Keyword-Based Approach for Lyrics Emotion Variation Detection

Ricardo Malheiro[1,2], Hugo Gonçalo Oliveira[1], Paulo Gomes[1] and Rui Pedro Paiva[1]
[1]*Center for Informatics and Systems of the University of Coimbra (CISUC), Pólo II, Pinhal de Marrocos 3030-290 Coimbra, Portugal*
[2]*Miguel Torga Higher Institute, Largo Cruz de Celas 1, 3000-132 Coimbra, Portugal*
{rsmal, panda, pgomes, ruipedro}@dei.uc.pt

Keywords: Music Information Retrieval, Lyrics Music Emotion Recognition, Lyrics Music Emotion Variation Detection, Keyword Based Approach.

Abstract: This research addresses the role of the lyrics in the context of music emotion variation detection. To accomplish this task we create a system to detect the predominant emotion expressed by each sentence (verse) of the lyrics. The system employs Russell's emotion model and contains 4 sets of emotions associated to each quadrant. To detect the predominant emotion in each verse, we propose a novel keyword-based approach, which receives a sentence (verse) and classifies it in the appropriate quadrant. To tune the system parameters, we created a 129-sentence training dataset from 68 songs. To validate our system, we created a separate ground-truth containing 239 sentences (verses) from 44 songs annotated manually with an average of 7 annotations per sentence. The system attains 67.4% F-Measure score.

1 INTRODUCTION

Music emotion recognition (MER) is gaining significant attention in the Music Information Retrieval (MIR) scientific community. In fact, the search of music through emotions is one of the main criteria utilized by users (Vignoli, 2004).

Real-world music databases from sites like AllMusic or Last.fm grow larger and larger on a daily basis, which requires a tremendous amount of manual work for keeping them updated. Unfortunately, manually annotating music with emotion tags is normally a subjective, expensive and time-consuming task. This should be overcome with the use of automatic recognition systems (Hu and Downie, 2010).

Most of the early-stage automatic MER systems were based on audio content analysis (e.g., (Lu et al., 2006)). Later on, researchers started combining audio and lyrics, leading to bi-modal MER systems with improved accuracy (e.g., (Hu and Downie, 2010), (Hu et al., 2009), (Laurier et al., 2008)). This does not come as a surprise since it is evident that the importance of each dimension (audio or lyrics) depends on music style. For example, in dance music audio is the most relevant dimension, while in poetic music (like Jacques Brel) lyrics are key.

Several psychological studies confirm the importance of lyrics to convey semantical information. Namely, according to Juslin and Laukka (2004), 29% of people mention that lyrics are an important factor of how music expresses emotions. Also, Besson et al. (1998) have shown that part of the semantic information of songs resides exclusively in the lyrics.

Each song is normally associated to a predominant emotion (e.g., happiness, sadness), which corresponds to the emotion perception of the listeners concerning that song. Music Digital Libraries (MDL) like AllMusic take this into account to classify songs in their sites.

There are songs in which the predominant emotion is easy to determine, i.e., for the majority of listeners the perceived emotion is the same or almost the same throughout the song, while in others the perceived emotion varies significantly along the song. The example below, from the song "Kim" by Eminem, illustrates emotion variation:

Aw look at daddy's baby girl
That's daddy baby
Little sleepy head
Yesterday I changed your diaper
Wiped you and powdered you.
How did you get so big?
Can't believe it now you're two

Baby you're so precious
Daddy's so proud of you

Sit down bitch
If you move again I'll beat the shit out of you
Don't make me wake this baby
She don't need to see what I'm about to do
Quit crying bitch, why do you always make me shout at you?
…

The lyric changes abruptly from emotions like serene joy and relaxation to anger and tension.

Thus it is important to investigate the time-varying relationship between music and emotion.

In the audio domain, there are a few studies tackling Music Emotion Variation Detection (MEVD), e.g., (Schubert, 1999), however, to the best of our knowledge, we are not aware of any research focused in this specific area of Lyrics Music Emotion Variation Detection (LMEVD).

In this work, we propose a novel a keyword-based approach (KBA) to classify song verses according to Russell's emotion model (Russell, 1980). To validate our model, we create a new manually annotated dataset with 239 sentences taken from 44 song lyrics.

This paper is organized as follows. In section 2, the related work is described and discussed. Section 3 presents the methods employed in this work, particularly the creation of the ground truth and the description of the architecture of our KBA model. The results attained by our system are presented and discussed in Section 4. Finally, section 5 summarizes the main conclusions of this work.

2 RELATED WORK

The relations between emotions and music have been a subject of active research in music psychology for many years. Different emotion paradigms (e.g., categorical or dimensional) and taxonomies (e.g., Hevner, Russell) have been defined (Hevner, 1936), (Russell, 1980) and exploited in different computational MER systems.

One of the most well-known dimensional models is Russell's circumplex model (Russell, 1980), where emotions are positioned in a two-dimensional plane comprising two axes, designated as valence (V) and arousal (A), as illustrated in Figure 1. According to Russell (Russell, 2003), valence and arousal are the "core processes" of affect, forming the raw material or primitive of emotional experience. We use in this research a categorical version of this Russell's model, so we consider that a sentence belongs to quadrant 1 if both dimensions are positive; quadrant 2 if V is smaller than 0 and A is bigger than 0; quadrant 3 if both dimensions are negative and quadrant 4 if V is bigger than 0 and A is smaller than 0. The main emotions associated to each quadrant are illustrated in Figure 1.

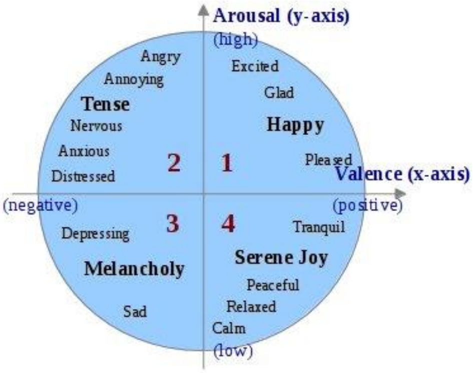


Figure 1: Main screen of the annotation platform.

