Ricardo Manuel da Silva Malheiro

Emotion-based Analysis and Classification of Music Lyrics

Doctoral Program in Information Science and Technology, supervised by Prof. Dr. Rui Pedro Pinto de Carvalho e Paiva and Prof. Dr. Paulo Jorge de Sousa Gomes and submitted to the Department of Informatics Engineering of the University of Coimbra

August 2016

· U C ·

Universidade de Coimbra

Ricardo Manuel da Silva Malheiro

Emotion-based Analysis and Classification of Music Lyrics

Doctoral Program in Information Science and Technology, supervised by Prof. Dr. Rui Pedro Pinto de Carvalho e Paiva and Prof. Dr. Paulo Jorge de Sousa Gomes and submitted to the Department of Informatics Engineering of the University of Coimbra

August 2016

· U C ·

UNIVERSIDADE DE COIMBRA

Thesis submitted to the
University of Coimbra
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Science and Technology

This work was carried out under the supervision of

Professor Doutor Rui Pedro Pinto de Carvalho e Paiva
Professor Auxiliar do
Departamento de Engenharia Informática da
Faculdade de Ciências e Tecnologia da
Universidade de Coimbra

and

Professor Doutor Paulo Jorge de Sousa Gomes
Professor Auxiliar do
Departamento de Engenharia Informática da
Faculdade de Ciências e Tecnologia da
Universidade de Coimbra

to Paula and Marta

ABSTRACT

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. 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 process and an expensive and time-consuming task. This should be overcome with the use of automatic systems. Besides automatic music classification, MER has several applications related to emotion-based retrieval tools such as music recommendation or automatic playlist generation. MER is also used in areas such as game development, cinema, advertising and health. Most of early-stage automatic MER systems were based on audio content analysis. Later on, researchers started combining audio and lyrics, leading to bimodal MER systems with improved accuracy.

This research addresses the role of lyrics in the music emotion recognition process. Feature extraction is one of the key stages of the Lyrics Music Emotion Recognition (LMER). We follow a learning-based approach using several state of the art features complemented by novel stylistic, structural and semantic features. To evaluate our approach, we created a ground truth dataset containing 180 song lyrics, according to Russell's emotion model. We conduct four types of experiments: regression and classification by quadrant, arousal and valence categories. To validate these systems we created a validation dataset composed of 771 song lyrics.

To study the relation between features and emotions (quadrants) we performed experiments to identify the best features that allow to describe and discriminate each quadrant. We also conducted experiments to identify interpretable rules that show the relation between features and emotions and the relation among features.

This research addresses also 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 with four sets of emotions (quadrants). To detect the predominant emotion in each verse, we proposed 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.

Finally, we measure the efficiency of the lyric features in a context of bimodal (audio and lyrics) analysis. We used almost all the state of the art features that we are aware of for both dimensions, as well as new lyric features proposed by us.

RESUMO

O reconhecimento de emoções a partir da música (Music Emotion Recognition – MER) está a ser alvo de uma atenção cada vez mais significativa por parte da comunidade científica que se dedica à Recuperação de Informação Musical (Music Information Retrieval). De facto, a pesquisa de música através de emoções é um dos tipos de pesquisa mais efetuados hoje em dia pelos utilizadores. Bases de dados musicais de sites como o AllMusic ou o Last.fm crescem grandemente todos os dias, o que requer uma enorme quantidade de trabalho para as manter atualizadas no que concerne ao processo de catalogação. Infelizmente, a anotação manual de música com etiquetas emocionais é normalmente um processo muito subjetivo e moroso. Isto pode ser ultrapassado com a utilização de sistemas de reconhecimento automático. Além de classificação automática de música, o MER tem várias outras aplicações como recomendação de música, geração automática de *playlists*, desenvolvimento de jogos, cinema, publicidade e saúde. Muitos dos primeiros sistemas automáticos de MER eram baseados apenas na análise do áudio. Estudos mais recentes passaram a combinar as duas dimensões (áudio e letra da música) conduzindo a análises bi-modais que melhoraram a eficácia dos sistemas MER.

Esta investigação foca-se em primeiro lugar no papel das letras musicais no processo de MER. A extração de características (feature extraction) é uma das etapas mais importantes no processo de MER a partir das letras. A nossa abordagem é baseada em aprendizagem e utiliza grande parte das características utilizadas no estado de arte complementadas por novas características estilísticas, estruturais e semânticas propostas por nós. Para avaliar a nossa abordagem, criámos um corpus contendo 180 letras de música anotadas de acordo com o modelo emocional de Russell. Realizámos quatro tipos de experimentos: regressão e classificação por quadrantes de emoções, por grau de valência (valence) e por grau de ativação (arousal). Para validar, criámos um corpus de validação composto por 771 letras de música anotadas através do mesmo modelo de Russell.

Para estudar a relação entre características das letras e as emoções (quadrantes), realizámos experimentos para identificar as melhores características que permitem descrever e discriminar cada quadrante. Conduzimos ainda experimentos para identificar regras interpretáveis que mostrem a relação entre características e emoções e entre características entre si.

Esta investigação foca-se ainda no papel das letras em contexto de deteção de variação de emoções na música. Para tal, criámos um sistema para detetar a emoção predominante transmitida por cada frase ou verso da letra. O sistema utiliza o mesmo modelo emocional de Russell com quatro conjuntos de emoções (quadrantes). Para detetar a emoção predominante em cada verso, propusemos uma abordagem baseada em palavras-chave, que tem como entrada uma frase (ou verso) e como saída a emoção (quadrante) correspondente. Para otimizar os parâmetros do sistema, criámos um corpus de treino constituído por 129 frases tiradas de 68 letras de música. Para o validar, criámos outro corpus com 239 frases tiradas de 44 letras.

No final, medimos a eficácia das características das letras num contexto de análise bimodal (áudio e letra). Utilizámos grande parte das características de letras e áudio presentes no estado de arte, assim como as novas características propostas por nós.

ACKNOWLEDGMENTS

The writing of this thesis is the culmination of a hard work of research that took several years and obviously had the help of several people who I want to thank now.

I would like to start by thanking my supervisors Rui Pedro Paiva and Paulo Gomes (Paulo Gomes was my supervisor together with Rui Pedro Paiva since the beginning until approximately one year ago – he left the Department of Informatics Engineering for new challenges in his career), who were there always for giving me their expert opinion whenever I needed and for encouraging me to go further. I worked more closely with Rui Pedro Paiva, who I know from my master's degree (since he was my co-advisor) and I would like to underline his great professionalism, his quality as supervisor and the friendship demonstrated throughout this great journey.

I would like to thank everybody in the MIR team, namely Renato Panda, Bruno Rocha and António Pedro Oliveira for all the stimulating discussions and share of knowledge. Special remarks to Renato Panda with whom I worked on more closely on specific parts of our projects and that has been always available to help.

I would like to give a special thanks to Hugo Gonçalo Oliveira for the stimulating and interesting meetings we had and for all the help he gave me mainly in the last two years of this work. The same thank to Nuno Seco, with whom I had interesting discussions about my research, normally at lunch.

I would like to thank to all the people who participate in the process of annotation of the three datasets (lyrics, audio and sentences). Their work was fundamental.

I would like to thank to the institution in which I work, Instituto Superior Miguel Torga (ISMT), for the help they gave me during this work. I would also like to thank to several colleagues

and teachers from ISMT but also from CISUC.

Personally, a work of this size and exigency, being a working student, was only possible with all the patience and love always demonstrated by the people to whom I dedicate this work: my wife Paula and my daughter Marta. I want to also highlight, for everything, my mother Adaltiva and my brother Flávio. I am also grateful to my close family and my friends, they know who they are.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

MAIN ABBREVIATIONS

ANEW	affective norms for english words	(defined on page 6)
BOW	bag of words	23
CBF	content based features	52
DAL	dictionary of affective language	6
GI	general inquirer	6
HA	hybrid approach	21
KBA	keyword based approach	21
KNN	k-nearest neighbors	25
LBA	learning based approach	20
LIWC	linguistic inquiry and word count	6
LMER	lyrics music emotion recognition	4
LMEVD	lyrics music emotion variation detection	5
LSA	latent semantic analysis	34
MDL	music digital libraries	35
MER	music emotion recognition	3
MEVD	music emotion variation detection	4
MIR	music information retrieval	3
MIREX	music information retrieval evaluation exchange	4
MLR	multiple linear regression	25
NB	naïve bayes	132
NLP	natural language processing	6
PLSA	probabilistic latent semantic analysis	34
PMI	pointwise mutual information	29
POS Tags	part-of-speech tags	6
SA	sentiment analysis	25
SemBF	semantic based features	54
StruBF	structure based features	53
StyBF	stylistic based features	53
SVC	support vector classification	26
SVD	singular value decomposition	32
SVM	support vector machines	25
SVR	support vector regression	25

TFIDF	term frequency–inverse document frequency	32
URL	uniform resource locator	37
WN	wordnet	28
WNA	wordnet affect	30

Chapter 1

INTRODUCTION

Music is the Shorthand of Emotion

Leo Tolstoy

The importance of music in different societies has been manifested since ancient civilizations. For example, in Egypt, music was very remarkable in daily life, musicians occupied a variety of positions in Egyptian society and music found its way into many contexts: temples, palaces, workshops, farms, battlefields and the tomb. Music was also an integral part of religious worship, so it is not surprising that there were gods associated with music, such as Hathor and Bes (Kelsey Museum, 2003). Also, in ancient Greece, music was almost universally present in society, from marriages and funerals to religious ceremonies, theatre, folk music and the ballad-like reciting of epic poetry. The word music comes from the Muses, the daughters of Zeus[1] and inspirational goddesses of literature, science and arts. They were considered the source of the knowledge embodied in the poetry, lyric songs and myths (Henderson, 1957). This relevant role of the music as a unifying factor of people and civilizations remains until the present day.

There is a strong relation between music and emotions, as we can see through authors like Juslin (Juslin, 2013) who affirms that "emotional expression has been regarded as one of the most important criteria for the aesthetic value of music" and other authors like Cooke (Cooke, 1959) who

[1] Main God of the Greek mythology

says that “Music has been described as a language of the emotions”.

Generally, music is connected to all the sectors of a society: cultural, political, religious and entertainment. We have specific types of songs to convey specific emotions, in normal situations of everyday life such as in an elevator, church, pub, supermarket, TV, radio, in political campaigns to pass normally a sense of credibility or enthusiasm, in yoga or meditation sessions to pass a sense of relaxation, in workout sessions to improve the motivation indices. Another interesting example is cinema, where, depending on the senses or feelings directors intend to convey in a scene or in a movie, the music is chosen accordingly. For example to convey a sense of fear in the movie “Halloween”, the song “Halloween Theme Song”[2] by John Carpenter or to convey a sense of freedom in the movie “Easy Rider”, the song “Born to be Wild”[3] by Steppenwolf, to convey a sense of discovery or grandiosity in the movie “2001: A Space Odyssey”, the song “Also Sprach Zarathustra”[4] by Richard Strauss or to convey a sense of sadness in the movie “Platoon”, the song “Adagio For Strings”[5] by Samuel Barber or, finally, to convey a sense of happiness in the movie “Despicable 2”, the song “Happy”[6] by Pharrell Williams.

People associate music to the most unique moments of their lives, so music is intrinsically connected to their lives. “The history of a people is found in its songs” (George Jellinek).

In this introductory chapter, we present the problem statement, the main motivations, research questions and contributions of this research work, as well as the overall organization of the dissertation. The chapter is structured as described in the following paragraphs.

Section 1.1 Problem Statement, Motivation and Scope

First of all, we state our problem and then we introduce the main motivations and scope of this project.

Section 1.2 Research Questions and Hypotheses

[2] https://www.youtube.com/watch?v=uu2igkV574I

[3] https://www.youtube.com/watch?v=rMbATaj7Il8

[4] https://www.youtube.com/watch?v=M0z_8Gj7wgE

[5] https://www.youtube.com/watch?v=ECQeLQURNuw

[6] https://www.youtube.com/watch?v=MOWDb2TBYDg

In the second section, we describe our research questions as well as the hypotheses.

Section 1.3 Results and Contributions

The main contributions of this work are summarized in connection with the main modules of our system: classification and regression of music lyrics; interpretability; lyrics music emotion variation detection; bimodal analysis (audio and lyrics).

Section 1.4 Thesis Structure

We finish this chapter presenting the structure of this dissertation.

1.1 Problem Statement, Motivation and Scope

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). As sustained by David Huron (Huron, 2000), "music's preeminent functions are social and psychological", and so "the most useful retrieval indexes are those that facilitate searching in conformity with such social and psychological functions. Typically, such indexes will focus on stylistic, mood, and similarity information".

Real-world music databases from sites like AllMusic[7] or Last.fm[8] 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 process and an expensive and time-consuming task. This should be overcome with the use of automatic recognition systems (Hu and Downie, 2010b).

Besides automatic music classification, MER has several applications related to emotion-based retrieval tools such as music recommendation or automatic playlist generation. MER is also

[7] http://www.allmusic.com/

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

used in areas such as game development, cinema, advertising and health.

