of each axis into 4 separate parts has been proposed, i.e. arousal and valence are categorized into four classes each [6], which is the taxonomy also adopted herein.

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 [13,14,15], i.e. the text is treated as a collection of unordered words, accompanied by their frequency (term frequency - tf). 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 document collection. van Zaanen et al. [6] treat the lyrics that belong to a particular mood as one document and calculate the tfidf score of the words of randomly chosen 10,000 songs and then apply nearest-neighbor learning for classifying songs into 4 or 16 mood classes.

The large size of the vocabulary, in addition to the data sparseness, in the bag-of-words model makes the discrimination process between the words very hard. Also on lyrics, the bag-of-words model leads to moderate performance [14], unless abundant amount of data is available [6]. To overcome this difficulty, approaches have experimented either with dimensionality reduction, or with language modeling techniques.

Regarding dimensionality reduction, Yang and Lee [16] make use of the Harvard General Inquirer in order to transform words into a limited set of psychological features and then apply tree-based learning to the resulting incidence matrix. Other, less knowledge-demanding, approaches achieve dimensionality reduction by applying Latent Semantic Analysis (LSA) to the lyrics text. Laurier et al. [14] apply LSA to the lyrics of last.fm songs and then perform binary classification for 4 mood categories and their negative counterparts (e.g. "angry" - "not angry").

Language modeling pertains to the identification of statistical properties of the text of each mood category. Laurier et al. [14] mine the 100 most frequent terms for each mood category and the 100 most frequent terms for its negative counterpart 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 Stylistics*

It has been shown [17] that stylistic markers are significant for stylometric analysis. Stylistic markers include the use of interjection words, certain punctuation marks (e.g. "!"), and text statistics (vocabulary richness, repetition rate, word and sentence length etc.). In the music domain, stylistic features have been employed for genre classification [15]. A proposed stylistic feature set for mood classification is presented in the thesis of Hu [2].

2.4. *Mood Classification Using Audio & 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 [14,2,11,18,19]. Several approaches have been proposed for the combination of the two modalities: obtaining separate mood predictions from each one and combining them through voting [14] or linear combination [19], training separate models from each modality on each subtask of mood classification (e.g. valence or arousal identification) and merge the outcome [19], and concatenating all features in the same feature space and using thus augmented feature vectors [14]. Laurier et al. [14] conclude that the combination of audio and lyrics features offer an improvement in the overall classification performance for the four Russell categories.

Yang and Lee [18], in one of the earlier works in the field, proposed the combination of lyrics and a number of audio features in order to maximize classification accuracy and minimize the mean error. Nevertheless, the significantly small data corpus (145 songs with lyrics) made the work unsafe to draw solid conclusions from. McVicar et al. [11] explore factors of both audio and lyrics that simultaneously affect the mood of a song.

In a dimensionality reduction framework, Yang et al. [19] extracted a number of low-level acoustic features from a 30-second part of the song and lyrics features, produced by unigram and bigram bag-of-words models, as well as Probabilistic LSA. Therein, songs are classified into four categories following the Russell model [20] to conclude that the use of textual features offers a significant accuracy amelioration of the methods examined. According to the aforementioned model, affect is represented in a two-dimensional space, where one dimension is the metaphor of positive/negative polarity (pleasure/displeasure) and the second of energy presence/absence (arousal/sleepiness). All emotions may be represented on this two-dimensional space, as combinations of values of the aforementioned dimensions. Using a similar methodology, but a different application setting, Logan et al. [21] combined Probabilistic LSA on the lyrics with audio features to detect artist similarity.

Apart from song lyrics, other sources of linguistic information have been experimented with in MIR, such as social tags [22]. 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). Lamere [23] presents a detailed description of the uses of social tags in MIR in general.

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 an automatic 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 purposes of experimentation in this work, two audio feature sets were utilized. The first set, audio1, was obtained using the jAudio application [24] that produces a set of, generic for the purposes of MIR, features. audio1 consists of the following features: 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.