Identification of musical emotions from lyrics is still in an embryonic stage. Most of the previous studies related to this subject used general text instead of lyrics and polarity detection instead of emotion detection. More recently, Lyrics MER (LMER) has gained significant attention in the MIR scientific community.

Each song is normally associated to a predominant emotion (e.g., happiness, sadness), which corresponds to the emotion-perception of the listeners concerning to that song. Music Digital Libraries (MDL) like AllMusic take this into account to classify songs in their sites.

Human perception of the emotions expressed by a song depends normally on several dimensions which compose a song (e.g., audio, lyrics).

In the audio domain, a few works have addressed MEVD. Namely, Schubert (Schubert, 1999) proposes a time series analysis method and Korhonen (Korhonen et al., 2006) tackles MEVD as a system identification method, exploiting the temporal information among the music segments. Other authors, e.g., (Yang et al., 2006) do not integrate temporal information and perform emotion prediction independently for each music segment.

Concerning Lyrics MEVD, we are not aware of any research of this type.

According to Chopade (2015), emotions may be expressed in lyrics by one word or a bunch of words. The sentence level emotion detection method plays a crucial role to trace emotions or to search out the cues for generating such emotions. Sentences are the basic information units of any document. For that reason, the document level emotion detection method depends on the emotion expressed by the individual sentences of that document that successively relies on the emotions expressed by the individual words.

According to Oxford Dictionaries (http://www.oxforddictionaries.com), a verse is a group of lines that form a unit in a poem or a song. Thus, regarding the typical structure of a lyric (based on verses as in poetry or based on sentences as in prose), we think composers convey ideas and emotions having, as basic unit of information, respectively the verses and the sentences. In our work we use interchangeably the terms sentence and verse.

There are basically three types of approaches to work with the task of detection of emotions in text (Binali et al., 2010):

- **Learning-based approaches (LBA)**. LBA is based on the use of a trained classifier to categorize input text into emotion classes by using keywords as features. To adapt to a new domain we have to supply a large training set to a machine learning algorithm to build a new classification model. Thus, we use the features extracted from the corpora. Here, the more difficult step is normally acquiring the corpora, e.g., (Yang et al., 2007).
- **Keyword-based approaches (KBA)**. KBA is based on the presence of keywords in text. It typically involves steps such as pre-processing with a parser and search based on an emotion dictionary. This technique is domain specific, relies on the presence of keywords for accurate results and requires pre-processing for improved accuracy results, e.g., (Chunling et al., 2005), (Hancock et al., 2007) and (Li et al., 2007).
- **Hybrid approaches (HA)**. Hybrid approaches are a combination of the previous methods. These approaches can improve results from training a combination of classifiers and adding knowledge-rich linguistic information from dictionaries and thesauri, e.g., (Aman and Szpakowicz, 2007), (Binali et al., 2010) and (Kao et al., 2009).

Some authors, e.g., (Chopade, 2015), consider lexicon-based approaches (which counts the number of words of a lexicon in the text) as a 4th independent approach, while others, e.g., Binalli et al., (2010), consider this approach as part of a KBA.

In our work, we use a KBA to detect emotions in sentences and, then, to understand how the emotions vary along the lyric.

There are some limitations associated normally to KBA, which we attempt to mitigate.

1) Ambiguity in keyword definitions, i.e., the meanings of keywords could be multiple and vague, as most words could change their meanings according to different usages and contexts. Our system performs disambiguation to some extent, since it retrieves the definitions of the words from Wordnet (WN) (Miller, 1995) and counts on their words to the emotion detection task. If we have for instance the word "crush" in "he had a crush on her", applying POS tags, "crush" is a noun and its definition from WN is "temporary love of an adolescent". If we have the same word in the sentence "He crushed the car", crushed here is a verb and the definition is "break into small pieces". Probably this will not work in all situations, even because WN may have more than one definition for each grammatical class (e.g., noun). We consider the most common case. Our system retrieves also from the WN synonyms of the words and the same happens here, i.e., depending on the grammatical class the synonyms list is different.
2) Emotions are recognized only in the presence of keywords. In our work, the retrieved synonyms and definitions help to extent our keyword list.

3 METHODS

3.1 Dataset Construction

To accomplish emotion variation detection based on song lyrics, we need a ground-truth composed of annotated sentences (verses). We consider the sentence as the basic unit for the lyric. Hence, through the variation of emotions along several consecutive sentences, we can observe the way the emotions vary along the lyric.

3.1.1 Validation Set

Data Collection and Pre-Processing

To construct our validation dataset, we collected 44 song lyrics, belonging to several genres. Musical genres are distributed as follows: pop/rock (6 songs), pop (18 songs), rock (8 songs), heavy-metal (3 songs), folk (2 songs), R&B (1 song), hip-hop (4 songs) and country (2 songs).

In the selection of the songs, we tried that the songs were distributed uniformly for the 4 quadrants of the Russell's emotion model, according to our a priori perception (11 for each quadrant).

The obtained lyrics were then pre-processed to improve their quality. Namely, we performed the following tasks:

- Correction of orthographic errors;
- Elimination of text not related with the lyric (e.g., names of the artists, composers, instruments);
- Elimination of common patterns in lyrics such as [Chorus x2], [Vers1 x2], etc.;
- Complementation of the lyric according to the corresponding audio (e.g., chorus repetitions in the audio are added to the lyrics).

Annotation and Validation

To simplify the sentence annotation process, we decided to create a web application in the Google App Engine. This app was disclosed for the annotators through direct invitations, mailing lists and social networks.