Most of the early-stage automatic MER systems were based on audio content analysis (e.g., (Lu et al., 2006a)). Later on, researchers started combining audio and lyrics, leading to bimodal MER systems with improved accuracy (e.g., (Hu and Downie, 2010b), (Hu et al., 2009a), (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 (Juslin and Laukka, 2004), 29% of people mention that lyrics are an important factor of how music conveys emotions. Also, Bessson et al. (Besson et al., 1998) have shown that part of the semantic information of songs resides exclusively in the lyrics.

Despite the recognized importance of lyrics, current research in Lyrics-based MER is facing the so-called glass-ceiling (Downie, 2008) effect (which also happened in audio). In our view, this ceiling can be broken with recourse to dedicated emotion-related lyrical features, as we will discuss in the following sections.

Another problem in MER research is that, because of the difficulties in manual annotation, researchers use distinct datasets created by each one, as well as distinct emotion taxonomies, making the comparison of results hard. In addition, none of the current datasets in the literature is public (e.g., (Laurier et al., 2008)). Some efforts have been made to address this issue, namely the Music Information Retrieval Evaluation eXchange (MIREX) mood classification dataset. Unfortunately this dataset does not contain lyrics.

Moreover, Music Emotion Recognition from lyrics (LMER) systems are black-box systems. We are not aware of any study that shows cause-effect relations between features and emotions in the form of interpretable linguistic rules. Such models have the potential of unveiling further knowledge pertaining to the discovery of unknown relations between lyrical features and emotions. Therefore, rule-based LMER models should be exploited.

Another relevant problem in MER is Music Emotion Variation Detection (MEVD). There are some MEVD studies, however they are based exclusively on audio. Therefore, MEVD in lyrics

(LMEVD) is also an open problem that needs to be studied.

1.2 Research Questions and Hypotheses

Regarding the ideas raised in the Section 1.1, we have defined one main research question related to the way we can classify music lyrics based on emotions. Therefore, in this work our main research question is

How to classify music lyrics based on emotions?

The main research question is quite overarching, thus, to answer it we have defined a set of specific research questions which will be described in the next points.

The following questions address a number of issues that we believe are key to answering the main research question. We can have a global classification of the emotions (RQ1) or have an instantaneous classification (RQ3) segment by segment. We can have a black-box system (RQ1) or have a system based on rules of interpretability (RQ2). Moreover, we also aim to evaluate the impact of our system on a bimodal setup (RQ4).

1.2.1 RQ1. Which are the most relevant lyric features to classify music by emotions?

Typical machine learning approaches are associated to the previous question. The lyrics are annotated according to an emotion model and the idea is to understand what are the most efficient features in a classification process.

To answer the RQ1 and the following RQs, we need a dataset annotated taking exclusively into consideration the emotion perceived[9] in lyrics, according to an emotion model, Russell's in our case (see Section 3.1). As previously mentioned, none of the current LMER datasets are public and, hence, the need to create a ground truth. Moreover, the datasets created in other research works do

[9] This work is focused on perceived emotions, rather than expressed or felt emotions.

not completely suit our needs, since they were not annotated taking into consideration exclusively the lyrics. In fact, in those research works, annotators exploit both audio and lyrical content in the annotation process. The weight of the lyric in the annotation is not clear, since the annotation provided belongs to the song as a whole and not specifically to the lyric or to the audio. Since one of the objectives of our work is related to aiming to understand the relations between lyrics and emotions, we believe that the results we would achieve could be more reliable compared to other works which were not carried out in datasets based on lyrics created for this purpose.

Hence, to support the answer to this question, we created a manual dataset firstly annotated through the lyrics (isolating the lyrics) and secondly through the audio (isolating the audio), suitable for the objectives of our work, which are to find better ways to detect emotions in real-world datasets with music, that is, a generic dataset. In addition, to the best of our knowledge, there are no emotion lyrics datasets in the English language that are annotated with continuous arousal and valence values. Therefore, our ground truth fills this gap.

To answer RQ1, we tested most of the features from the state of the art of MER and Sentiment Analysis, namely:

- Content-based features (e.g., n-grams) with and without typical Natural Language Processing (NLP) transformations (e.g., stemming, Part-of-Speech Tags - POS tags);
- Stylistic-based features such as number of occurrences of punctuation marks, grammatical classes, etc.;
- Features based on known lexicons such as Affective Norms for English Words (ANEW) and Dictionary of Affective Language (DAL);
- Features based on frameworks such as General Inquirer (GI), Linguistic Inquiry and Word Count (LIWC), Synesketch and ConceptNet.

Furthermore, we proposed also new features, namely:

- Features based on emotion gazetteers created by us;
- Stylistic-based features, such as the number of occurrences of slang words;

- Structure-based features based on the structure of the lyric (e.g., number of occurrences of verses, chorus).

Next, we tested the features through classification and regression processes.

Finally, we analyzed the impact of the features in each classification problem, including not only the new features but the other features from the state of the art.

RQ1 is answered on Section 3.

1.2.2 RQ2. Can we extract rules from lyrics to help the understanding and improvement of the emotion detection process?

To have a deeper understanding of the emotions conveyed by the lyric, we can complement the black-box systems (mostly used in RQ1) with rule-based systems. Thus, RQ2 arises.

The idea is to detect interpretable rules which relate certain music features to specific emotions or sets of emotions. For that we will use association learning algorithms.

RQ2 is answered on Section 3.

1.2.3 RQ3. How can the variation of emotions along the lyric be captured?

Most of the studies referenced in the state of the art assign a global emotion to each song. However, knowing that the lyric is composed of several segments (e.g., title, chorus), to which specific patterns can be associated, it is natural to think that the emotions conveyed evolve throughout the lyric. Thus, we have another research question.

To answer this RQ3, we used a keyword-based approach, using emotion gazetteers to detect emotions on sentences (verses) of the lyric (Section 4). We start to apply some pre-processing and cleaning operations to the sentence. The final emotion associated to the original sentence depends on the values of arousal and valence from the selected words extracted from the emotion gazetteers. Each word has a specific weight. The selected words may belong to the original sentence, to their synonyms or to their definitions.

We have also created a manual dataset of sentences annotated according to Russell's model.

1.2.4 RQ4. Does our system confirm that bimodal approaches improve the accuracy of MER systems?

The last research question is not so directly linked to our main research question. However, as previously mentioned, past works have suggested that bimodal approaches improve the classification performance of MER systems. Therefore, we believe it is important to address this issue in our work.

In this RQ, we applied the results obtained in RQ1 to a bimodal system. To this end, we extended the dataset created for RQ1 with the annotation of the corresponding audio excerpts and evaluated our system accordingly. RQ4 is answered on Section 5.

1.3 Results and Contributions

This work offers a number of contributions to extend the state of the art in the MER research area, namely:

- Creation of a ground-truth dataset manually annotated through the audio and the lyrics (Russell's model);
- Creation of a larger dataset annotated from AllMusic (Russell's model);
- Creation of a ground-truth dataset of manually annotated sentences (Russell's model);
- Proposal of novel features and/or adaptation of features from other domains;
- Proposal of a novel approach (adapted from NLP research) for music emotion variation detection in lyrics;

- Derivation of a set of rules that relate lyric features and emotions.

The main contributions of this project are summarized in the following publications:

1. Malheiro, R., Panda, R., Gomes, P., Paiva, R. (2016). "Emotionally-Relevant Features for Classification and Regression of Music Lyrics". IEEE Transactions on Affective Computing.

2. Malheiro, R., Panda, R., Gomes, P., Paiva, R. (2013). "Music Emotion Recognition from Lyrics: A Comparative Study". In: 6th International Workshop on Machine Learning and Music (MML13). Held in Conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPPKDD13), Prague, Czech Republic.

3. Panda, R., Malheiro, R., Rocha, B., Oliveira, A., Paiva, R. (2013). "Multi-Modal Emotion Music Recognition (MER): A New Dataset, Methodology and Comparative Analysis". In: 10th International Symposium on Computer Music Multidisciplinary Research (CMMR13), Marseille, France.

4. Malheiro, R., Oliveira, H., Gomes, P., Paiva, R. (2016). "Keyword-Based Approach for Lyrics Emotion Variation Detection". In: 8th International Conference on Knowledge Discovery and Information Retrieval, Porto, Portugal. Note: an extended version of this paper will be submitted to a journal.

5. Malheiro, R., Panda, R., Gomes, P., Paiva, R. (2016). "Classification and Regression of Music Lyrics: Emotionally-Significant Features". In: 8th International Conference on Knowledge Discovery and Information Retrieval, Porto, Portugal.

6. Malheiro, R., Panda, R., Gomes, P., Paiva, R. (2016). "Bimodal Music Emotion Recognition: Novel Lyrical Features and Dataset". In: 9th International Workshop on Machine Learning and Music, Riva del Garda, Italy.

1.4 Thesis Structure

In this introductory chapter we present the premises of our work, namely the problem, motivation and scope, the research questions and approaches we are considering in our research and the contributions we offer to extend the state of the art in this research area.

- Chapter 2: presents an overview about the background and knowledge associated to this work.

- Chapter 3: presents our machine learning system, including the creation of the lyrics dataset, feature extraction, classification and regression experiments, experiments to identify the best models and features for each problem and finally the identification of a set of interpretable rules that relate features and emotions.

- Chapter 4: presents our keyword-based approach to detect emotions in sentences. The chapter includes also the process of creation of the sentences dataset.

- Chapter 5: presents our bimodal analysis and include also the process of creation of the audio dataset. Includes also other classification experiments, using state of the art features and using a dataset annotated from the AllMusic platform

- Chapter 6: presents the main conclusions of our work as well as some perspectives for future work.

Chapter 2

LITERATURE REVIEW

Music evokes a lot of different emotions and triggers different senses

Kaskade

Music and emotions are intimately connected. This connection is synthesized by Juslin and Laukka in (Juslin and Laukka, 2004) who affirm that, for most people, emotions are one of the main motivations to listen to music, and by Pratt in (Pratt, 1950) who summarizes music as the language of emotion. This relation has long been studied by researchers in the field of psychology ((Juslin and Laukka, 2004), (Russell, 1980), (McKay et al., 2002)).

This chapter introduces the background related to our work. The chapter is structured as described in the following paragraphs.

Section 2.1 Music and Emotion: Context and Overview

This section starts with the definition of the term "emotion". Then, we present the different types of emotions (e.g., expressed, perceived, felt). Next, we discuss the subjectivity of emotions, namely, regarding social or cultural issues. Finally, we end up presenting the different emotion representation paradigms, namely, the categorical and the dimensional.

Section 2.2 Detection of Emotions from Text

In this section, we present the approaches commonly employed in the state of the art to construct a ground-truth. Then we show the different methodologies for LMER: learning-based approach, keyword-based approach and hybrid approach. Finally we describe the related work using the prior methodologies.

Section 2.3 Music Lyrics Emotion Variation Detection

This section presents the methods used to detect emotions throughout the songs. This is already used in the state of the art for the audio dimension. For lyrics, as far as we know, this is a new research field.

Section 2.4 Our Approach at a Glance: Comparison to the State of the Art

This section shows a brief analysis from the works related and presents a short overview about the approaches used in this research work.

Section 2.5 Resources

This section shows a brief description of the tools and frameworks used throughout this work.

2.1 Music and Emotion: Context and Overview

In this section we discuss the main aspects involved in the concept of emotion: its definition, types, models and so forth.

2.1.1 Emotion Definition

The etymology of the word "emotion" according the Online Etymology Dictionary[10] says that the word comes from Middle French *émotion* (16th century), from Old French *émouvoir* "to stir up" (12th

[10] http://www.etymonline.com/index.php?term=emotion

century), from Latin *emovere* “move out, remove, agitate”, from assimilated form of *ex-* “out” + *movere* “to move”.

The concept of emotion is not easy to define as we can see in statements such as from Fehr and Russell (Fehr and Russell, 1984), “everybody knows what an emotion is, until you ask them a definition”. Although there are different opinions, we can say that emotions are mental and psychological states associated with several feelings, thoughts and behaviors (Martinazzo, 2010).

Emotion is tightly related to concepts such as mood or affect. In particular, the terms emotion and mood have been used interchangeably, as they have a close meaning. However, there are differences that we must point out.

Starting with the definitions in the American Oxford Dictionary[11], an emotion is “a natural instinctive state of mind deriving from one’s circumstances, mood, or relationships with others”, while a mood is “a temporary state of mind or feeling”.

An emotion arises usually from known causes like for example joy, when we hear a specific song, or anger in the traffic, when we discuss with another driver. On the contrary, a mood arises often from unknown causes like for example a person with depression or that waked up sad and doesn't know why.

An emotion in a person can be enhanced if that person already is in a particular mood. We can synthesize this relation between moods and emotions through the following excerpt from Paul Ekman (Ekman, 2003): “A mood resembles a slight but continuous emotional state. If it is irritability, it is like being mildly annoyed all the time, ready to become angry. If it is a blue mood, we are slightly sad, ready to become very sad. A mood activates specific emotions. When we are irritable, we are seeking an opportunity to become angry; we interpret the world in a way that permits, or even requires, us to become angry. We become angry about matters that do not typically get us angry, and when we become angry, the anger is likely to be stronger and last longer than it would if we were not in an irritable mood”.