The second set, audio2, was extracted using the MIRtoolbox [25] and the "mirfeatures" routine. This routine, computes a large set of high level features organized along the main musical dimensions of dynamics, rhythm, timbre and tonality. In detail, the features include the mean values of the frame-based RMS, the fluctuation summary with its highest peak and centroid, a frame-based tempo estimation, the attack times of the onsets as well as the envelope curve used for the onset detection, spectral frame-based characteristics such as centroid, brightness, spread, skewness, kurtosis, roll-off (using 95% and 85% threshold), entropy, flatness, roughness and irregularity, frame-decomposed MFCCs, frame-decomposed delta-MFCCs, frame-decomposed delta-delta-MFCCs and timbre frame-decomposed characteristics such as zero-crossing rate, low energy rate, spectral flux, unwrapped chromagram with its highest peak and centroid, key clarity, mode and HCDF.

3.2. *Lyrics Features*

Lyrics text, especially in the dataset used in the present approach that includes all genre categories (from rock to ethnic and folklore songs as well), is highly problematic (i.e. it contains linguistic errors, disfluencies, uncommon word forms, truncations, etc.). To overcome these difficulties, the song lyrics underwent a series of pre-processing steps that included:

- removal of punctuation marks, symbols, exclamations, apostrophes and commas
- dealing manually with truncated words (e.g. "μου είπες"- "you told me" written as "μου ΄πες"), separate words that are erroneously joined together (e.g. "εγώ πήγα" - "I went" 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", "Ζωή" - "Life" written 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 mood 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 [26]

Bag-of-words For the lyrics features extracted using 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 (after stop word removal). Unlike previous approaches that represent each song as an incidence vector with as many dimensions as the collection vocabulary size, the aforementioned model is adopted herein in an attempt to address the sparse data problem. Almost 3% of the songs had a vocabulary of less than 20 (distinct) words in total, and less than 1% of the songs had a vocabulary of less than 15 words. Each word is accompanied by its frequency in the song and its tfidf score. The total number of linguistic features in this approach is 60.

Language Modeling Unlike the work by Laurier et al. [14], 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. To this end, the 50 most frequent words in the lyrics of a given category are computed, leading to a total of 218 words (excluding duplicates) words for the eight (four valence and four arousal) categories. This constitutes the absolute language model (ALM) (in contrast to the language model distances approach proposed by Laurier et al. [14]), i.e. 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, again in an attempt to address the sparse data problem. Each of these terms constitutes a linguistic learning feature, and its value is the tfidf metric of the given term in the given song (ALM tfidf). Following the rationalization of Zaanen et al. [6], who claim that tf is a good backoff estimate in the case of a very small idf value, a second ALM dataset was constructed that, apart from the tfidf values, also contains the tf values of a word in a song (ALM tf+tfidf), i.e. has twice the number of lyrics features.

It was interesting to observe in the ALM, that, from the 50 most frequent words in each mood category, that, the 20 top ranked words were shared among most of the categories, and therefore showed low discriminative power, i.e. they were weak indicators of the differences among the categories. The words below a certain rank position (30th) appeared in only one, at most two categories, so their discriminative power was clearer and they could better distinguish one category from the other. Therefore experiments were run taking into account only the 30 last of the 50 most frequent terms of each category (relative language model-RLM), leading to a total of 240 words as linguistic features. Again, two datasets were constructed, one containing only the tfidf values of the lyrics features (RLM tfidf), and one containing also their tf values (RLM tf+tfidf)

Latent Semantic Analysis LSA [27] is a matrix singular value decomposition (SVD) technique, initially proposed for reducing the size of the term-document matrix in information retrieval (IR) applications. SVD decomposes the initial matrix A into a product of three matrices and "transfers" matrix A into a new semantic space:

$$A = TSD^T \tag{1}$$

T is the matrix with rows the lexicon terms, and columns the dimensions of the new semantic space. The columns of D represent the initial documents and its rows the new dimensions, while S is a diagonal matrix containing the singular values of A. Multiplication of the three matrices will reconstruct the initial matrix. The product can be computed in such a way that the singular values are positioned in S in descending order. The smaller the singular value, the less it affects the product outcome. By maintaining only the first few (k) singular values, setting the remaining ones to zero and calculating the resulting product, a low-rank approximation A_k of the initial matrix A may be calculated as a least-squares best fit. The reduced number of dimensions of the new matrix is equal to the number k of selected singular values.