Initially, the annotators have to register in the web application and then confirm the email sent by the application for their emails. The session starts after authentication. The following items shows some characteristics of the platform:

- The start-up screen shows information about the goals of the research and instructions to accomplish the task;
- The sentences are presented randomly to the annotators;
- The same sentence does not appear twice for the same annotator, even in different sessions;
- If a song has several repetitions of the same sentence (e.g., chorus), the sentence only appears once to the annotator;
- The annotator can continue his work in different sessions;
- The annotator can classify any number of sentences;
- If the annotator classifies all the sentences in the database, the system shows, at the end, a message saying that there are no more sentences to annotate.

Figure 2 shows the application interface. The annotator should read the sentence and then pick the most appropriated choice with the mouse in the pie chart.

If the user hovers with the mouse the several regions in the pie chart (e.g., Q1, Q2, Q3, Q4), the system shows the most predominant emotions from that quadrants.

Finally, the application provides instructions on how to correctly perform the task:

1. Read the sentence with attention;
2. Try to identify the basic predominant emotion expressed by the sentence, according to the sets of emotions (quadrants) in the figure (Figure 3);
3. If you think the sentence does not convey any emotion, select the option Neutral.

To further improve the quality of the annotations, the users were recommended not to use any known previous knowledge about the lyric when they recognized the song through the sentence, not to search for information about the lyric neither the song on the Internet or another place and to avoid tiredness by taking a break and continuing later.

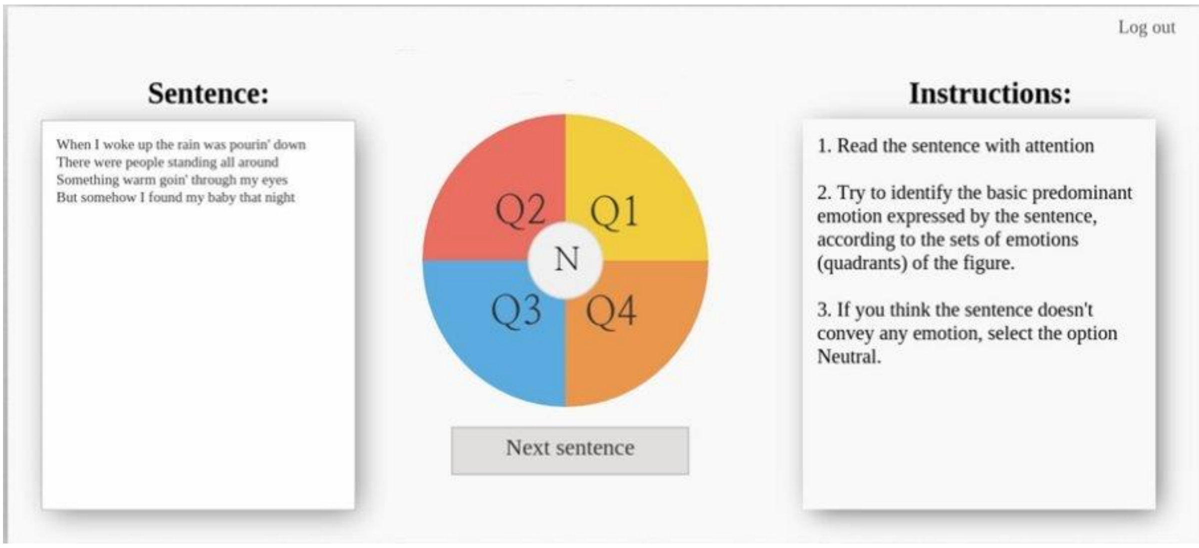
The 44 employed lyrics have a total of 330 sentences and we obtained an average of 7 annotations per sentence.

The classification of each sentence corresponds to the most representative class among all the annotations. In case of a draw the sentence is ignored. This situation happened in 9 sentences.

Since our goal is to build a system to classify sentences in 1 of the 4 possible quadrants, we ignore the sentences annotated as neutral sentences, which happened 18 times. In the future we intend to expand our model to detect previously if a sentence is emotional or non-emotional.

Additionally, we also ignore the repetitions of verses and chorus, that is, we consider only one occurrence of each repeated section. This excludes more 64 sentences. So, at the end, we obtained 239 sentences in total (330 – 9 – 18 – 64).

The following examples illustrate the process of annotation for some of these sentences: 1) the sentence "*I've got peace like a river, I've got peace like a river in my soul*" from the song "*Peace like a river*" (Veggie Tales) has 7 annotations, all of them in Q4; 2) the sentence "*Well now she's gone; even though I hold her tight, I lost my love, my life, that nigh*t" from the song "*Last kiss*" (Pearl Jam) has 6 annotations, all of them in Q3; 3) the sentence "*At the end of all this hatred lies even deeper hate, their darkness has defeated you, your lifeline running backwards*" from the song "*Blood on your hands*" (Arch Enemy) has 10 annotations, 9 on Q2 and 1 on Q3, so the sentence was annotated in Q2; 4) the sentence "*You're the light, you're the night, you're the color of my blood, you're the cure, you're the pain, you're the only thing I wanna touch, never knew that it could mean so much*" from the song "*Love me*



Log out

Sentence:

When I woke up the rain was pourin' down
There were people standing all around
Something warm goin' through my eyes
But somehow I found my baby that night

Instructions:

1. Read the sentence with attention

2. Try to identify the basic predominant emotion expressed by the sentence, according to the sets of emotions (quadrants) of the figure.

3. If you think the sentence doesn't convey any emotion, select the option Neutral.

Figure 2: Main screen of the annotation platform.