According to Ekman, this relation between moods and emotions is bidirectional, since a

[11] http://www.oxforddictionaries.com/

determined mood may appear when a person is subjected to a highly dense emotional experience. For example, dense joy can result in a high or euphoric mood. In those situations we know why we have this specific mood.

Another difference is that usually an emotion is short-lived (e.g., seconds, minutes), while a mood can last longer (e.g., days). An emotion is also generally a stronger feeling than a mood, i.e., a person may be experiencing a depressive mood and have moments of joy. Clearly the emotions that we have throughout the day will normally be influenced by mood.

Although they are often used interchangeably in the MIR research community, the concept of emotion is more popular in the area of music psychology ((Meyer, 1956), (Juslin et al., 2006)) while the concept of mood is normally more popular in the area of Music Information Retrieval (MIR) ((Feng et al., 2003a), (Mandel et al., 2006), (Hu and Downie, 2007)). Nevertheless, in our opinion the term "emotion" is more accurate and, thus, we will employ it preferably in this document.

2.1.2 Emotion Types: Expressed, Perceived and Felt

Emotions are commonly divided into three categories: expressed emotions, perceived emotions, and felt emotions (or induced emotions) (Gabrielsson,2002).

- Expressed emotion: refers to the emotion the performer tries to communicate to the listeners (Gabrielsson and Juslin, 1996).

- Perceived emotion: regards the emotion one perceives as being expressed in a song (which may be different than the emotion the performer tries to communicate) (Gabrielsson and Juslin, 1996) or the emotion felt by the listener.

- Felt (induced) emotion: occurs when one actually feels an emotion in response to a song (Scherer and Zentner, 2001), (Sloboda and Juslin, 2001).

Albeit perceived emotions and felt emotions are both related to the emotional responses of the listeners, on emotion perception we may perceive an emotion being expressed in a song, while on emotion induced we actually feel an emotion in response to the song. Both perceived emotion and felt emotion, especially the latter, are dependent on an interplay between the musical, personal, and

situational factors (Gabrielsson,2002). MIR researchers tend to focus on the perceived emotion, for it is relatively less influenced by the situational factors (environment, mood, etc.) of listening (Yang and Chen, 2012).

All the three types, although inter-related, might be different. For example, one performer might attempt to transmit happiness, but one listener might perceive serenity, despite the fact that that song might make that listener feel depressed.

In this work, we are focused on emotion perception, not on emotion expression or induction. Hence, from this point on, unless explicitly stated, we will focus on perceived emotion.

2.1.3 Emotion and Subjectivity

In all the emotion types described in the previous section, emotion can be regarded as a subjective construct (Yang and Chen, 2011). For instance, a specific song may be associated to a sad moment of our life (e.g., the favorite song of a dear relative who passed away) while for other person, the same song can be associated to a happy moment (e.g., be associated to a personal conquer). Memories, experiences, culture, age, gender, personality and other factors might influence how emotion is perceived, felt or expressed. In addition, as mentioned in (Sloboda, 2001), it is commonly accepted that some people are more "emotional" than others and, hence, able to perceive, feel or express emotions with more clarity. In fact, emotion is by nature ambiguous and subjective.

There are social factors that can potentiate the emotions conveyed by music. In a rock concert, the audience is exultant and, thus, there is a contagious effect in the emotions, i.e., the social factor is relevant to the induction of the emotion, as well as its intensity.

Cultural issues can also influence the way people perceived/express/feel emotions. For example in the western cultures, and in a musical context, it is usual not to express emotions when we hear some types of music, as classical music or in Portugal the *Fado*[12], although these genres can convey, possibly, strong emotions (Gabrielsson,2002). That is, social and cultural factors can affect the way we express/feel/perceive emotions.

[12] https://en.wikipedia.org/wiki/Fado

Regarding specifically perceived emotion, music perception is intrinsically subjective and is influenced from many factors referenced above such as cultural background, age, gender, personality, and training, among others (Abeles and Chung, 1996). The interactions between music and listener may also involve the listener's familiarity with the music (Jargreaves and North, 1997) and his musical preferences ((Jargreaves and North, 1997), (Holbrook and Schindler, 1989)). Because of this subjectivity issue, it is difficult to achieve consensus concerning the choice of the best affective term to characterize a specific piece of music (Yang and Chen, 2011).

Furthermore, different emotions can be perceived along the same song. This is more usual in specific types of music (e.g., classical music) than in others.

2.1.4 Emotion Representation Paradigms: Categorical vs Dimensional

Studies in the area of psychology have identified two emotion representation paradigms: categorical and dimensional. The main difference between the two models is that while in the categorical paradigm, emotions are represented as a set of discrete categories or emotional descriptors (Kim et al., 2010) identified by adjectives, in the dimensional models, emotions are organized along 2 or 3 axes as discrete adjectives or as continuous values. (Russell, 1980).

Categorical models

In this type of models, people experience emotions as categories that are distinct from each other (Yang and Chen, 2012). The most known model in this paradigm is probably Ekman's model (Ekman, 1982). In this model, emotions are classified in six categories: anger, disgust, fear, happiness, sadness and surprise. These emotions are known as basic emotions (Ekman, 1992), however, as this model was developed for encoding facial expressions, some of these categories may not be adequate for the musical case (e.g., disgust), while some moods usually associated to music are not present (e.g., calm, soothing) (Hu, 2010).

Another known model is Hevner's (Hevner, 1936), which divides emotions into eight categories or clusters using a total of 67 adjectives (Figure 2.1).

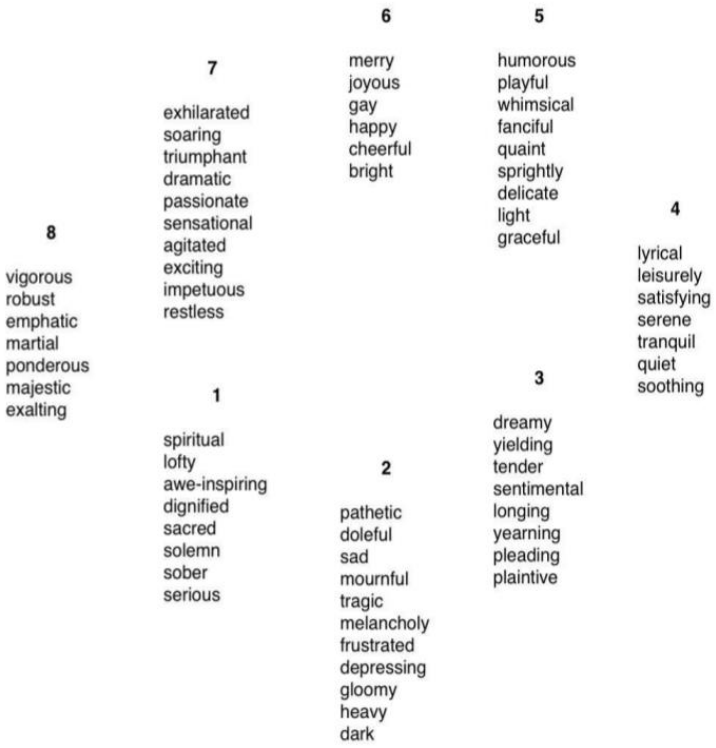

Figure 2.1. Hevner's model (Hevner, 1936) (adapted from (Yang and Chen, 2012)).

We consider that there is a great intra-cluster similarity, which means that inside each cluster the adjectives are very close in meaning for classification effects. On the other hand, the closeness of meanings between adjectives from adjacent clusters is bigger than from adjectives from distant clusters (e.g., the adjectives sad (cluster 2) and serious (cluster 1) are closer in meaning than the adjectives sad (cluster 2) and happy (cluster 6)).

Hevner's model was later adapted by Farnsworth (Farnsworth, 1954) to included ten adjective groups and by Schubert (Schubert, 2003), who defined nine adjective groups.

MIREX (Music Information Retrieval Evaluation eXchange) is the framework employed by the Music Information Retrieval (MIR) scientific community for the formal evaluation of systems and algorithms (Downie, 2008). In MIREX songs are categorized into one of five mood clusters, shown in Table 2.1. MIREX: The five clusters and respective subcategories. The five categories were derived by performing clustering on a co-occurrence matrix of mood labels for popular music from

the AllMusic[13] (Kim et al., 2010).

Clusters	Mood Adjectives
Cluster 1	Passionate, Rousing, Confident, Boisterous, Rowdy
Cluster 2	Rollicking, Cheerful, Fun, Sweet, Amiable/Good Natured
Cluster 3	Literate, Poignant, Wistful, Bittersweet, Autumnal, Brooding
Cluster 4	Humorous, Silly, Campy, Quirky, Whimsical, Witty, Wry
Cluster 5	Aggressive, Fiery, Tense/anxious, Intense, Volatile, Visceral

Table 2.1. MIREX: The five clusters and respective subcategories.

According to Yang and Chen (Yang and Chen, 2012) the major drawback of this categorical approach is that the number of primary emotion classes is too small in comparison to the richness of music emotion perceived by humans. Moreover, according to Laurier et al., (Laurier et al., 2008), there is a semantic overlap between clusters 2 and 4, and an acoustic overlap between clusters 1 and 5.

Dimensional models

In this type of models, emotions are organized along 2 or 3 axes. These models correspond to internal human representations of emotions. Russell (Russell, 1980) even went as far as claiming that valence and arousal are the "core processes" of affect, constituting the raw material or primitive of emotional experience (Yang and Chen, 2012).

Russell's dimensional model (Russell, 1980), (Thayer, 1989) (Figure 2.2) is the most well-known model in this category and is broadly used in several Music Emotion Recognition (MER) studies (Juslin and Sloboda, 2001), (Laurier et al., 2009).

13 http://www.allmusic.com/

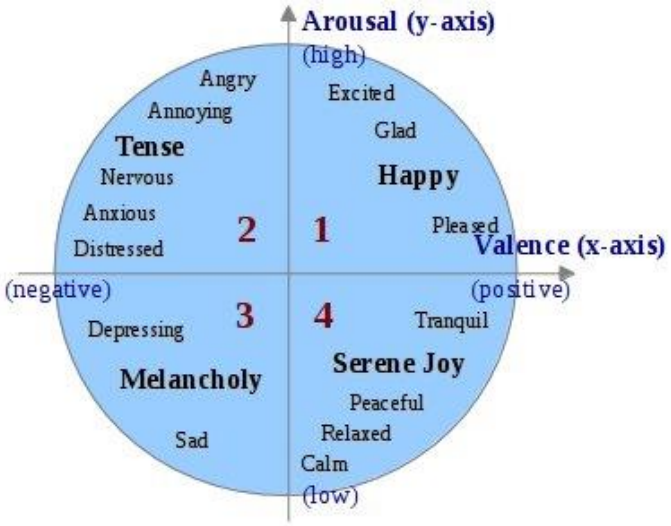

Figure 2.2. Russell's circumplex model (adapted from Yang et al., 2008).

In its abbreviated and more usual form, this model represents emotions using a Cartesian space composed by the two emotional dimensions: arousal and valence. The Y-axis represents arousal (also known as activation, energy and stimulation level) while the X-axis represents valence, i.e., the polarity of emotion (positive and negative affective states, also known as pleasantness). The complete model contains a third dimension: dominance or potency (a sense of control or freedom to act ((Tellegen et al., 1999), (Schimmack and Reisenzein, 2002)). However, for the sake of simplicity, this dimension is not usually employed in most MER works.

Dimensional models can be categorized into discrete or continuous. In discrete models, different regions of the emotion plane represent different emotions, described by different emotion tags, as previously described in the categorical paradigm. For example, Russell's model may be regarded as containing 4 emotions, one for each quadrant (happy in quadrant 1, angry in 2, sad in 3 and relaxed in 4). Besides this representation, Russell proposes a number of adjectives that are distributed in the Cartesian plane (Figure 2.2).

In continuous models there are no specific emotion tags. On the contrary, emotions are regarded as a continuum, and so each point in the plane can represent a different emotion. For this reason, it is argued that the continuous paradigm entails lower ambiguity since no subjective tags are employed (Yang et al., 2008).

Two-dimensional continuous model entail, nevertheless one important limitation. In fact, since the dominance or potency axis is typically discarded, some important aspects of emotion might be obscured. Namely, anger and fear are closely placed in the arousal-valence plane, but they have opposite dominance. Hence, excluding this dimension might lead to ambiguous emotion characterization, as illustrated in this example (Yang et al., 2008).

2.2 Detection of Emotions from Text

In recent years, the task of emotion detection from text has deserved growing attention by the scientific community. Still, there is a paucity of research in emotion detection from text in comparison to the other areas of emotion detection (Binali et al., 2010) (e.g., audio, speech and facial emotion detection). Emotion research has recently attracted increasing attention from the Natural Language Processing (NLP) community – it is one of the tasks at Semeval-2007[14]. A workshop on emotional corpora was also held at LREC-2006[15].