As an interesting side effect, dimensionality reduction reduces or increases the frequency of words in certain documents, or may even set the occurrence of words to higher than zero for documents that they initially did not appear in. Thereby semantic relations between words and documents are revealed that were not apparent at first (latent). It needs to be noted that LSA is fully automatic, i.e. the latent semantic relations are learned in an unsupervised manner.

Herein LSA is applied in two different ways. First, in two experiments, it is applied in the "traditional" way to the ALM (LSA on ALM) and RLM (LSA on RLM) lyrics datasets. Songs (rows) and words (columns) form the initial term-document matrix and several dimensionality reduction ratios are tested (from 30 to 200 latent dimensions). Audio features are

then simply appended to the latent dimensions (LSA lyrics features). In another pair of experiments, "words" are defined in a broader sense as generic items, each carrying part of the song semantics. A similar definition of words is applied to game modeling [28]. In this sense, "words" do not only include text words from the songs' lyrics, but are also carriers of musical information, i.e. the audio features are treated as "words", and LSA is applied to the augmented feature vectors that contain the ALM and the audio features (LSA on both ALM and audio features) or the RLM and the audio features (LSA on both RLM and audio features), after normalizing the feature values. Like before, several dimensionality reduction ratios are tested. In both cases, LSA is combined for the first time, to the authors' best knowledge, with language modeling in the same dataset for the task at hand.

3.3. *Stylistic Features*

The 15 stylistic features experimented with in the present work are based on the feature set proposed by Hu and Downie [2]. They include the number of interjections in the song text, the size of the song lyrics in words, the number of unique words, the word repetition ratio, the average word length, the number of lines, the number of unique lines, the number of blank lines, the blank line ratio, the average line length, the standard deviation of the line length, the number of unique words per line, the line repetition ratio, the average word repetition ratio per line and the standard deviation of the word repetition ratio per line.

4. Experimental Results

As mentioned earlier, in our experimental setup, each song is represented as a feature-value vector and constitutes a learning example. Each learning example is to be classified into one of four valence and into one of four arousal classes (i.e. mood recognition is viewed as two separate classification tasks). All experiments were run using 10-fold cross validation.

4.1. *Experimental Setup*

The dataset utilized in this work consists of 943 Greek songs from various genres that include lyrics collected from http://www.stixoi.info. The songs were manually annotated (assigned a valence and an arousal value) by three listeners. Every annotator listened to at least one verse and one chorus of every verse. In case of diverse annotations (occurring in approximately 8% of the annotations) majority voting and/or discussion followed to ensure agreement.

Experiments were run using the Weka Machine Learning Workbench [29]. In an earlier version of the presented approach [30] several learning algorithms (nearest neighbors, Naive Bayes, decision trees, random forests, support vector machines)

were experimented with for investigative purposes and the random forest classifier turned out to achieve the best performance. Regarding the learning parameters, a number of 100 trees were chosen to from the forest and 100 randomly chosen features were taken into account each time for creating a tree. Random forests have been used previously for the task at hand [14], also with very satisfactory results. Moreover, the descriptive power of tree-based learners is of great importance, as they encode qualitative information regarding learning features and learning examples.

Regarding the stemming process, its negative impact is described and explained in detail in the preliminary experiments of the work [30]. Stemming increases the term frequency range (stem frequency increases accumulatively from the word forms that belong to it), making the learning process more difficult. Regarding the stemmer itself, its accuracy is bounded; several errors occur by assigning different lemmata erroneously to the same stem, partly attributed to the tool, and partly to the idiosyncratic morphology of the Modern Greek language (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. For all aforementioned reasons, only results with the unstemmed version of the lyrics features are presented herein.

The bag-of-words model [30], reaches valence and arousal accuracy values between 35-50%, much lower than the language model results. The superiority of the language model approach is evident, and partly attributed to the nature of the numerical features involved in the ALM and RLM datasets, and the lack of nominal word-based features that take many unique values, and that are present in the bag-of-words dataset. But mostly, it is attributed to the discriminative power of the features. Thus, the results included herein are only with the language model datasets.