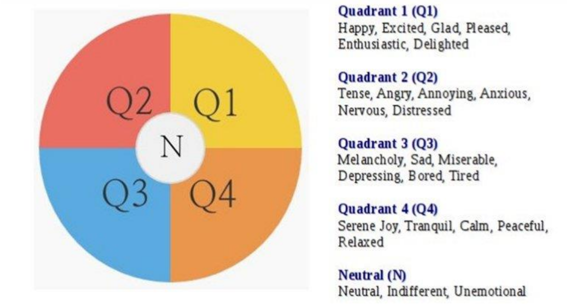

Q2
Q1
N
Q3
Q4
Quadrant 1 (Q1)
Happy, Excited, Glad, Pleased, Enthusiastic, Delighted
Quadrant 2 (Q2)
Tense, Angry, Annoying, Anxious, Nervous, Distressed
Quadrant 3 (Q3)
Melancholy, Sad, Miserable, Depressing, Bored, Tired
Quadrant 4 (Q4)
Serene Joy, Tranquil, Calm, Peaceful, Relaxed
Neutral (N)
Neutral, Indifferent, Unemotional

Figure 3: Predominant emotions by quadrant.

like you do” (Ellie Goulding) has 7 annotations, 6 in Q1 and 1 in Q2, so the sentence was annotated in Q1.

The consistency of the ground truth was evaluated using Krippendorff’s alpha (Krippendorff, 2004), a measure of inter-coder agreement. This measure achieved, for the classes Q1, Q2, Q3, Q4 and N, a value of 53%. This is considered a moderate agreement among the annotators (Landis and Koch, 1977).

According to quadrants, the sentences are distributed in the following way (Table 1).

As can be observed in Table 1, the final validation dataset is not very balanced. Particularly, quadrants 3 and 4 turned out to obtain a much lower number of samples. However, as described below, the training set is nearly balanced.

Table 1: Distribution of the sentences by quadrant.

Quadrant	# **Sentences**
Q1	86
Q2	67
Q3	47
Q4	39
Total	239

3.1.2. Training Set

As will be described later on, our system employs a number of parameters that need to be tuned. To this end, we have additionally created a training dataset. This dataset was annotated according to Russell’s model (4 quadrants) by 2 persons and we just considered sentences in which there were unanimity. We considered a total of 129 lyric sentences from 68 songs, distributed across the four quadrants according to Table 2. As can be seen, this training is nearly balanced.

Table 2: Distribution of the sentences by quadrant.

Quadrant	# **Sentences**
Q1	35
Q2	36
Q3	27
Q4	31
Total	129

3.2 Sentence Emotion Recognition Model (SERM)

We use a knowledge-based approach to create a Sentence Emotion Recognition Model (SERM). This model uses NLP techniques to assign to each sentence an emotion quadrant in Russell's plane, following an unsupervised approach.

Figure 4 shows the architecture of our system.

We use two lexicons to retrieve the values of valence and arousal from the words: Emotion Dictionary (ED) and Dictionary of Affect in Language (DAL) (Whissell, 1989).

To create de ED dictionary:

1. We define as seed words the emotion terms defined for each quadrant and based on Russell's plane (see Figure 2).
2. From these terms, we consider for the dictionary only the ones present in the DAL or the ANEW (Bradley and Lang, 1999) dictionaries. In the DAL, we assume that pleasantness corresponds to valence, and activation to arousal, based on (Fontaine, 2013). We employ the scale defined in the DAL: arousal and valence (AV) values from 1 to 3. If the words are not in the DAL dictionary but are present in ANEW, we still consider the words and convert the arousal and valence values from the ANEW scale to the DAL scale.
3. We then extend the seed words through Wordnet Affect (Strapparava and Valitutti, 2004), where we collect the emotional synonyms of the seed words (e.g., some synonyms of joy are exuberance, happiness, bonheur and gladness). The process of assigning the AV values from DAL (or ANEW) to these new words is performed as described in step 2.
4. Finally, we search for synonyms of the gazetteer's current words in Wordnet and we repeat the process described in step 2. Steps 2, 3 and 4 are repeated iteratively while we add at least a word in an iteration.

Before the insertion of any word in the dictionary (from step 1 on), each new proposed word is validated or not by two persons, according to its emotional value. There should be unanimity between the two subjects. The two persons involved in the validation were not linguistic scholars but were sufficiently knowledgeable for the task.

Based on the procedure above, the emotion dictionary ended up with 1246 words.

Next, we will explain in detail each one of the modules.

After reading a directory containing the lyrics, the lyrics are divided into sentences (verses) and the system processes one sentence at a time.

Removal of Punctuation Marks

The punctuation marks of are first removed. For example the sentence: "Martha, are you playing cello?" is transformed in "Martha are you playing cello"

Word Transformation

In this step, the words in the sentence are transformed according to the rules below, if necessary:

- Verbs in gerund finished by the character "'". The character "'" is replaced by the character "g" (e.g., sittin' → sitting, sippin' → sipping);
- Ended by the characters "'s". These two characters are removed from the word (e.g., the sentence "my mother's house" changes to "my mother house");
- Contraction of verbs or simplification of words due to informal text or slang. These words are corrected according to a dictionary (e.g., ain't → am not, couldn't → could not, won't → will not, they're → they are, hadn't → had not, gonna → going to, gotta → got to, 'cause → because, 'til → until, cuz → because, 'em → them).

VANA Detection

Several works such as (Lu et al., 2006) consider that only verbs (V), adjectives (Adj), nouns (N) and adverbs (A) can convey emotions or can help to understand the emotions.

We follow the same assumption, so we applied a POS tagger (Taylor et al., 2003) to identify the VANA words.

For example, applying a POS tagger to the sentence "Martha. Are you playing cello?" we obtain "Martha/NNP are/VBP you/PRP playing/VBG cello/NN", so the VANA words are "Martha", "are", "playing" and "cello".

SVANA Detection (Selected VANA)

Among the VANA words from the original sentence, we consider for the calculation of the emotion conveyed by the sentence, the adjectives, the nouns (except proper nouns) and the verbs (except auxiliary verbs). So, from the sentence "Martha/NNP are/VBP you/PRP playing/VBG cello/NN", only two words (playing and cello) are selected words to go to the next level.