Similarly, in the music domain, the area of Music Information Retrieval (MIR) has significantly more work devoted to tasks such as genre detection. In fact, the identification of musical emotions is still in its early stages, though it has received increasing attention in recent years (Kim et al., 2010). Moreover, most of the research on Music Emotion Recognition (MER) is devoted to the audio modality and significantly less attention has been devoted to the detection of emotion from lyrics (LMER).

Current LMER systems employ different emotion taxonomies, datasets and methodologies, according to the specific problems they address, e.g., learning-based, keyword-based or hybrid approaches (Binali et al., 2010):

- **Learning-based approach (LBA).** It 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

14 Semeval-2007 - http://nlp.cs.swarthmore.edu/semeval/tasks/task14/summary.shtml

15 LREC-2006 - http://www.lrec-conf.org/lrec2006/IMG/pdf/programWSemotion-LREC2006-last1.pdf

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 approach (KBA).** It 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). Some authors, e.g. (Chopade, 2015), consider the lexicon-based approach as a 4th independent approach, while other authors as Binalli (Binali et al., 2010) consider this approach, which counts the number of words of a lexicon into the text, included in the keyword-based method.

- **Hybrid approach (HA).** It is 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).

In this section, we review the main approaches for the creation of a ground-truth and the main emotion detection methodologies employed. Moreover, as lyrics can be regarded as a type of text, LMER works are reviewed along with works on emotion detection from general text.

2.2.1 Ground Truth Construction

Presently, there is no common benchmark to validate and compare LMER research works. Each researcher usually builds his/her own dataset, making comparisons difficult between different approaches. Moreover, there is no consensus on the emotion taxonomy to employ (e.g., inside a categorical model we may have different number of categories: 4 categories (Feng et al., 2003b), 8 categories (Hevner, 1936)). Each researcher has its own vision.

The quality of the ground-truth is fundamental for the quality of the emotion detection process.

To build a ground truth we have to take decisions related to: i) the emotion representation paradigm; ii) the type and number of categories and/or dimensions; iii) the number of instances; iv) the type of instances, i.e., guarantee the representativeness of the instances.

Regarding the employed emotion representation paradigms, different models are used: Hu created a model with 18 categories of emotions (Hu, 2010), Laurier created a model with 4 categories of emotions (Laurier, 2011), MIREX is a model with 5 categories of emotions (Downie, 2008) and Yang created a dimensional model with 2 axes - arousal and valence) (Yang et al., 2008).

Regardless of the employed taxonomy, a set of song samples must be collected and annotated. To this end, different approaches might be followed.

One typical way to obtain annotations is by conducting manual annotation studies. These can be divided into two categories: expert-based or subject-based. In the expert-based annotation, the song is annotated by experts (typically less than 5) (Yang and Chen, 2011) and unanimity is often a requirement, i.e., when there is no consensus among the experts the song is often abandoned. In the subject-based annotation, non-expert subjects (typically more than 10) (Yang and Chen, 2011) annotate the songs. Then, each song is annotated using the average of the opinions of all subjects (Yang et al., 2008). As a result, datasets constructed following this approach tend to be controlled but small.

As manual annotation is a time-consuming task, some researchers use other methods, e.g., tags are obtained directly from Internet sites like AllMusic or Last.fm[16]. For example, through the AllMusic web service, we can easily obtain the more representative songs for a specific mood tag.

Comparing to manual annotation, with this method it is easier and faster to collect the ground truth data, but the quality of the annotations may not be so reliable because, for example, the tags in Last.fm are assigned by online users and the annotation process in AllMusic remains more or less unknown. There are several works using this approach ((Hu, 2010), (Laurier et al., 2009)).

Another method to annotate emotions is through collaborative games on the web, also termed Games with a Purpose (GWAP) (Kim et al., 2010). The following example is described by Yang and

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

Chen (Yang and Chen, 2012) about the online multiplayer game called Listen Game (details in (Turnbull et al., 2007)): "When playing the game, a player sees a list of semantically related words (e.g., instruments, emotions, usages, genres) and is asked to pick both the best and worst word to describe a song. Each player's score is determined by the amount of agreement between the player's choices and the choices of all other players".

2.2.2 Learning-Based Approach (LBA)

Feature Extraction

The features extracted from text are divided into various categories (Hu, 2010): i) Content-based features; ii) text stylistic features based on the style of the written text; iii) linguistic features based on lexicons.

Content-Based Features

The most used features in text analysis (and, consequently, in lyric analysis) are content-based features (CBF), namely the bag-of-words (BOW) (Sebastiani, 2002). In this model, the text in question (e.g., lyrics) is represented as a set of bags which normally corresponds, in most cases, to unigrams, bigrams and trigrams.

Illustrating, in the sentence below, the unigrams, bigrams and trigrams representation would be the following:

She looked really lovely in the blue dress

Unigrams: *She; looked; really; lovely; in; the; blue; dress*

Bigrams: *She looked; looked really; really lovely; lovely in; in the; the blue; blue dress*

Trigrams: *She looked really; looked really lovely; really lovely in; lovely in the; in the blue; the blue dress*

The bag-of-words are typically associated to a set of transformations that are applied

immediately after the tokenization of the original text. These transformations usually involve, for example, stemming and stopwords removal.

Stemming allows each word to be reduced to its stem or to its root and it is assumed that there are no differences, from the semantic point of view, in words that share the same stem. Through stemming, words like “argue”, “argued”, “argues”, “arguing” and “argus” would be reduced to the same stem “argu”.

The *stopwords* (e.g., the, is, in, at), which may also be called function words, include mainly determinants, pronouns and other grammatical particles, which, by their frequency in a large quantity of documents, are not discriminative. Function words removal have been used successfully applied in works such as (Argamon et al., 2003).

POS tags are another type of features used. They consist in attributing a corresponding grammatical class to each word. Some of the most known grammatical classes of the English language (using Penn Treebank (Taylor et al., 2003)) are: noun (NN), determiner (DT), adjective (JJ), verb (VB), adverb (RB).

Illustrating, the grammatical tagging of the sentence “The student read the book” would be “The/DT student/NN read/VBZ the/DT book/NN”.

The POS tagging is typically followed by a BOW analysis. This technique was used in studies such as (Li and Ogihara, 2004) and (Mayer et al., 2008).

Text Stylistic Features

These features are related to stylistic aspects of the language. Some known studies (e.g., (Hu, 2010)) include, for instance: the number of interjections such as “yeah” or “ah”; punctuation marks, such as “...” or “!”; types of words such as adjectives or verbs; text statistics, such as the number of unique words or the number of lines.

One of the issues related to the written style is the choice of the type of the words to convey a certain idea (or emotion, in our study). Concerning music, those issues can be related to the style of the composer, the musical genre or the emotions that we intend to convey.

Linguistic Features based on lexicons

Since the 1930s, psychologists have interpreted the affective value of words based upon empirical surveys and expert judgments (Hu, 2010). There are some lexicons that measure words in several dimensions (e.g., arousal, valence and dominance) and diverse rating scales for the words. The documents are rated by averaging the ratings of the individual words. Other lexicons assign simply affective or psychological states to each word. These lexicons, such as ANEW (Bradley and Lang, 1999) and General Inquirer[17] (GI), will be described in Section 2.5.

One particular sub-field of emotion detection from text is Sentiment Analysis (SA), which is the extraction of positive and negative emotions from unstructured text (Pang and Lee, 2008). Most of the features used in LMER were used since the beginning in the SA area. These features may be for instance a subset of the words of a document, parts of speech or n-grams (Abbasietal.,2008a), (Ng et al., 2006) and (Tang et al., 2009).

There are important features to LMER, which do not exist in the state of the art of detection of emotions from text. For example features related to the written style of the composer such as the use of slang words or other specific features from the lyrics such as the number of repetitions of the chorus and the title into the lyric. We aim to close this gap in the state of the art.

Classification and Regression

After pre-processing the text, which may include, for example, tasks such as detection of typos, we start the machine learning process. Some of the algorithms more commonly used for classification include decision trees such as C4.5 (Cohen and Hirsh, 1998), K-Nearest Neighbors (KNN) (Altman, 1992), Naïve Bayes algorithm (Russell and Norvig, 2003) and Support Vector Machines (SVM) (Boser et al., 1992). For regression some of the algorithms more frequently used are Multiple Linear Regression (MLR) (Sen and Srivastava, 1990), Support Vector Regression (SVR) (Smola and Schölkopf, 2004), and AdaBoost.RT (BoostR) (Solomatine and Shrestha, 2004).

[17] http://www.wjh.harvard.edu/~inquirer/

For text categorization, Naïve Bayes and SVM (or Support Vector Classification – SVC) are almost always considered. Naïve Bayes often serves as a baseline, while SVM seems to achieve top performances (Yu, 2008). SVMs are more computationally expensive than the Naïve Bayes classifier. However, SVMs are very robust with noisy examples and can achieve very good performance with relatively few training examples because only support vectors are taken into account. SVMs are one of the most widely used classification algorithms because they generally obtain good results in many types of machine learning tasks ((Abbasietal.,2008a), (Abbasi et al., 2008b), (Argamon et al., 2007), (Binali et al., 2010), (Gamon, 2004), (Mishne, 2005), (Teng et al., 2006) and (Wilson et al., 2006)).

For text regression, we employ SVR (SVM for Regression) for the same reasons.

Comparisons using several classification algorithms can be seen in (Airoldi et al., 2006), (Das and Chen, 2007), (Dave et al., 2003), (Gamon, 2004), (Matsumoto et al., 2005) and (Mullen and Collier, 2004).

To improve the performance of the classifiers, feature selection is usually performed to reduce the number of features. One of the most known algorithms is ReliefF (Robnik-Šikonja and Kononenko, 2003).

2.2.3 Keyword-based Approach (KBA)

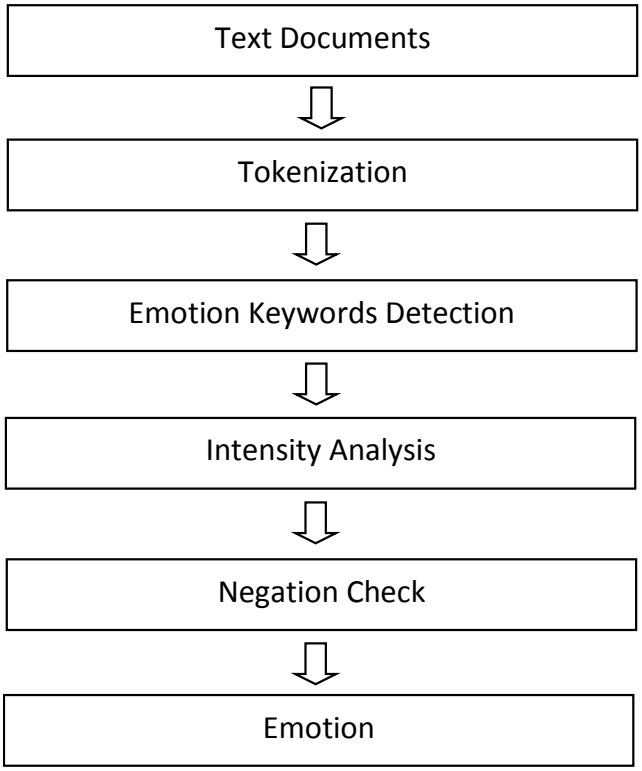

Figure 2.3. Keyword Spotting Technique.

This technique is based on the presence of keywords in the text. These keywords are associated to specific emotions according to the emotion model used. The strength of the emotion can be also associated to the keywords.

This approach involves typically the stages illustrated in Figure 2.3.

The general idea is to receive as input a text document and to generate as output an emotion class. First the text document is converted into tokens. Then emotion words are identified in these tokens. Next, analysis of the intensity of the emotional words is performed. This phase is followed by checking of negation in the sentences. Finally the output is generated through an emotion class. The type of emotion depends on the emotion model used. For example the Ekman's emotion model (Ekman, 1982) contains 6 emotions: disgusted, sad, happy, angry, fearful and surprised.

There are some known limitations in this approach (Hirat and Mittal, 2015):

- **Ambiguous definition of keywords.** Often, the same word has different meanings according to several contexts and usages. For example the word “crush” may be a synonym of destruction in “he crushed the car” or a synonym of a person in love in “he had a crush on her”. It is not possible to include all these possible meanings in the emotion dictionary. Even words clearly associated to certain emotions may have different meanings if the type of discourse is, for example, ironic.

- **Emotions are recognized only in the presence of keywords.** The presence in the sentence of emotional keywords is fundamental to assign an emotion to the sentence. If the sentence does not have any emotional word is because the sentence is a non-emotional sentence. We know that this is not true because the emotion is often passed by the idea conveyed and not specifically by the keywords used. We can see this, for example, in the sentence “he left us for a better place”.

- **Lack of linguistic information.** The expressed emotions are influenced by the used syntactic and semantic structures. The sentence “I laughed at him” and “He laughed at me” suggest different emotions, positive and negative respectively.

In our work, we use a KBA to detect emotions in sentences and, then, to understand how the emotions vary along the lyric. Our work aims to mitigate the first two previous limitations.

- 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.

- Emotions are recognized only in the presence of keywords. In our work, the retrieved synonyms and definitions to help to extend our keyword list.