4.2. *Results and Discussion*

Figure 1 shows the results with the raw datasets (prior to LSA). The improved results of the ALM datasets may partly be attributed to the inclusion of frequent words, across all categories. This commonality helps make the learning process easier. Interestingly, the datasets that include term tf values, achieve improved results compared to the ones that include only tfidf. Even though the linguistic features double in number in the tf+tfidf datasets (complicating thus the learning process), the improved accuracy results show the importance of the tf backoff process mentioned earlier. Regarding the two audio feature sets, audio2 seems to outperform audio1 in most cases. This result can be attributed to the higher-level character of the features included within audio2 dataset in contrast to the low-level characteristics described by the features of audio1. As aforementioned, the nature of the

features selected in order to represent the data is of paramount importance to the processes discussed herein. Thus, the more specialised character towards MIR of the features in audio2 outperform the generic audio processing features of audio1.

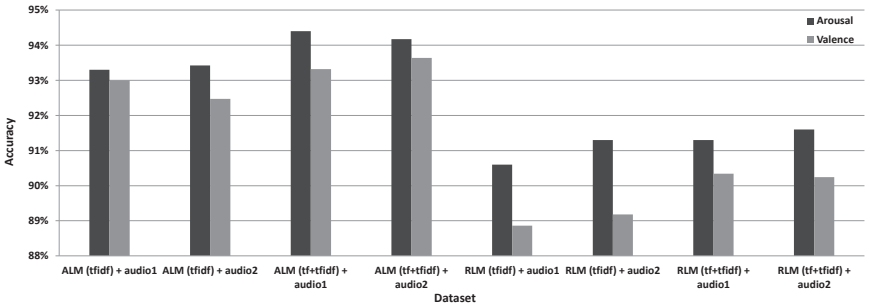

Figure 1. Arousal and valence accuracy with all combinations of lyrics and audio datasets.

Figure 2 shows the results with the datasets that include all three modalities, i.e. linguistic, audio and stylistic features. Despite the increased number of features, accuracy is, in several cases, higher compared to the bimodal experiments. The significance of the stylistic features to mood classification is thereby indicated.

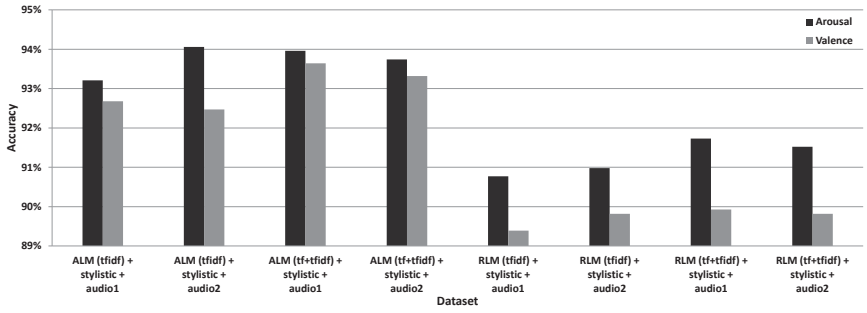

Figure 2. Arousal and valence accuracy with all combinations of lyrics, stylistic and audio datasets.

The first results with LSA (Figure 3) correspond to the traditional manner of its application, i.e. only to the lyrics words. audio1 features are simply concatenated. Results are quite low, and therefore, only results with 60 latent semantic dimensions, tfidf values and audio1 features are shown.

Applying LSA to "words", leads to much higher accuracy values, as can be seen in the following graphs.

Figures 4 and 5 show arousal and valence accuracy values when applying LSA to both RLM linguistic and audio features, for various dimensionality reduction ratios.

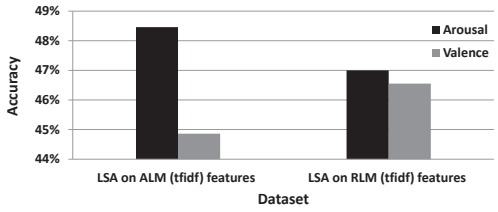

Figure 3. Arousal and valence accuracy with LSA on lyrics and concatenated audio1 features.