Modifiers Detection

In this step we will identify words that can change the emotion of the other sentence's words. In this class of words (modifiers) we may include:

- Negations such as for example not, no, never;
- Adverbs such as for example very, extremely, little.

In these modifiers we have always a cause and an object. The cause is the modifier and the object is the word where we can apply the modifier (see Table 3).

Table 3: Example of modifiers in sentences.

Sentence	Modifier	Object
I'm not sad	not	sad
I'm very happy	very	happy

Our system detects automatically the modifiers and the corresponding objects.

When the modifier is a negation, the object is not considered anymore for the calculation of the sentence's emotion (Agrawal and An, 2012). In the sentence "I'm not sad", the emotion conveyed is not necessarily an emotion from the 1st quadrant (e.g., happiness). It can be for example an emotion from the 4th quadrant (e.g., serene joy), i.e., the emotion conveyed is not necessarily the antonym of the object. So we have decided to not consider this kind of words.

To the best of our knowledge, we did not find any dictionary of adverbs classified by intensity. Hence, we decided to create one, so the modifiers were classified according to its intensity in a range between -5 (minimum intensity) and 5 (maximum intensity) by one person, who is not linguistic scholar but is sufficiently knowledgeable for the task. The dictionary has 102 adverbs.

Table 4 shows some examples of adverbs classified according to its intensity.

Table 4: Examples to the weight of the word "happy" in sentences with adverb modifiers.

Sentence	Intensity
extremely	5
very	3
little	-3
rarely	-5

Assignment of Word Weights

These words get a specific weight (WL1), whose value is set as described below (the same value for each word). However, the weights can be modified if they are objects of specific modifiers.

They may increase or decrease if the modifier is an adverb or it may become zero if the modifier is a negation. We have also other different possible weights according to the provenience and the emotional features of the words.

Therefore, we consider the following weights:

- WL1: Represents the weight of the SVANA words – adjectives, nouns (except proper nouns) and verbs (except auxiliary verbs) – that belong to the original sentence.
- WL2: If the selected words from the original sentence belong to the lexicon ED, then the SVANA words of their definitions (see the "retrieval of definitions" step, below) get a weight with value WL2. Words that do not belong neither to ED nor DAL, but their synonyms belong to ED, also get a weight with value WL2.
- WL3: If the selected words from the original sentence do not belong to the lexicon ED then the SVANA words of their definitions get a weight with value WL3. Words that do not belong neither to ED nor DAL, but their synonyms do not belong to ED but belong to DAL, get a weight with value WL3.
- WL4 and WL5: Represent weights to multiply additionally by the initial weight of the words, when these words belong to ED (WL4) and to DAL (WL5).

After the assignment of the word weights, we have to update the weights according the detection of modifiers seen previously. If the selected word is object of a modifier of the type negation then the word will have the weight zero (e.g., the word "happy" in the sentence "I'm not happy").

When the modifier is an adverb, the weight of the object, for the calculation of the emotion, can be increased or decreased. Suppose for instance that the word "happy" in the sentence "I'm happy" has an initial weight of 10 and suppose that in our dictionary the adverbs, extremely, very, little and rarely have respectively the intensity values of 5, 3, -3 and -5. We can see in Table 5 the weight of the object "happy" for sentences using the previous adverbs as modifiers.

Observing the table, the weight of the object in the first sentence (15) is obtained from the sum of the weight associated to the word "happy" (10) by the value associated to the modifier "extremely" (5).