2.2.4 Analysis of Current Systems in the Literature

In Table 2.2, we show some of the studies that are more relevant to our research, using the learning-based, keyword-based and hybrid approaches. We present only the information useful for our work. In Section 2.4 we describe in which way they have influenced our work.

Reference; Type	Dataset; Emotion Model	Methodology
(Agrawal and An, 2012); KBA	2 types of datasets 1) a) Wikipedia data b) Gutenberg corpus (36000 ebooks) 2) Alm's dataset (Alm et al., 2005) (1207 sentences annotated with 5 emotions taken from 176 fairytale stories); 6 emotions (Ekman's model)	"Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations" 1. Preprocessing the sentence including POS tags detection. 2. Detection of NAVA (Nouns, Adverbs, Verbs, Adjectives) words. 3. Exploitation of syntactic dependency parsing as, for example, detection of modifiers (e.g., negation modifiers: "he is not happy" – "not" influences the mood associated to "happy"). 4. Representation of each word by a 6th position vector (one position for each emotion). 5. Use of Pointwise Mutual Information (PMI) (Read, 2004) to assign an emotion to a word that co-occurs with another word with that emotion.

		6. Through the word's vectors, calculation of the emotion vector of the sentence.
(Alm et al., 2005); LBA	185 children stories; 2 problems: a) emotional vs non-emotional sentences b) positive vs negative emotions	"Emotions from text: machine learning for text-based emotion prediction" 1. Feature Extraction: BOW and other 30 features such as number of "?", "!" and percentage of POS tags (e.g., percentage of adjectives). 2. Classification: Naïve Bayes.
(Aman and Szpakowicz, 2007); LBA	Blog data annotated with emotion category, intensity, emotion indicators; 6 emotions (Ekman's model)	"Identifying Expressions of Emotion in Text" 1. Feature Extraction: General Inquirer and WordNet Affect (WNA). 2. Classification: SVM and Naïve Bayes. In other study the authors do the same study but using Roget's thesaurus[18] to build a lexicon of emotion related words.
(Binali et al., 2010); HA	Blog data; Two problems: a) 6 emotions (Ekman) b) 2 classes (positive and negative)	"Computational Approaches for Emotion Detection in Text" 1. KBA: a) tokenizer, sentence splitter, POS tagger b) based on some previous keywords, construction of gazetteers lists and rules to automatically classify sentences into classes. 2. LBA: SVM algorithm application to build a prediction model.

[18] http://www.thesaurus.com/Roget-Alpha-Index.html

(Chaffar and Inkpen, 2011); LBA	3 datasets: news headlines, fairytales; blogs; 6 emotions (Ekman's model)	"Using a Heterogeneous Dataset for Emotion Analysis in Text" 1. Feature Extraction: BOW, lexical emotion features. 2. Classification: SVM, C4.5 and Naïve Bayes. SVM got better results.
(Chuang and Wu, 2004); KBA-bimodal	1085 sentences in 227 dialogues manually annotated (Chinese); 6 emotions (Ekman's model)	"Multi-Modal Emotion Recognition from Speech and Text" 1. Detection of emotion keywords in the sentences. The authors assume that every sentence has at least one emotion keyword. For each emotion keyword, the corresponding emotion descriptor (emotional state label and intensity value) is manually defined (e.g., the word "depressed" has the emotional state "sadness" and an intensity value of 0.6. 2. Detection of modification values i.e., detection of adverbs (e.g., very, extremely) and negations (e.g., not, never). The authors change the emotional descriptor (e.g., "very happy" is stronger than "happy" and "not happy" may be "sad" or "angry". 3. Calculation of the final emotional state from acoustic features and the prior lyric features.
(del-Hoyo et al., 2009); HA	3878 Spanish movie reviews; NA	"Hybrid Text Affect Sensing System for Emotional Language Analysis" 1. LBA: a) Preprocessing module: sentence correction and cleaning; lemma extraction; POS

		tagger; stopwords removal. b) Feature Extraction: statistical features are constructed in base of a term vector space model (unigrams). Features are represented by Term Frequency–Inverse Document Frequency[19] (TFIDF). Feature reduction through the Singular Value Decomposition (SVD) algorithm. c) Classification: SVM and Multilayer Perceptron. 2. KBA: In the point 1b) they consider only words in DAL (Spanish version). The rest is equal.
(Hu et al., 2009b); LBA-bimodal	5585 songs (En) annotated through Last.fm with WordNet Affect (WNA); 18 emotion categories	"Lyric Text Mining in Music Mood Classification" 1. Lyrics Feature Extraction: a) BOW with and without stemming b) POS tags c) Stopwords removal. 2. Classification: Binary classification for each one of the 18 categories. Some of the songs are in more than one category; SVM algorithm. 3. Bimodal analysis comparing the best lyric features to the best audio features. Audio features do not always outperform lyric features, but combining both dimensions improve the results in comparison to each dimension separately.
(Laurier et al., 2008); LBA-bimodal	1000 songs (En) annotated through Last.fm; 4 emotion	"Multimodal Music Mood Classification using Audio and Lyrics" 1. Technique to select the most discriminative terms

[19] Is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus

	categories (Russell's model)	looking at the differences between language models. 2. Classification: Binary classification; SVM algorithm. 3. Bimodal analysis. Audio analysis achieved better results than lyrics-only analysis, however audio+lyrics is better than each one of the dimensions separately.
(Lu et al., 2006a); KBA	Affective Chatting Room; 6 emotions (Ekman's model)	"Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques" 1. Detection, through Semantic Role Labeling in sentences, of subjects and objects (e.g., "a girl found a tiger"; girl – subject, tiger – object). 2. Through the google function "define", establishment of connection of these words respectively to the adjectives "young" and "predatory". 3. Combination of the previous adjectives with the verb find (consulting a DB always in construction), resulting the emotion "fear".
(Seol et al., 2008); HA	3200 sentences manually annotated; 8 emotions (anger, fear, hope, sadness, happiness, love, thank, neutral)	"Emotion Recognition from Text Using Knowledge-based ANN" If the sentence has emotional keywords, use of a KBA through an emotional keyword dictionary; if the sentence has no emotional keywords, use of an LBA

		through an Artificial Neural Network (ANN).
(Strapparava and Mihalcea, 2008); KBA	Corpus of 250 news headlines; 6 emotions (Ekman's model)	"Learning to Identify Emotions in Text" 1. Techniques used: a) WNA presence b) some variations of the Latent Semantic Analysis (LSA) algorithm c) Naïve Bayes trained on blogs.
(Tao and Tan, 2004); KBA	Chinese sentences; 6 emotions (Ekman's model)	"Emotional Chinese Talking Head System" 1. Detection of emotional words in the sentences. For example, the authors consider that "unhappy" has equal possibilities of belonging to the states "angry" and "sad", so the word's weight is 0.5 for "angry" and 0.5 for "sad". They use a total of only 390 emotional words. 2. Detection of modifier words as in (Chuang and Wu, 2004). 3. Detection of metaphor words (related to synonyms of the basic emotions). 4. Emotion calculation.
(Yang and Chen, 2011); LBA-bimodal	1240 pop songs (Chinese); Russell's model	"Music Emotion Recognition" 1. Lyrics Feature Extraction: BOW (with and without stemming and stopwords removal). 2. Application of the Probabilistic Latent Semantic Analysis (PLSA) algorithm to discover synonyms and *polysems*[20].

[20] words that have multiple senses and multiple usages in different contexts

		3. Classification: SVM. Audio-only features perform higher for arousal while lyrics perform better for valence.

Table 2.2. Summary of related work.

2.3 Music Lyrics Emotion Variation Detection

Each song is normally associated with 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 relatively easy to determine, i.e., the perceived emotion is the same or is 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.

In some musical genres, the variation of the emotion throughout the song is more common

than in others (e.g., classical music), thus it is important to investigate the time-varying relationship between music and emotion.

We know that human perception from the emotions expressed by a song depends on several dimensions that compose a song (e.g., audio, lyrics).

In audio, according to (Yang and Chen, 2011), there are two known approaches to Music Emotion Variation Detection (MEVD) (Schmidt et al., 2010). The first approach, based on time series analysis (Schubert, 1999) and system identification (Korhonen et al., 2006), exploits the temporal information among the music segments while computing the arousal and valence values. The second approach makes a prediction independently for each music segment as in (Yang et al., 2006). This approach does not consider the temporal dynamical information underlying the music signals.

Concerning lyrics, we are not aware of any research of this kind. However, this is an important issue, as emotion may vary throughout a song, both in the audio and lyrical dimensions.

According to Chopade (Chopade, 2015), emotions may be conveyed by one word or a bunch of words. Sentence level emotion detection 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 transmitted by the individual sentences of that document that successively relies on the emotions transmitted by the individual words. Emotions could be conveyed typically by the person's speech, the face expression and the text (Chopade, 2015).

According to the typical structure of a lyric, based on verses as in poetry or based on sentences as in prose, composers convey ideas and emotions having, as basic unit of information, respectively the verses and the sentences.

The method used to detect emotions in sentences, and after that to understand the way the emotion varies along the lyric, is explained in (Section 4.2 Sentence Emotion Recognition Model (SERM)).

2.4 Our Approach at a Glance: Comparison to the State of the Art

Based on the analysis of the previous sections we will present in the following paragraphs conclusions about some studies from the state of the art (Section 2.2.4) and simultaneously present our approaches.

Concerning the ground truth, all the state of the art datasets are different, as we can see in Section 2.2.4, both in content and in type. They contain different kinds of information, such as song lyrics (Hu et al., 2009b), blogs (Aman and Szpakowicz, 2007), news (Chaffar and Inkpen, 2011), children stories (Alm et al., 2005), books (Agrawal and An, 2012), etc. Even in works tackling the analysis of song lyrics, datasets differ (e.g., different authors, different genres, different type of discourse and different songs). Moreover, the employed emotion models generally vary from work to work. As a consequence, it is difficult to compare different works. This happens mainly because datasets are generally not made public, and so researchers have to create their own datasets.

There are exceptions, as the dataset created by Alm, which is public and is used in several works such as (Alm et al., 2005) and (Agrawal and An, 2012). However, this is a dataset of children stories and we think that this kind of datasets, as well as, for example, datasets composed by news headlines as in (Chaffar and Inkpen, 2011), can in principle attain better classification results than a corpus of generic song lyrics in the task of emotion detection. The reason for this is that the vocabulary is more limited and the discourse is more direct, in the case of the children stories, and more objective, in the case of the news. Lyrics have tendency to a more poetic style and consequently to a more subjective discourse. In any case, the employed dataset cannot be used in our work, which is devoted to music emotion recognition.

Concerning our goals, we have decided to make our dataset public to allow future comparisons. In rigor, due to copyright issues, it will only be partially public: instead of including the lyric, we will only include the Uniform Resource Locator (URL) from which we obtained it. This will allow other researchers to use the exact same text we employed. This contrasts with the approach followed by some authors (e.g., (Hu and Downie, 2007)), who only provide the features extracted from their private lyrics datasets. These are of little use when the research focus is on the proposal of novel, more accurate features.

To annotate the ground truth, we may have, as we have seen in Section 2.2.1, manual or semi-automatic approaches. Semi-automatic approaches (e.g., annotations taken from Last.fm or AllMusic), as in (Hu et al., 2009b) and (Laurier et al., 2008), have as advantage the ease of creation of big dimension datasets with little effort in comparison to the manual approach. However, these kind of approaches may cause some ambiguity (e.g., when a person uses the tag "hate" in Last.fm, this means that the song is about "hate" or this means that the person hates the song?). Manual annotations, following the procedure applied by us and authors like Yang (Yang et al., 2008), allows better control at that level. Another reason to use manual annotation is related to the fact that when someone reads a lyric without knowing the corresponding song, there are emotions conveyed by the reading, which could change if the audio is also listened to. Authors like Hu (Hu et al., 2008) study the relations between the audio and the perceived emotions, and so they ask the annotators to ignore the lyrics in the annotation process. We have worked the other way around, i.e., we "shield" the lyrics annotation by giving the annotators only the lyrics, since our goal is to research exclusively the relations between lyric features and emotions. To make bimodal analysis we also ask users (different users) to annotate the corresponding audio dataset, using the same premise (Hu et al., 2008). We have not seen this perspective in any work in the state of the art and these kind of goals can only be achieved with manual annotation.

As for the features, we use the three categories from the state of the art (Section 2.2.2), renaming linguistic features based on lexicons for semantic-based features and adding a new category that is specific from the lyrics domain and has not existed yet in the state of the art. The category is called structural-based features.

Therefore, in our work we use most of the state of the art features from sentiment analysis and LMER areas. We categorize the employed features into 4 categories:

- **Content-based features.** These features are the baseline in most works and in our case include n-grams with and without stemming and stopwords removal. We also use BOW from POS tags. In this case, unigrams from POS tags correspond to the number of occurrences of each grammatical class in the text (e.g., number of adjectives, when we use frequency as feature representation). In BOW, we use n-grams from 1 to 3.