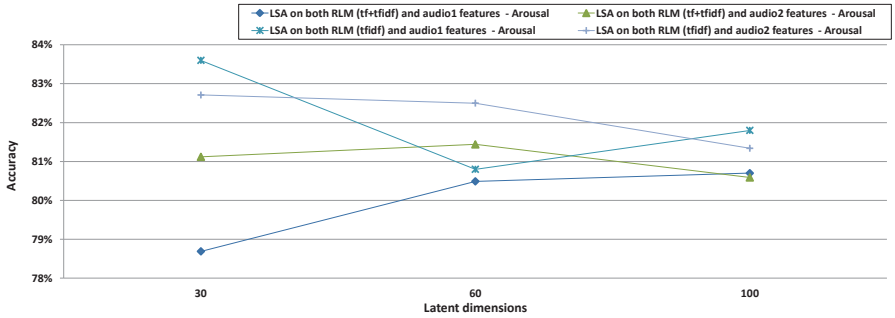

Figure 4. Arousal accuracy with LSA on both RLM lyrics and audio features for 30, 60 and 100 latent dimensions.

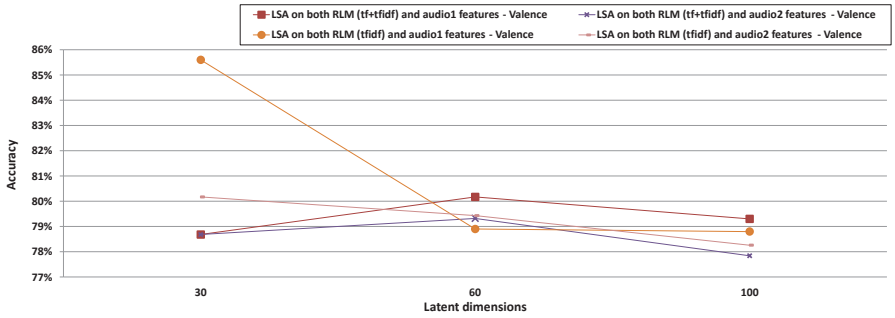

Figure 5. Valence accuracy with LSA on both RLM lyrics and audio features for 30, 60 and 100 latent dimensions.

In most cases, 60 latent semantic dimensions achieve the highest score, fewer dimensions seem to be less able to capture the song semantics, more dimensions seem to be superfluous, redundant and thus sometimes misleading. Figures 6 and 7 show the corresponding results with the ALM datasets. Compared to the raw datasets results are lower, due to the more abstract numeric data values, that are not limited in number and in range, like tf and tfidf. However, the primary goal, when applying LSA, is not the increase of classification accuracy (rarely do latent features in literature improve classification performance), but the discovery of hidden semantic

information regarding the relations between learning features and instances. For example, when depicting the learning instances of the LSA-RLM-audio1 dataset with 30 latent semantic dimensions in a two-dimensional plot, where one axis (e.g. the horizontal) is the first latent dimension and the second (e.g. the vertical) is the second latent dimension, it is interesting to observe how songs are grouped together. The upper left corner of the plot consists of songs that are of more modern acoustics, younger artists, mostly hip-hop-like, alternative language, and underground in genre, while the lower right corner is formed by sounds more common to the Greek music history and mainstream artists of the Greek popular music culture. This underlying semantics has been revealed through LSA and was not observable beforehand.

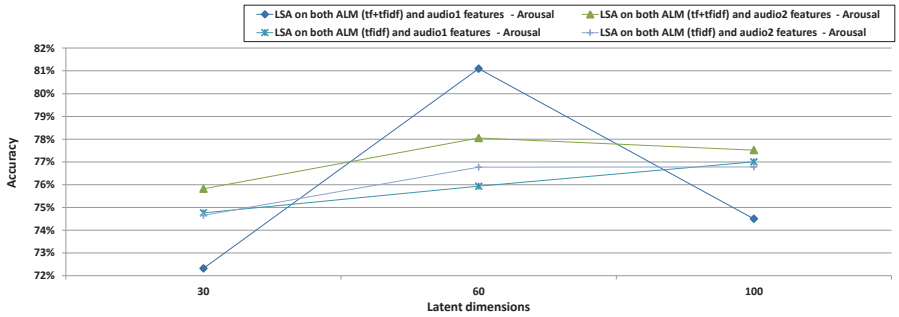

Figure 6. Arousal accuracy with LSA on both ALM lyrics and audio features for 30, 60 and 100 latent dimensions.