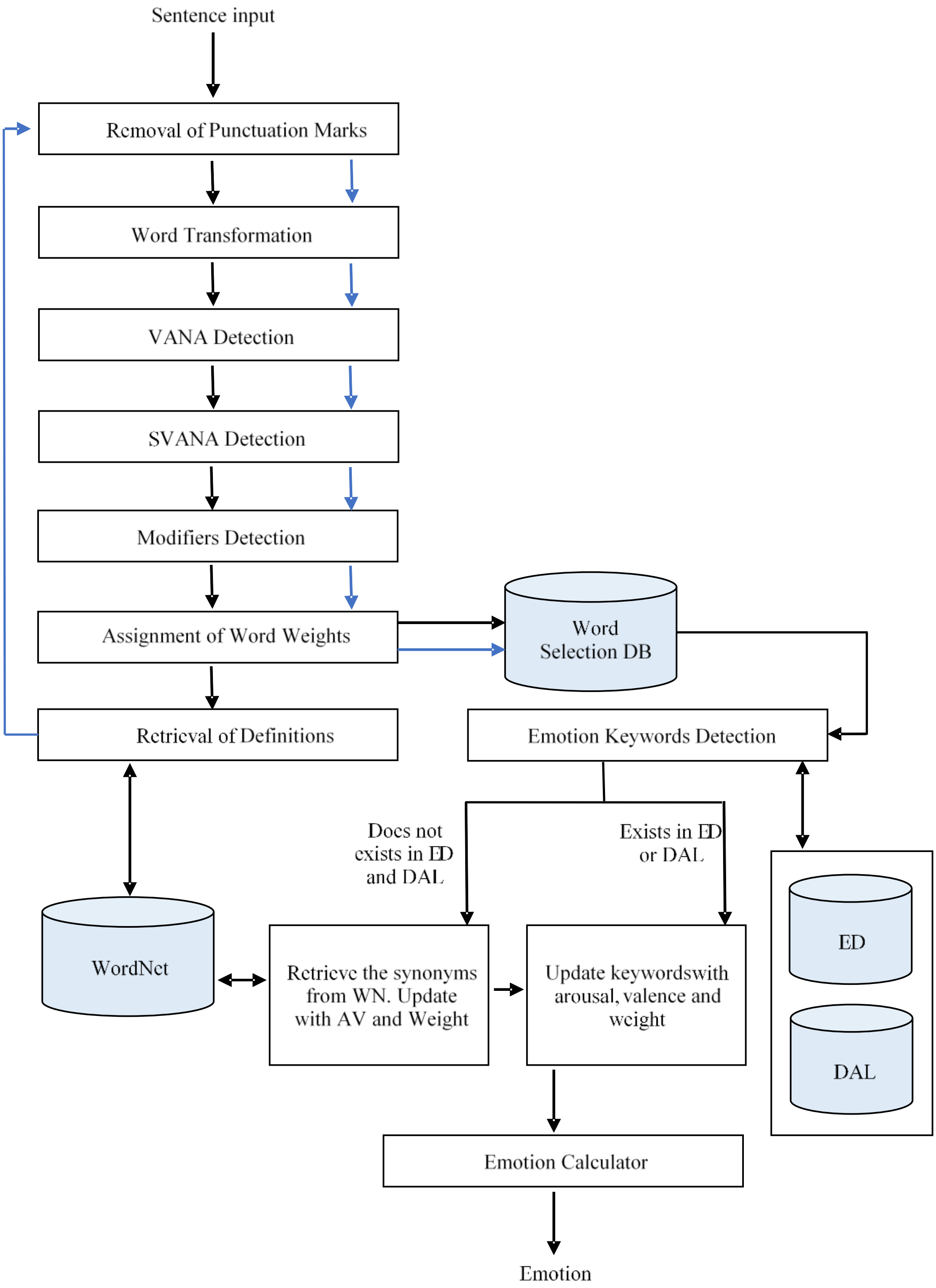

Sentence input
Removal of Punctuation Marks
Word Transformation
VANA Detection
SVANA Detection
Modifiers Detection
Assignment of Word Weights
Word Selection DB
Retrieval of Definitions
Emotion Keywords Detection
Does not exists in ED and DAL
Exists in ED or DAL
WordNet
Retrieve the synonyms from WN. Update with AV and Weight
Update keywordswith arousal, valence and weight
ED
DAL
Emotion Calculator
Emotion

Figure 4: Architecture of the Sentence Emotion Recognition Model (SERM).

Table 5: Examples to the weight of the word “happy” in sentences with adverb modifiers.

Sentence	Weight of the word happy
I’m extremely happy	15
I’m very happy	13
I’m happy	10
I’m little happy	7
I’m rarely happy	5

Retrieval of Definitions

The system retrieves the definition of the selected words (adjectives, nouns (except proper nouns) and verbs (except auxiliary verbs) taken from the original sentence) from Wordnet. We then apply all the prior steps to this definition (sentence): Remove punctuation marks, word transformation, VANA detection, modifiers detection and word weight update. The selected words from definition are then added to database of selected words.

Emotion Keywords Detection

In this step, each one of the originally selected words, as well as the selected words in the definitions, is searched first in the ED, and if it not exists, searched in the DAL.

If the word is in one of these two dictionaries, the corresponding valence and arousal values will be assigned to it.

If the word is not in any of the dictionaries, we retrieve from Wordnet all of its synonyms and then we search them on the ED and the DAL. If they are in the dictionaries, we retrieve valence and arousal.

Emotion Calculator

At this point, the database of selected words contains all the words found in the dictionaries. The predominant emotion (valence and arousal) is then calculated. The final emotion (valence and arousal) is the weighted valence/arousal average of all the selected words, taking into account the weight of each word. The sentence is then classified in one quadrant depending on the obtained valence and arousal values.

4 RESULTS AND DISCUSSION

4.1 Discovering the Best Weights

To build our non-supervised model, we have to find out the optimum values for the weights (WL1, WL2, WL3, WL4 and WL5), which maximize the performance (F-Measure) of the system, when this is applied to new sentences.

To this end, we perform exhaustive tests with the 129 training sentences, combining different values for the different weights in a specific range for each type of weight.

First, we defined experimentally the allowed range for each weight:

- WL1: between 10 e 1500.
- WL2: between 10 and 110.
- WL3: between 2 and 22.
- WL4: between 2 and 5
- WL5: between ½ and 1.

We then performed an iterative local search to look for each optimum. We start with an initial large granularity, which is decreased in the later iterations until the possible minimum level, to find out the best values for the 5 parameters. Illustrating, for WL1 in the first iteration we went from 10 to 1500 in 50-unit steps. Then, if the maximum performance were achieved in the interval between 300 and 400, we would test between 300 and 400 with 10-unit steps. This was repeated until 1-unit granularity was attained. We observed that our system has low parameter sensitivity, as desired. In fact, the system performance changed very slowly for different parameters (see Table 8).

Tables 6 and 7 show respectively the best values for each weight and the confusion matrix for these parameters.

Table 6: Statistics for the best training model.

Weight Level	Value
1	350
2	10
3	10
4	4
5	0.5

Table 7: Statistics for the best training model.

WL1	**WL2**	**WL3**	**WL4**	**WL5**
350	10	10	4	0.5
CM	**Q1**	**Q2**	**Q3**	**Q4**
Q1	27	2	2	4
Q2	3	28	5	0
Q3	6	3	14	4
Q4	6	2	0	23
	Precision	**Recall**	**F-Measure**	
Q1	64.3%	77.1%	70.1%	
Q2	80.0%	77.8%	78.9%	
Q3	66.7%	51.2%	58.3%	

Q4	74.2%	74.2%	74.2%
Average	71.3%	70.2%	**70.3%**

We can see that this combination of weights achieved a performance of 70.38% (F-measure) in the training set.

A possible cause for the lower results of quadrant 3 (13 sentences from quadrant 3 were incorrectly classified in other quadrants) can be related to the fact that this is a keyword-based approach. Quadrants 1, 2 and 4 are more influenced by keywords than quadrant 3, which is more influenced by ideas (e.g., he goes to heaven), as discussed in another work by our team (Malheiro et al., 2016).