- **Stylistic-based features.** We use 36 features representing the number of occurrences of 36

different grammatical classes in the lyrics. These features were also used in (Hu, 2010). We employ two features related to the use of capital letters and a novel proposed feature, which counts the number of slang words in the lyrics. Our hypothesis is that this feature may be influent because it is known that some music styles (e.g., hip-hop) have more slang than other style.

- **Structural-based features.** To the best of our knowledge, the features proposed here are completely new in the state of the art. These features are related to the structure of the lyric, such as for example, the number of repetitions of chorus and title or the way the chorus and the other verses are organized.

- **Semantic-based features.** We consider here the linguistic features based on lexicons from the state of the art, because they contain semantic information. Among these lexicons and frameworks are GI, LIWC, ConceptNet, Synesketch, ANEW, DAL (more information about these resources on Section 2.5). We propose 14 new features based on gazetteers built for each one of the 4 quadrants of the Russell's model. These gazetteers were built using resources such as DAL, ANEW, WordNet and WordNet Affect.

We believe we use the most complete set of features of the state of the art, since features extracted from known platforms such as General Inquirer (GI), Linguistic Inquiry and Word Count (LIWC), Synesketch, ConceptNet, as well as novel features proposed by us, are employed.

In each feature category (or set), we use two types of feature representation: frequency and TF-IDF (Zaanen and Kanters, 2010).

Initially, we used four different classification algorithms: SVM, KNN, C4.5 and Naïve Bayes. We performed tests with all the different kinds of features and the performance with SVM was almost always better in comparison to the other algorithms. Therefore, we decided to focus our experiments using only this algorithm. For regression, we resorted to the corresponding algorithm (SVR), as described previously. Most of the state of the art studies use the same algorithms (e.g., (Hu et al., 2009b), (Yang et al., 2008)).

In our experiments, results were validated with repeated stratified 10-fold cross validation (Duda et al., 2000) with 20 repetitions and reported the average performance. During this process,

we performed feature selection and ranking with the ReliefF algorithm (Robnik-Šikonja and Kononenko, 2003).

Hu (Hu et al., 2009b) and Laurier (Laurier et al., 2008) are closer to our work, but they only use binary classification. We use both, binary classification and also multiclass classification to better understand the quality of the features in a real scenario.

They do not have a local perspective about the emotions in lyrics (e.g., in a specific sentence of the lyric) because they just analyze the lyrics as a whole. Besides that approach, we also analyze the emotion in lyrics sentence by sentence, to understand how the emotion varies along the lyric. We have seen this latter approach in several authors such as (Chuang and Wu, 2004), (Agrawal and An, 2012) and (Shaheen et al., 2014) but none of them apply it to song lyrics.

Almost all the studies from the state of the art perform LMER making use of black-box models. Besides using these models, we also aim to create rule-based human-comprehensible models to better understand the relation between features and emotions. The work by Yang and Lee (Yang and Lee, 2009) is the only study, to the best of our knowledge, which uses a similar approach.

2.5 Resources

In this section we present the frameworks for text-feature extraction that we use in our work. Namely, General Inquirer (GI), Linguistic Inquiry and Word Count (LIWC), ConceptNet and Synesketch.

General Inquirer (GI)[21] is a psycholinguistic lexicon composed by a total of 11700 words of the English language. These words were manually annotated into 182 categories (Stone et al., 1966). There are words annotated in more than one category. Therefore, from each lyric a 182-dimension feature vector is extracted.

Linguistic Inquiry and Word Count (LIWC)[22] is a dictionary composed of 2290 words and word stems. Each word or word-stem defines one or more word categories or subdictionaries. For

[21] http://www.wjh.harvard.edu/~inquirer/

[22] http://liwc.wpengine.com/

example, the word 'cried' is part of four word categories: sadness, negative emotion, overall affect, and a past tense verb. Hence, if that word is found in the target text, each of these four subdictionary scale scores will be incremented. As in this example, many of the LIWC categories are arranged hierarchically. All anger words, by definition, will be categorized as negative emotion and overall emotion words[23]. We extract a total of 82 features from this framework.

ConceptNet[24] is a freely available commonsense knowledge-base and natural-language-processing toolkit that supports many practical textual-reasoning tasks over real-world documents right out-of-the-box (without additional statistical training) including the weight for 6 basic emotions: Happy, Sad, Angry, Fearful, Disgusted and Surprised.

Synesketch[25] algorithms analyze the emotional content of text sentences in terms of emotional types (happiness, sadness, anger, fear, disgust, and surprise), weights (how intense the emotion is), and valence (positive or negative). The recognition technique is grounded on a refined keyword spotting method which employs a set of heuristic rules, a WordNet-based word lexicon, and a lexicon of emoticons and common abbreviations.

Besides the previous frameworks we use some known dictionaries such as ANEW and DAL.

The Affective Norms for English Words (ANEW) dictionary (Bradley and Lang, 1999) provides a set of normative emotional ratings for a large number of words (1034) in the English language. This set of verbal materials have been rated in terms of pleasure, arousal, and dominance in order to create a standard for use in studies of emotion and attention.

The Dictionary of Affect Language (DAL) (Whissell, 1989) is an instrument designed to measure the emotional meaning of words and texts. It does this by comparing individual words to a word list of 8742 words which have been rated by people for their activation, evaluation and imagery.

Finally, we use WordNet and WordNet Affect.

WordNet (Miller, 1995) is a large lexical database of English. Nouns, verbs, adjectives and

[23] http://www.kovcomp.co.uk/wordstat/LIWC.html

[24] http://alumni.media.mit.edu/~hugo/conceptnet/

[25] http://krcadinac.com/synesketch/#

adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet Affect (Strapparava and Valitutti, 2004) is an extension of Wordnet, which includes a set of synsets adapted to represent affective concepts correlated with affective words.

Chapter 3

LYRICS CLASSIFICATION AND REGRESSION

I never got lessons. I took influence from Chet Baker, Ian Dury, and Joe Strummer. I don't hear my voice and think, 'Yeah, that's a banging voice!' It's more about putting the right emotions into the right words and the lyrics than anything else to me.

King Krule

In this chapter we present our machine learning system, including the creation of the lyrics dataset, feature extraction, classification and regression experiments to identify the best features for each problem and finally the identification of a set of interpretable linguistic rules that relate features and emotions.

The chapter is structured as described in the following paragraphs.

Section 3.1 Lyrics-Dataset Construction (DT1-L)

This section presents the process of creation of the manual lyrics dataset: Data collection, annotation and validation and assignment of emotion categories to each lyric according to Russell's emotion model. Additionally, we show the process of creation of our larger validation dataset, which is based

on AllMusic's annotations.

Section 3.2. Feature Extraction

We present in this section the four types of features extracted from the lyrics: Content-based, stylistic-based, structure-based and semantic-based. We explain how these features are grouped in the different feature sets.

Section 3.3 Classification and Regression

This section explains the methodology used in the process of classification and regression.

Section 3.4 Results and Discussion

In this section, we present the results achieved for regression (we have two regressors –arousal and valence), and for classification (we have three classification problems – by quadrant categories, by arousal hemispheres and by valence meridians). We present the results achieved in binary classification for each one of the quadrants, then compare our new proposed features with the baseline features to measure its quality. We identify the best features for each one of the previous classification problems and we end up identifying interpretable rules that relate features and emotions and relate features each other. We finish the chapter showing some examples of misclassified lyrics suggesting explanations for this fact.

3.1 Lyrics-Dataset Construction (DT1-L)

As abovementioned, current MER systems either follow the categorical or the dimensional emotion paradigm. It is often argued that dimensional paradigms lead to lower ambiguity, since instead of having a discrete set of emotion adjectives, emotions are regarded as a continuum (Yang et al., 2008). 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 and arousal, as illustrated in Figure 3.1 (repeated here from Figure 2.2 for convenience). This is the emotion model used in our work.

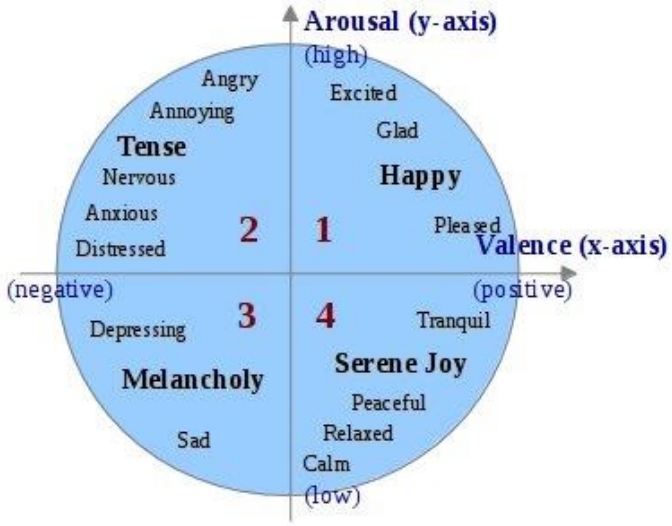

Figure 3.1. Russell's circumplex model (adapted from Yang et al., 2008).

3.1.1 Data Collection

To construct our ground truth, we started by collecting 200 song lyrics. The criteria for selecting the songs were the following:

- Several musical genres and eras (see Table 3.1);
- Songs distributed uniformly by the 4 quadrants of the Russell emotion model;
- Each song belonging predominantly to one of the 4 quadrants in the Russell plane.

To this end, before performing the annotation study described in the next section, the songs were pre-annotated by our team and were nearly balanced across quadrants.

Next, we used the Google API to search for the song lyrics. In this process, three sites were

used for lyrical information: lyrics.com[26], ChartLyrics[27] and MaxiLyrics[28].

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

- Correction of orthographic errors;
- Elimination of songs with non-English lyrics;
- Elimination of songs with lyrics with less than 100 characters;
- 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).

3.1.2 Annotations and Validation

The annotation of the dataset was performed by 39 people with different backgrounds. To better understand their background, we delivered a questionnaire, which was answered by 62% of the volunteers. 24% of the annotators who answered the questionnaire have musical training and, regarding their education level, 35% have a BSc degree, 43% have an MSc, 18% a PhD and 4% have no higher-education degree. Regarding gender balance, 60% were male and 40% were female subjects.

During the process, we recommended the following annotation methodology:

1. Read the lyric;

[26] http://www.lyrics.com/

[27] http://www.chartlyrics.com/

[28] http://www.lyricsmania.com/maxi_lyrics.html

2. Identify the basic predominant emotion expressed by the lyric (if the user thought that there was more than one emotion, he/she should pick the predominant);

3. Assign values (between -4 and 4) to valence and arousal. The granularity of the annotation is the unit, which means that annotators could use 9 possible values to annotate the lyrics, from -4 to 4;

4. Fine tune the values assigned in 3) through ranking of the samples.

To further improve the quality of the annotations, the users were also recommended 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.

We obtained an average of 8 annotations per lyric. Then, the arousal and valence of each song were obtained by the average of the annotations of all the subjects. In this case we considered the average trimmed by 10% to reduce the effect of outliers.

To improve the consistency of the ground truth, the standard deviation (SD) of the annotations made by different subjects for the same song was evaluated. Songs with an SD above 1.2 were excluded from the original set. As a result, 20 songs were discarded, leading to a final dataset containing 180 lyrics. This leads to a 95% confidence interval (Montgomery et al., 1998) of about ±0.4. We believe this is acceptable in our -4.0 to 4.0 annotation range. We can see in the following figure (Figure 3.2) the distribution of the standard deviations in the validated songs.

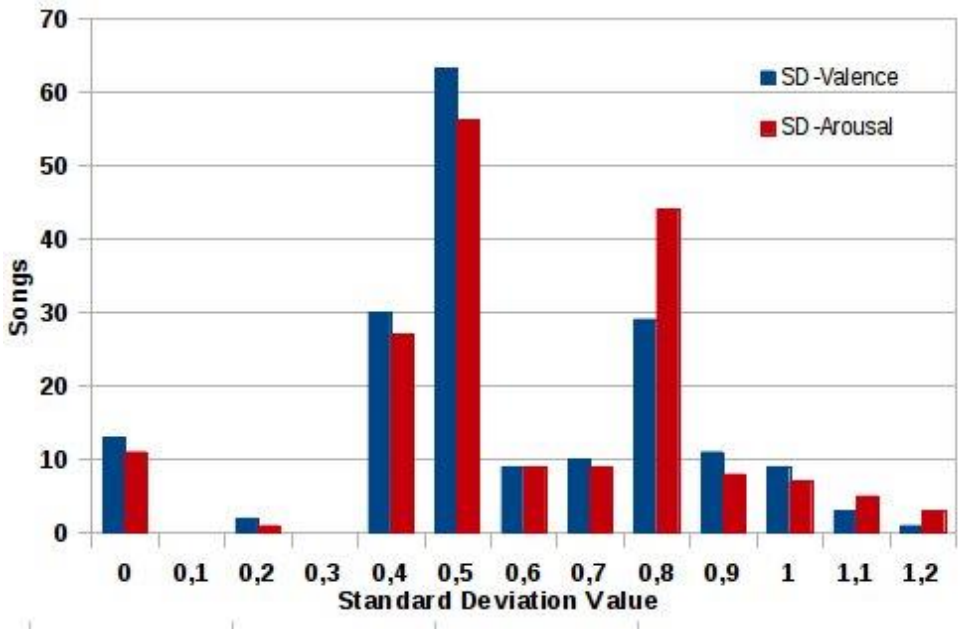

Figure 3.2. Lyrics: Distribution of the Standard Deviations in the Validated Songs.