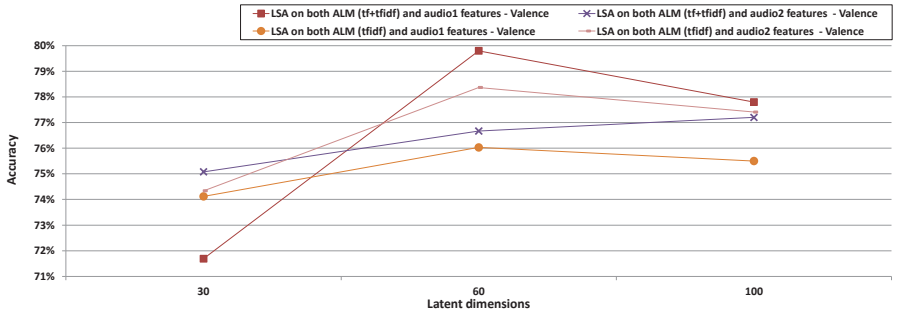

Figure 7. Valence accuracy with LSA on both ALM lyrics and audio features for 30, 60 and 100 latent dimensions.

For a more informative representation of results, Figure 8 shows precision and recall scores for every class label for some indicative datasets.

It is interesting to observe how the application of LSA swaps the performance between ALM and RLM. Due to the dimensionality reduction process, LSA is not so

Dataset		Arousal			Valence	
		Precision	Recall		Precision	Recall
LSA on both RLM(tfidf) and audio1 features	A	0,87	0,63	1	0,762	0,833
	B	0,753	0,918	2	0,83	0,679
	C	0,841	0,79	3	0,804	0,736
	D	0,881	0,783	4	0,798	0,822
LSA on both RLM(tfidf) and audio2 features	A	0,83	0,654	1	0,756	0,827
	B	0,801	0,904	2	0,842	0,737
	C	0,816	0,839	3	0,829	0,764
	D	0,881	0,788	4	0,802	0,802
RLM(tfidf)+audio2	A	0,897	0,819	1	0,887	0,906
	B	0,898	0,956	2	0,898	0,898
	C	0,917	0,895	3	0,854	0,854
	D	0,928	0,907	4	0,918	0,896
ALM(tfidf)+audio2	A	0,896	0,882	1	0,9	0,955
	B	0,965	0,956	2	0,976	0,905
	C	0,918	0,952	3	0,905	0,91
	D	0,928	0,912	4	0,944	0,909
RLM(tfidf)+stylistics+audio1	A	0,855	0,835	1	0,858	0,912
	B	0,904	0,959	2	0,938	0,876
	C	0,932	0,883	3	0,849	0,882
	D	0,919	0,898	4	0,95	0,889
RLM(tfidf)+stylistics+audio2	A	0,914	0,835	1	0,899	0,912
	B	0,896	0,953	2	0,904	0,891
	C	0,899	0,899	3	0,872	0,876
	D	0,944	0,898	4	0,912	0,899

Figure 8. Precision and recall for every class value for some indicative datasets.

sensitive to the limited number of tf and tfidf values and presence of the same terms in many feature vectors (the commonality mentioned earlier) that help achieve the high results with the ALM datasets. Thereby, LSA helps identify the discriminative power of the less common words in the RLM datasets.

Results are highly satisfactory, even when compared to approaches that are knowledge-demanding (employ more sophisticated morphological information, like part-of-speech tags of the lyrics words, or semantic thesauri, like Wordnet-Affect, the General Inquirer dictionary etc.). For a rough comparison with techniques utilizing similar resources, Yang et al. [19], for example, report a valence classification accuracy of 74.8% without and 72.9% with PLSA. van Zaanen and Kanters [6] report 77.2% arousal accuracy and 76.3% valence accuracy using tf and tfidf metrics for the lyrics features.

Figure 9 shows statistical significance between the accuracy results, by pairing one-by-one 20 datasets. The two-tailed t-test with a 0.05 confidence level has been employed for performing the significance tests. In the Figure the first row indicates that the accuracies of datasets m and n are significantly better than the accuracy of dataset a (cell value 1), while all other dataset accuracies are not significantly

better than the accuracy of dataset a (cell values 0). Statistical significance results were obtained using the WEKA Experimenter mode and running ten-fold cross validation experiments 10 times each, a total of 100 runs.