We can see the comparison of results of the 10 best models in Table 8.

Table 8: Statistics for the best 10 training models.

WL							
1	**2**	**3**	**4**	**5**	**Prec.**	**Recall**	**FM**
350	**10**	**10**	**4**	**0.5**	**71.29**	**70.24%**	**70.38%**
250	10	10	6	1	70.38	69.55%	69.65%
450	10	2	2	0.5	69.77%	68.98%	69.14%
650	10	2	2	0.5	69.77%	68.98%	69.14%
450	10	2	4	1	69.77%	68.98%	69.14%
450	90	2	2	0.5	69.77%	68.98%	69.14%
550	10	2	2	0.5	69.77%	68.98%	69.14%
550	10	2	4	1	69.77%	68.98%	69.14%
650	10	2	4	1	69.77%	68.98%	69.14%
350	10	2	2	0.5	69.56%	68.98%	69.09%

4.2 Classification of Sentences

We applied SERM with the selected parameters to our sentence validation dataset. The achieved results are summarized in Table 9.

Table 9: Statistics for the validation model.

WL1	**WL2**	**WL3**	**WL4**	**WL5**
350	10	10	4	0.5
CM	**Q1**	**Q2**	**Q3**	**Q4**
Q1	68	5	4	9
Q2	7	44	14	2
Q3	14	0	22	11
Q4	3	0	4	32
	Precision	**Recall**	**F-Measure**	
Q1	73.9%	79.1%	76.4%	
Q2	89.8%	65.7%	75.9%	
Q3	50.0%	46.8%	48.4%	
Q4	59.3%	82.1%	68.8%	
Average	68.2%	68.4%	**67.4%**	

The average F-measure results (67.35%) are very close to the results achieved in the training set (70.82%).

In Table 9, we can also see the confusion matrix. The validation dataset confirms the lower performance of Q3 in comparison to the other quadrants. This is shown by the amount of songs from Q3 erroneously classified in other quadrants (recall is 48.35%) namely Q1 and Q4 (14 and 11 sentences respectively). It is also shown by the amount of sentences from Q2 (14) incorrectly classified in Q3. This fact leads to a low precision for Q3 (50%). Q4 also has low precision (59.26%). This is due to the sentences from Q1 and Q3 being erroneously classified in Q4 (see example below).

At the end of section 3.1.1, we illustrated the annotation results for 4 sentences of the dataset. Table 10 and the text below show the predicted classes for these sentences and possible explanations for the errors.

Table 10: Statistics for the validation model.

Sentences	**Actual**	**Predicted**
I’ve got peace like a river, I’ve got peace like a river in my soul	Q4	Q4
Well now she’s gone, even though I hold her tight, I lost my love, my life, that night	Q3	Q1
At the end of all this hatred lies even deeper hate, their darkness has defeated you, your lifeline running backwards	Q2	Q2
You’re the light, you’re the night, you’re the color of my blood, you’re the cure, you’re the pain, you’re the only thing I wanna touch, never knew that it could mean so much	Q1	Q2

Possible explanations for the wrong classifications in the 2nd and the 4th sentences are related to the vocabulary used. In the 2nd sentence, affective words are almost absent. We can point out only the word *love*, which is a word more related to Q1. This confirms our conclusion that Q3 is more influenced by ideas than keywords in comparison to the other quadrants which are more influenced by the keywords. We can see this typical behaviour in other sentences like “*Oh where, oh where, can my baby be? The Lord took her away from me, she’s gone to heaven, so I’ve got to be good so I can see my baby when I leave this world*” and “*The stars are burning I* *hear your voice in my mind, can't you hear me calling? My heart is yearning like the ocean that's running dry, catch me, I'm falling*", both of them have essentially positive keywords (e.g., baby, heart, ocean). The general idea conveyed by both sentences is associated with Q3 (according to the annotators), but our system classified them in Q1. An example which explains the low recall from Q3 and low precision from Q4 is the sentence "*I lifted her head, she looked at me and said – hold me darling just a little while – I held her close, I kissed her our last kiss, I found the love that I knew I had missed*" from Q3 incorrectly classified in Q4. We can see the predominance of words with positive valence, namely kiss, darling, love, but the general idea for most annotators was associated with Q3.

The 4th sentence belongs to Q1, but our system classified it in Q2. This was probably due to the fact that the sentence uses antithesis and some of the negative words are normally associated with Q2 (e.g., blood, pain).

Another example which can explain the amount of sentences from Q2 erroneously classified in Q3 and consequently imply a low precision for Q3, is the sentence "*Shut up when I'm talking to you, shut up, shut up, shut up, shut up when I'm talking to you, shut up, shut up, shut up, I'm about to break*". This sentence has a predominance of the word shut, and our system has the limitation of not recognizing phrasal verbs (e.g., shut up – more associated with Q2) and the verb shut is associated with Q3, according to DAL. We will address this issue in our future work.

We cannot directly compare the results to other works, because the datasets are different and ours is only one composed by sentences from lyrics that we are aware (the others are composed by other types of text, such as children stories and less subjective text such as journalistic text). Nevertheless the results seem promising in comparison with approaches using machine learning for complete song lyrics, e.g., 73.6% F-measure in another work from our team (Malheiro et al., 2016).

5 CONCLUSIONS

This research addresses the role of the lyrics in the context of music emotion variation detection. To accomplish this task we created a system to detect the predominant emotion expressed by each sentence (verse) of the lyrics, using a use a keyword-based approach, which receives a sentence (verse) and classifies it in the appropriate quadrant, according to Russell's emotion model. To validate our system, we created a training set containing 129 verses and a validation set with 239, annotated manually with an average of 7 annotations per sentence. We attained 67.4% F-measure performance.