Finally the consistency of the ground truth was evaluated using Krippendorff's alpha (Krippendorff, 2004), a measure of inter-coder agreement. This measure achieved, in the range -4 up to 4, 0.87 and 0.82 respectively for the dimensions valence and arousal. This is considered a strong agreement among the annotators (Landis and Koch, 1977).

One important issue to consider is how familiar are the lyrics to the listeners. 13% of the respondents reported that they were familiar with 12% of the lyrics (on average). Nevertheless, it seems that the annotation process was sufficiently robust regarding the familiarity issue, since there was an average of 8 annotations per lyric and the annotation agreement (Krippendorff's alpha) was very high (as discussed in the following sections). This suggests that the results were not skewed.

Although the size of the dataset is not large, we think it is acceptable for experiments and is similar to other datasets manually annotated (e.g., (Yang et al., 2008) has 195 songs).

Figure 3.3 and Figure 3.4 show the histogram for arousal and valence dimensions as well as the distribution of the 180 selected songs for the 4 quadrants.

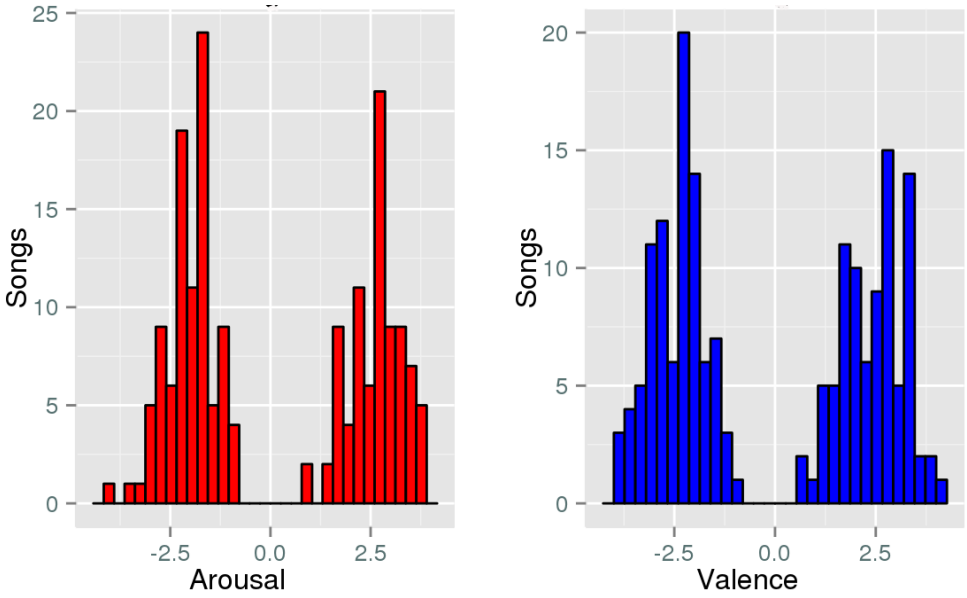

Figure 3.3. Arousal and valence histogram values.

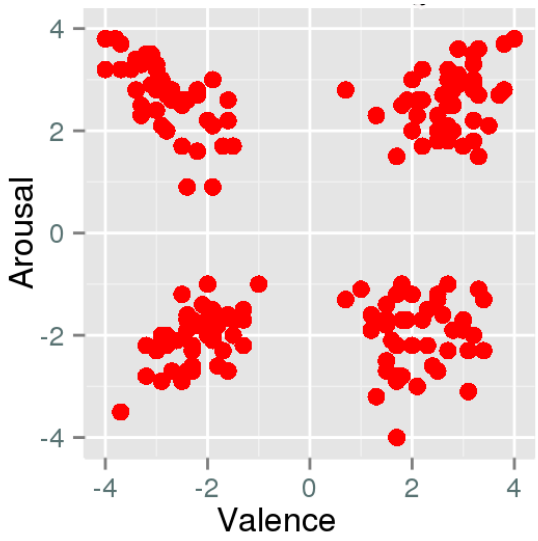

Figure 3.4. Distribution of the songs for the 4 quadrants.

Finally, the distribution of lyrics across quadrants and genres is presented in Table 3.1. We can see that, except for quadrant 2 where almost half of the songs belong to the heavy metal genre, the other quadrants span several genres.

Genre	Q1	Q2	Q3	Q4
Pop/Rock	6	1	15	11
Rock	5	13	13	1
Heavy-metal	0	20	1	0
Pop	1	0	10	6
Jazz	2	0	3	11
R&B	12	0	4	0
Dance	16	0	0	0
New-age	0	0	1	14
Hip-hop	0	7	0	0
Country	1	0	4	1
Reggae	1	0	0	0
Total by Quadrant	44	41	51	44

Table 3.1. Distribution of lyrics across quadrants and genres.

3.1.3 Emotion Categories

Finally, each song is labeled as belonging to one of the four possible quadrants, as well as the respective arousal hemisphere (north or south) and valence meridian (east or west). In this work, we evaluate the classification capabilities of our system in the three described problems.

According to quadrants, the songs are distributed in the following way: quadrant 1 – 44 lyrics; quadrant 2 – 41 lyrics; quadrant 3 – 51 lyrics; quadrant 4 – 44 lyrics (see Table 3.1).

As for arousal hemispheres, we ended up with 85 lyrics with positive arousal and 95 with negative arousal.

Regarding valence meridians we have 88 lyrics with positive valence positive and 92 with negative valence.

3.1.4 Validation Set

To further validate our system, we have also built a larger validation set. This dataset was built in the following way:

1. First, we mapped the mood tags from AllMusic into the words from the ANEW (Affective Norms for English Words) dictionary (ANEW has 1034 words with values for arousal (A) and valence (V)). Depending on the values of A and V, we can associate each word to a single Russell's quadrant. So, from that mapping, we obtained 33 words for quadrant 1 (e.g., fun, happy, triumphant), 29 words for quadrant 2 (e.g., tense, nervous, hostile), 12 words for quadrant 3 (e.g., lonely, sad, dark) and 18 words for quadrant 4 (e.g., relaxed, gentle, quiet).

2. Then, we considered that a song belongs to a specific quadrant if all of the corresponding AllMusic tags belong to this quadrant. Based on this requirement, we initially extracted 400 lyrics from each quadrant (the ones with a higher number of emotion tags), using the AllMusic's web service.

3. Next, used again the Google API to search for the song lyrics (using the sites Lyrics.com, ChartLyrics and MaxiLyrics).

4. Finally, this initial set was validated by three people. Here, we followed the same procedure employed by Laurier (Laurier et al., 2008): a song is validated into a specific quadrant if at least one of the annotators agreed with AllMusic's annotation (Last.fm in his case). This resulted into a dataset with 771 lyrics (211 for Q1, 205 for Q2, 205 for Q3, 150 for Q4). Even though the number of lyrics in Q4 is smaller, the dataset is still nearly balanced.

3.2. Feature Extraction

3.2.1 Content-Based Features (CBF)

The most commonly used features in text analysis, as well as in lyric analysis, are content-based features (CBF), namely the bag-of-words (BOW) (Sebastiani, 2002).

In this model the text in question is represented as a set of bags which normally correspond, in most cases, to unigrams, bigrams or trigrams. The BOW are normally associated to a set of transformations such as stemming and stopwords removal which are applied immediately after the tokenization of the original text. Stemming allows each word to be reduced to its stem and it is assumed that there are no differences, from the semantic point of view, in words which share the same stem. Through stemming the words “argue”, “argued”, “argues”, “arguing” e “argus” would be reduced to the same stem “argu”. The stopwords (e.g., the, is, in, at) which may also be called as function words are very common words in a certain language. These words bring normally little knowledge. The words include mainly determinants, pronouns and other grammatical particles which, by their frequency in a large quantity of documents, are not discriminative. The BOW may also be applied without any of the prior transformations. This technique was used, for example, in (Hu et al., 2009b).

Part-of-speech (POS) tags are another type of state of the art features. They consist in attributing a corresponding grammatical class to each word. For example the grammatical tagging of the following sentence “The student read the book” would be “The/DT student/NN read/VBZ the/DT book/NN”, where DT, NN and VBZ mean respectively determiner, noun and verb in 3rd person singular present. The POS tagging is typically followed by a BOW analysis. This technique was used in studies such as (Mayer et al., 2008).

In our research we use all the combinations of unigrams, bigrams and trigrams with the aforementioned transformations. We also use n-grams of POS tags from bigram to 5-grams.

3.2.2 Stylistic-Based Features (StyBF)

These features are related to stylistic aspects of the language. One of the issues related to the written style is the choice of the type of the words to convey a certain idea (or emotion, in our study). Concerning music, those issues can be related to the style of the composer, the musical genre or the emotions that we intend to convey.

We use 36 features representing the number of occurrences of 36 different grammatical classes in the lyrics. We use the POS tags in the Penn Treebank Project (Taylor et al., 2003) such as for instance JJ (adjectives), NNS (noun plural), RB (adverb), UH (interjection), VB (verb). Some of these features are also used by authors like (Hu et al., 2009b).

We use two features related to the use of capital letters: All Capital Letters (ACL), which represents the number of words with all letters in uppercase and First Capital Letter (FCL), which represents the number of words initialized by an uppercase letter, excluding the first word of each line.

Finally, we propose a new feature: the number of occurrences of slang words (abbreviated as *#Slang*). These slang words (17700 words) are taken from the Online Slang Dictionary[29] (American, English and Urban Slang). We propose this feature because, in specific genres like hip-hop, the ideas are expressed normally with a lot of slang, so we believe that this feature may be important to describe specific emotions associated to specific genres.

3.2.3 Song-Structure-Based Features (StruBF)

To the best of our knowledge, no previous work on LMER employs features related to the structure of the lyric. However, we believe this type of features has relevance for LMER. Hence, we propose novel features of this kind, namely:

- *#CH*, which stands for the number of times the chorus is repeated in the lyric;

[29] http://onlineslangdictionary.com/

- *#Title*, which is the number of times the title appears in the lyric.
- 10 features based on the lyrical structure in verses (V) and chorus (C):
    - *#VorC* (total of sections - verses and chorus - in the lyrics);
    - *#V* (number of verses);
    - *C...* (the lyric starts with chorus – boolean);
    - *#V/Total* (relation between Vs and the total of sections);
    - *#C/Total* (relation between C and the total of sections);
    - *>2CAtTheEnd* (lyric ends with at least two repetitions of the chorus – boolean);
    - (3 features) alternation between versus and chorus, e.g., *VCVC...* (verses and chorus are alternated), *VCCVCC...* (between 2 verses we have at least 1 chorus), *VVCVC* (between 2 chorus we have at least 1 verse).

Usually more danceable songs have more repetitions of the chorus. We believe that the different structures that a lyric may have, are taken into account by the composers to express emotions. That is the reason why we propose these features**.**

3.2.4 Semantic-Based Features (SemBF)

These features are related to semantic aspects of the lyrics. In this case, we used features based on existing frameworks like Synesketch[30] (8 features), ConceptNet[31] (8 features), LIWC[32] (82 features) and GI[33] (182 features).

In addition to the previous frameworks, we use features based on known dictionaries: DAL

[30] http://synesketch.krcadinac.com/blog/

[31] http://web.media.mit.edu/~hugo/conceptnet/

[32] http://www.liwc.net/

[33] http://www.wjh.harvard.edu/~inquirer

(Whissell, 1989) and ANEW (Bradley and Lang, 1999). From DAL (Dictionary of Affect in Language) we extract 3 features which are the average in lyrics of the dimensions pleasantness, activation and imagery. Each word in DAL is annotated with these 3 dimensions. As for ANEW (Affective Norms for English Words) we extract 3 features which are the average in lyrics of the dimensions valence, arousal and dominance. Each word in ANEW is annotated with these 3 dimensions.

Additionally, we propose 14 new features based on gazetteers, which represent the 4 quadrants of the Russell emotion model. We constructed the gazetteers according to the following procedure:

1. We define as seed words the emotion terms defined in Russell's plane (see Figure 3.1).

2. From these emotion terms, we consider for the gazetteers only the ones present in the DAL or the ANEW dictionaries. In DAL, we assume that pleasantness corresponds to valence, and activation to arousal, based on (Fontaine et al., 2013). We employ the scale defined in 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.

Before the insertion of any word in the gazetteer (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 annotators. The two persons involved in the validation were not linguistic scholars but were sufficiently knowledgeable for the task.

Table 3.2 and Table 3.3 illustrate some of the words for each quadrant.

Quadrant 1	Valence	Arousal
Dance	2.29	2.3
Excited	2.5	2.91
Fun	2.84	2.56
Glad	2.75	2.5
Joy	2.88	2.31

Quadrant 2	Valence	Arousal
Afraid	1.25	2.42
Agony	1.36	2.27
Anger	1	2.89
Anxiety	1	2.8
Distressed	1.24	2.35

Table 3.2. Examples of words from the gazetteers 1 and 2.