Dataset	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
a=RLM(tf+tfidf)+audio1	-	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
b=RLM(tfidf)+audio1	1	-	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
c=RLM(tfidf)+audio2	1	0	-	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
d=RLM(tf+tfidf)+audio2	1	0	0	-	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
e=LSAonRLM(tfidf)+audio2-100dim	1	1	1	1	-	1	1	0	0	0	1	1	1	1	0	0	0	1	0	0
f=LSAonRLM(tfidf)+audio2-30dim	1	1	1	1	0	-	1	0	0	0	1	1	1	1	0	0	0	1	0	0
g=LSAonRLM(tfidf)+audio2-60dim	1	1	1	1	0	0	-	0	0	0	1	1	1	1	0	0	0	1	0	0
h=LSAonALM(tfidf)+audio2-60dim	1	1	1	1	1	1	1	-	0	1	1	1	1	1	0	1	0	1	0	1
i=LSAonALM(tfidf)+audio2-30dim	1	1	1	1	1	1	1	1	-	1	1	1	1	1	1	1	0	1	1	1
j=LSAonALM(tfidf)+audio2-100dim	1	1	1	1	1	1	1	0	0	-	1	1	1	1	0	1	0	1	0	1
k=RLM(tfidf)+Stylistic+audio1	1	0	0	1	0	0	0	0	0	0	-	0	1	1	0	0	0	0	0	0
l=RLM(tfidf)+Stylistic+audio2	1	0	0	1	0	0	0	0	0	0	1	-	1	1	0	0	0	0	0	0
m=ALM(tfidf)+Stylistic+audio1	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0
n=ALM(tfidf)+Stylistic+audio2	0	0	0	0	0	0	0	0	0	0	0	0	1	-	0	0	0	0	0	0
o=LSAonALM(tfidf)+audio1-100dim	1	1	1	1	1	1	1	1	0	1	1	1	1	1	-	1	0	1	1	1
p=LSAonRLM(tfidf)+audio1-100dim	1	1	1	1	1	1	1	0	0	0	1	1	1	1	0	-	0	1	0	1
q=LSAonALM(tfidf)+audio1-30dim	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	1	1	1
r=LSAonRLM(tfidf)-30dim	1	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	-	0	0
s=LSAonALM(tfidf)+audio1-60dim	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	-	1
t=LSAonRLM(tfidf)+audio1-60dim	1	1	1	1	1	1	1	0	0	0	1	1	1	1	0	0	0	1	0	-

Figure 9. Statistical significance results between 20 datasets.

The ALM model with tfidf, audio1 and stylistic features (dataset m) significantly outperforms all others. When LSA is not applied, using both tf and tfidf metrics for representing lyrics features significantly outperforms using only tfidf. Results with LSA and RLM are significantly better than those with LSA and ALM. In most cases, results with audio2 features significantly outperform those with audio1 features.

5. Conclusion

In this work, a novel approach to automatic mood classification is presented that combines the application of LSA and language modeling to songs' data. Songs' data include audio, lyrics and stylistic data, and in an innovative broad definition of song "words", where the augmented vocabulary includes, apart from lyrics words, also audio words. The methodology is applied to Greek songs of varying genre, and Random Forests are chosen to classify songs into four valence and four arousal categories.

The impact of the dimensionality reduction ratio on classification performance is explored, two different audio feature sets are experimented with for comparison purposes, and the ability of LSA to identify and reveal the discriminative strength of the relative language model compared to the absolute language model is established.

A dataset that is comprised of more song data, more sophisticated text preprocessing techniques, e.g. the inclusion of part-of-speech information or bigram lyrics features, more fine-grained mood classification into more classes, as well as viewing mood recognition as a multi-class classification task, and applying LSA to a