The main contributions of our work are the KBA methodology proposed, as well as the ground-truth of sentences created. In the future, we intend to improve our methodology including the improvement of the ED dictionary and a mechanism to detect beforehand if the sentence is emotional or non-emotional.

Moreover, we intend to study emotion variation detection along the lyric to understand the importance of the different structures (e.g. chorus) along the lyric. Additionally, we intend to make music emotion variation detection in a bimodal scenario, including audio and lyrics. This implies an audio-lyrics alignment.

ACKNOWLEDGEMENTS

This work was supported by CISUC (Center for Informatics and Systems of the University of Coimbra).

REFERENCES

Agrawal, A., An, A., 2012. Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations. In Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, 346-353.

Aman, S., Szpakowicz, S. 2007. Identifying Expressions of Emotion in Text. In Proceedings 10th International Conference on Text, Speech and Dialogue TSD 2007, Plzen, Czech Republic, Lecture Notes in Computer Science 4629, Springer, pp. 196-205.

Besson, M., Faita, F., Peretz, I., Bonnel, A., Requin, J. 1998. Singing in the brain: Independence of lyrics and tunes, *Psychological Science*, 9.

Binali, H., Wu, C., Potdar, V. 2010. Computational Approaches for Emotion Detection in Text. 4th IEEE International Conference on Digital Ecosystems and Technologies.

Bradley, M., Lang, P. 1999. Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings. Technical report C-1, The Center for Research in Psychophysiology, University of Florida.

Chopade, C. 2015. Text based Emotion Recognition. International Journal of Science and Research (IJSR), 4(6), 409-414.

Chunling, M., Prendinger, H., Ishizuka, M. 2005. Emotion Estimation and Reasoning Based on Affective Textual Interaction. In Affective Computing and Intelligent

Interaction, Vol. 3784/2005: Springer Berlin / Heidelberg, pp. 622-628.

Fontaine, J., Scherer, K., Soriano, C. 2013. Components of Emotional Meaning. A Sourcebook. Oxford University Press.

Hancock, J., Landrigan, C., Silver, C. 2007. Expressing emotions in text-based communication. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 929-932.

Hevner, K. 1936. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48: 246-268.

Hu, Y., Chen, X., Yang, D. 2009. Lyric-Based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method. Tenth Int. Society for Music Information Retrieval Conference.

Hu, X., Downie, J. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. Proc. Tenth Ann. joint conf. on Digital libraries, pp. 159-168.

Juslin, P., Laukka, P. 2004. Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. Journal of New Music Research, 33 (3), 217–238.

Kao, E., Chun-Chieh, L., Ting-Hao, Y., Chang-Tai, H., Von-Wun, S. 2009. Towards Text-based Emotion Detection. In International Conference on Information Management and Engineering, pp. 70-74.

Korhonen, M., Clausi, D., Jernigan, M. 2006. Modeling emotional content of music using system identification. IEEE Transactions Systems Man Cyber, 36(3), 588-599.

Krippendorff, K. 2004. *Content Analysis: An Introduction to its Methodology.* 2nd edition, chapter 11. Sage, Thousand Oaks, CA.

Landis, J., Koch, G. 1977. *The measurement of observer agreement for categorical data*. Biometrics, 33:159–174.

Laurier, C., Grivolla, J., Herrera, P. 2008. Multimodal music mood classification using audio and lyrics. *Proc. of the Int. Conf. on Machine Learning and Applications.*

Li, H., Pang, N., Guo, S. 2007. Research on Textual Emotion Recognition Incorporating Personality Factor. In International conference on Robotics and Biomimetics, Sanya, China.

Lu, C., Hong, J-S., Cruz-Lara, S. 2006. Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques. Third Taiwanese-French Conf. on Information Technology.

Malheiro, R., Panda, R, Gomes, P., Paiva, R. 2016. Emotionally-Relevant Features for Classification and Regression of Music Lyrics. IEEE Transactions on Journal Affective Computing, Vol 8.

Miller, G. 1995. WordNet: A Lexical Database for English Communications of the ACM Vol. 38, No 11: 39-41.

Russell, J. 1980. A circumspect model of affect. Journal of Psychology and Social Psychology, vol. 39, no. 6, p. 1161.

Russell, J. 2003. Core affect and the psychological construction of emotion. Psychology Review, 110, 1, 145–172.

Schmidt, E., Turnbull, D., Kim, Y. 2010. Feature selection for content-based, time-varying musical emotion regression. In Proceedings of the ACM International Conference on Multimedia Information Retrieval, 267-274.

Schubert, E., 1999. Measurement and time series analysis of emotion in music. Ph.D. dissertion, School of Music Education, University of New South Wales, Sydney, Australia.

Strapparava, C., Valitutti, A. 2004. Wordnet-affect: an affective extension of wordnet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1083-1086, Lisbon.

Taylor, A., Marcus, M., Santorini, B. 2003. The Penn Treebank: an overview. Series Text, Speech and Language Technology. Ch1. 20, 5-22.

Vignoli, F. 2004. Digital Music Interaction concepts: a user study. Proc. of the 5th Int. Conference on Music Information Retrieval.

Whissell, C., 1989. Dictionary of Affect in Language. In Plutchik and Kellerman Emotion: Theory, Research and Experience, vol 4, pp. 113–131, Academic Press, NY.

Yang, Y-H., Liu, C., Chen, H. 2006. Music emotion classification: A fuzzy approach. In Proceedings of the ACM International Conference on Multimedia, 81-84.

Yang, C., Lin, K., Chen, H. 2007. Emotion Classification Using Web Blog Corpora. In IEEE/WIC/ACM International Conference on Web Intelligence.

Yang Y-H. Chen H. 2012. Machine recognition of music emotion: a review. In ACM Transactions on Intelligent Systems and Technology (TIST). Vol. 3, Issue 3.