multi-class experiment are challenging future research directions

Bibliography

1. Calvin K. M. Lam and Bernard C. Y. Tan. The internet is changing the music industry. *Communications ACM*, 44(8):62–68, 2001.
2. Xiao Hu and J. Stephen Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of Joint Conference on Digital Libraries*, pages 159–168, 2010.
3. Don Byrd. Organization and searching of musical information, course syllabus, 2008.
4. Erik M. Schmidt and Youngmoo E. Kim. Prediction of time-varying musical mood distributions from audio. In *Proceedings of International Society for Music Information Retrieval*, pages 465–470, 2010.
5. Brian McFee and Gert R. G. Lanckriet. The natural language of playlists. In *Proceedings of International Society for Music Information Retrieval*, pages 537–542, 2011.
6. Menno van Zaanen and Pieter Kanters. Automatic mood classification using tf*idf based on lyrics. In *Proceedings of International Society for Music Information Retrieval*, pages 75–80, 2010.
7. Schmidt E. Migneco R. Morton B.G. Richardson P. Scott J. Speck J. A. Kim, Y.E. and Turnbull D. Music emotion recognition: A state-of-the-art review. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 255–266, 2010.
8. K. Hevner. Experimental studies of the elements of expression in music. *Proceedings of American Journal of Psychology*, 48(2):246–267, 1936.
9. D. Grandjean M. Zentner and K. R. Scherer. Emotions evoked by the sound of music: Characterization, classification, and measurement. 8(4):494–521, 2008.
10. C. Laurier M. Bay X. Hu, J. Downie and A. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of International Conference on Music Information Retrieval*, 2008.
11. Matt McVicar, Tim Freeman, and Tijl De Bie. Mining the correlation between lyrical and audio features and the emergence of mood.
12. R.E. Thayer. *The biopsychology of mood & arousal*. Oxford University Press, 1989.
13. Xiao Hu, J. Stephen Downie, and Andreas F. Ehmann. Lyric text mining in music mood classification. In *Proceedings of International Society for Music Information Retrieval*, 2009.
14. Cyril Laurier, Jens Grivolla, and Perfecto Herrera. Multimodal music mood classification using audio and lyrics. In *Proceedings of International Conference on Machine Learning and Applications*, pages 688–693, 2008.
15. A. Rauber R. Mayer, R. Neumayer. Rhyme and style features for musical genre classification by song lyrics. In *Proceedings of International Conference on Machine Learning and Applications*, pages 337–342, 2008.
16. Dan Yang and Won-Sook Lee. Music emotion identification from lyrics. In *Proceedings of IEEE International Symposium on Multimedia*, pages 624–629, 2009.
17. M. Argamon, S. abd Saric and S. S. Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 475–480, 2003.
18. Dan Yang and Won-Sook Lee. Disambiguating music emotion using software agents. In *Proceedings of International Conference on Music Information Retrieval*, 2004.
19. Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and

Homer H. Chen. Toward multi-modal music emotion classification. In *Proceedings of Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, pages 70–79, 2008.

20. J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.
21. A. Kositsky B. Logan and P. Moreno. Semantic analysis of song lyrics. In *Proceedings of IEEE International Conference on Multimedia and Expo*, volume 2, pages 827–830, 2004.
22. Panagiotis Symeonidis Alexandros Nanopoulos, Dimitrios Rafailidis and Yannis Manolopoulos. Musicbox: Personalized music recommendation based on cubic analysis of social tags. 18(2):407–412, 2010.
23. P. Lamere. Social tagging and music information retrieval. 37(2):101–114, 2008.
24. D. McEnnis, C. McKay, and I. Fujinaga. jAudio: A feature extraction library. In *Proceedings of International Conference on Music Information Retrieval*, 2005.
25. Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of International Conference on Digital Audio Effects*, 2007.
26. Spyridon Saroukos. Enhancing a greek language stemmer - efficiency and accuracy improvements. Master's thesis, Dept. of Computer Sciences, University of Tampere, Finland, 2008.
27. P. Foltz T. Landauer and D. Laham. An introduction to latent semantic analysis. 25:259–284, 1998.
28. Katia Lida Kermanidis, Panagiotis Pandis, Costas Boletsis, and Dimitra Chasanidou: LSA for Mining Hidden Information in Action Game Semantics. Proceedings of european conference on artificial intelligence. In *Proceedings of European Conference on Artificial Intelligence*, pages 967–968, 2012.
29. G. Holmes, A. Donkin, and I. H. Witten. Weka: A machine learning workbench. In *Proceedings of Intelligent Information Systems*, pages 357–361, 1994.
30. Spyros Brilis, Evangelia Gkatzou, Antonis Koursoumis, Karolos Talvis, Katia Lida Kermanidis, and Ioannis Karydis. Mood classification using lyrics and audio: A case-study in greek music. In *Proceedings of Artificial Intelligence Applications and Innovations Conference*, pages 421–430, 2012.