Quadrant 3	Valence	Arousal
Depressed	1.55	1.83
Gloom	1.25	1.38
Lonely	1	1.27
Sad	1.38	1.43
Sorrow	1.2	1.77

Quadrant 4	Valence	Arousal
Comfort	3	1.33
Cozy	2.6	1.58
Peace	2.68	1.49
Relaxed	2.5	1.35
Serene	2.6	1.22

Table 3.3. Examples of words from the gazetteers 3 and 4.

Overall, the resulting gazetteers comprised 132, 214, 78 and 93 words respectively for the quadrants 1, 2, 3 and 4.

The features extracted are:

- *VinGAZQ1* (average valence of the words present in the lyrics that are also present in the gazetteer of the quadrant 1);
- *AinGAZQ1* (average arousal of the words present in the lyrics that are also present in the gazetteer of the quadrant 1);

- *VinGAZQ2* (average valence of the words present in the lyrics that are also present in the gazetteer of the quadrant 2);
- *AinGAZQ2* (average arousal of the words present in the lyrics that are also present in the gazetteer of the quadrant 2);
- *VinGAZQ3* (average valence of the words present in the lyrics that are also present in the gazetteer of the quadrant 3);
- *AinGAZQ3* (average arousal of the words present in the lyrics that are also present in the gazetteer of the quadrant 3);
- *VinGAZQ4* (average valence of the words present in the lyrics that are also present in the gazetteer of the quadrant 4);
- *AinGAZQ4* (average arousal of the words present in the lyrics that are also present in the gazetteer of the quadrant 4);
- *#GAZQ1* (number of words of the gazetteer 1 that are present in the lyrics);
- *#GAZQ2* (number of words of the gazetteer 2 that are present in the lyrics);
- *#GAZQ3* (number of words of the gazetteer 3 that are present in the lyrics);
- *#GAZQ4* (number of words of the gazetteer 4 that are present in the lyrics);
- *VinGAZQ1Q2Q3Q4* (average valence of the words present in the lyrics that are also present in the gazetteers of the quadrants 1, 2, 3, 4);
- *AinGAZQ1Q2Q3Q4* (average arousal of the words present in the lyrics that are also present in the gazetteers of the quadrants 1, 2, 3, 4).

3.2.5 Feature Grouping

The proposed features are organized into four different feature sets:

CBF. We define 10 feature sets of this type: 6 are BOW (1-gram up to 3-grams) after tokenization with and without stemming (st) and stopwords removal (sw); 4 are BOW (2-grams up to 5-grams) after the application of a POS tagger without st and sw. These BOW features are used as the baseline, since they are a reference in most studies such as (Hu and Downie, 2010b), (Yang et al., 2008).

StyBF. We define 2 feature sets: the first corresponds to the number of occurrences of POS tags in the lyrics after the application of a POS tagger (a total of 36 different grammatical classes or tags); the second represents the number of slang words (*#Slang*) and the features related to words in capital letters (*ACL* and *FCL*).

StruBF. We define one feature set with all the structural features.

SemBF. We define 4 feature sets: the first with the features from Synesketch and ConceptNet; the second with the features from LIWC; the third with the features from GI; and the last with the features from gazetteers, DAL and ANEW.

We use the term frequency and the term frequency inverse document frequency (TFIDF) as representation values in the datasets.

3.3 Classification and Regression

For classification and regression, we use Support Vector Machines (SVM) (Boser et al., 1992), since, based on previous evaluations, this technique performed generally better than other methods. A polynomial kernel was employed and a grid parameter search was performed to tune the parameters of the algorithm. Feature selection and ranking with the ReliefF algorithm (Robnik-Šikonja and Kononenko, 2003) were also performed in each feature set, in order to reduce the number of features. In addition, for the best features in each model, we analyzed the resulting feature probability density functions (pdf) to validate the feature selection that resulted from ReliefF, as described below.

For both classification and regression, results were validated with repeated stratified 10-fold cross validation (Duda et al., 2000) (with 20 repetitions) and the average obtained performance is

reported.

3.4 Results and Discussion

3.4.1 Regression Analysis

The regressors for arousal and valence were applied using the feature sets for the different types of features (e.g., SemBF). Then, after feature selection, ranking and reduction with the ReliefF algorithm, we created regressors for the combinations of the best feature sets.

To evaluate the performance of the regressors the coefficient of determination R^2 (Montgomery et al., 1998) was computed separately for each dimension (arousal and valence). This is a statistic that gives information about the goodness of fit of a model. This measure indicates how well data fit a statistic model. If value is 1, the model perfectly fits the data. A negative value indicates that the model does not fit the data at all.

Suppose a dataset with n values marked as $y_1 \cdots y_n$ (known as y_i), each associated with a predicted value $f_1 \cdots f_n$ (known as f_i). $\bar{y}$ is the mean of the observed data. R^2 is calculated as in (1).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (1)$$

R^2 was computed separately for each dimension (arousal and valence).

The results were 0.61 (with 234 features) for arousal and 0.64 (with 340 features) for valence. The best results were achieved always with RBFKernel (Keerthi and Lin, 2003).

Yang (Yang et al., 2008) made an analogous study using a dataset with 195 songs (using only the audio). He achieved a R^2 score of 0.58 for arousal and 0.28 for valence. We can see that we

obtained almost the same results for arousal (0.61 vs 0.58) and much better results for valence (0.64 vs 0.28). Although direct comparison is not possible, these results suggest that lyrics analysis is likely to improve audio-only valence estimation. Thus, in the near future, we will evaluate a bimodal analysis using both audio and lyrics.

In addition, we used the obtained arousal and valence regressors to perform regression-based classification (discussed below).

3.4.2 Classification Analysis

We conduct three types of experiments for each of the defined feature sets: i) classification by quadrant categories; ii) classification by arousal hemispheres; iii) and classification by valence meridians.

Classification by Quadrant Emotion Categories

Table 3.4 shows the performance of the best models for each one of the features categories (e.g., CBF). For CBF, we considered for example the two best models (M11 and M12). The field *#Features-SelFeatures-FMeasure(%)* represents respectively the total of features, the number of selected features and the results accomplished via the F-measure metric after feature selection.

In the table below (Table 3.4), M1x stands for models that employ CBF features, M2x represents models with StyBF features, M3x StruBF features and M4x SemBF features. The same code is employed in the tables in the following sections.

The model M41 is not significantly better comparing to M11, but is significantly better than the model M42 (at $p < 0.05$). As for statistical significance we use the Wilcoxon rank-sum test.

As we can see, the two best results were achieved with features from the state of the art, namely BOW and LIWC. The results were close to the novel semantic features in M42 (62.7%). The results of the other novel features (M22 and M31) were not so good in comparison to the baseline at least when evaluated in isolation.

Model ID	Description	#Features-SelFeatures-F-Measure(%)
M11(CBF)	BOW (unigrams)	3567-200-**67.9**
M12(CBF)	POS+BOW(trigrams)	4687-700-**59.8**
M21(StyBF)	#POS_Tags	34-20-**49.5**
M22(StyBF)	#Slang+ACL+FCL	3-3-**36.3**
M31(StruBF)	Structural Lyric Features	12-11-**33.5**
M41(SemBF)	LIWC	82-39-**68.2**
M42(SemBF)	Features based on gazetteers	20-20-**62.7**
M43(SemBF)	GI	182-90-**60.3**

Table 3.4. Classification by Quadrants: Best F-Measure results for model.

Table 3.5 shows the results of the combination the best models for each of the features categories. For example C1Q is the combination of the CBF's best models after feature selection, i.e., initially, for this category, we have 10 different models (see Section 3.2.5 Feature Grouping). After feature selection, the models are combined (only the selected features) and the result is C1Q. Then C1Q has 900 features and after feature selection we got a result of 68.2% for F-measure. The classification process is analogous for the other categories.

In Table 3.5, *#Features* represents the total of features of the model, *Selected Features* is the number of selected features and *F-measure* represents the results accomplished via the F-measure metric.

Model ID	#Features	Selected Features	F-Measure (%)
C1Q (CBF)	900	812	68.2
C2Q (StyBF)	23	20	50.4
C3Q (StruBF)	11	11	33.8
C4Q (SemBF)	163	39	72.2
Mixed C1Q+C2Q+C3Q+C4Q	1006	609	77.1

Table 3.5. Classification by Quadrants: Combination of the best models by categories.

As we can see, the combination of the best models of BOW (baseline) keep the results close to the 70% (model C1Q) with a high number of features selected (812). The results of the SemBF (C4Q) are significantly better since we obtain a better performance (72.20%) with much less features (39). It seems that the novel features (M42) have an important role in the overall improvement of the SemBF since the overall results for this type of features is 72.20% (C4Q) and the best semantic model (M41) achieved 68.2%. Finally the mixed classifier (77.1%) is significantly better than the best classifiers by type of feature: C1Q, C2Q, C3Q and C4Q (at $p < 0.05$). As for statistical significance we use the Wilcoxon rank-sum test.

Additionally, we performed regression-based classification based on the above regression analysis. An F-measure of 76.1% was achieved, which is close to the quadrant-based classification (77.1%). Hence, training only two regressor models could be applied to both regression and classification problems with reasonable accuracy.

Finally, we trained the 180-lyrics dataset using the mixed C1Q+C2Q+C3Q+C4Q features, and validated the resulting model using the new large dataset (comprising 771 lyrics). We obtained 73.6% F-measure, which shows that our model, trained in the 180-lyrics dataset, generalizes reasonably well

Classification by Arousal Hemispheres

We perform the same study for the classification by arousal hemispheres. Table 3.6 shows the results attained by the best models for each feature set.

Model ID	Description	#Features-SelFeatures-F-Measure (%)
M11(CBF)	BOW (unigrams)	3567-404-**75.4**
M12(CBF)	POS+BOW(trigrams)	4687-506-**80.7**
M13(CBF)	POS+BOW(bigrams)	700-290-**77.4**
M21(StyBF)	#POS_Tags	34-24-**74.2**
M22(StyBF)	#Slang+ACL+FCL	3-2-**71.5**
M31(StruBF)	Structural Lyric Features	12-8-**67.8**
M41(SemBF)	LIWC	82-50-**77.7**
M42(SemBF)	Features based on Gazeteers	20-8-**78.9**
M43(SemBF)	GI	182-79-**74.8**
M44(SemBF)	SYN+CN	16-8-**59.2**

Table 3.6. Classification by Arousal Hemispheres: Best F-Measure results for model.

The best results (80.7%) are obtained for trigrams after POS (M12). This suggests that the way the sentences are constructed, from a syntactic point of view, can be an important indicator for the arousal hemispheres of the lyrics. The trigram *vb+prp+nn* is an example of an important feature for this problem (taken from the ranking of features of this model). In this trigram, "vb" is a verb in the base form, "prp" is a preposition and "nn" is a noun. Observing the values we find a tendency for higher values for the class Arousal Positive (AP), that is, it seems that quadrants 1 and 2 use more phrasal verbs followed by nouns in the sentences construction, than quadrants 3 and 4. This model (M12) is significantly better than the other classifiers (at $p < 0.05$).

The novel features in StruBF (M31) and StyBF (M22) achieved respectively 67.8% with 8

features and 71.5% with 2 features. These results are above some state of the art features like the features in M44 and these results are accomplished with few features (2 and 8 respectively). The results of the novel features in M42 seem promising since they are close to the best model M12 and with similar values compared to known platforms like LIWC and GI and with less features (8 to 50 and 79 respectively for LIWC and GI). In comparison to the other SBF, the model M42 (new features) is not significantly better than LIWC (M41) but is significantly better than the other semantic models: M43 and M44 (at $p < 0.05$).

Table 3.7 shows the combinations by feature sets and the combination of the combinations respectively.

Model ID	#Features	Selected Features	F-Measure (%)
C1A (CBF)	1690	1098	**79.6**
C2A (StyBF)	26	26	**75.5**
C3A (StruBF)	8	8	**67.8**
C4A (SemBF)	66	64	**81.1**
Mixed C1A+C2A+C3A+C4A	1274	377	**86.3**

Table 3.7. Classification by Arousal Hemispheres: Combination of the best models by categories.

Comparing to best state of the art features (BOW), the best results with the combinations were improved from 79.6% to 86.3%. The mixed classifier (86.3%) is significantly better than best classifiers by type of feature: C1A, C2A, C3A and C4A (at $p < 0.05$).

Classification by Valence Meridians

We perform the same study for the classification by valence meridian. Table 3.8 shows the results of the best models by type of features.

Model ID	Description	#Features-SelFeatures-FMeasure (%)
M13(CBF)	POS+BOW(bigrams)	700-100-**68.5**
M14(CBF)	BOW (unigrams+stemming)	2856-395-**80**
M15(CBF)	BOW(bigrams - tfidf)	18139-600-**62.5**
M22(StyBF)	#Slang+ACL+FCL	3-3-**49.5**
M23(StyBF)	#POS_Tags – tfidf	34-11-**66.3**
M31(StruBF)	Structural Lyric Features	12-4-**56.4**
M41(SemBF)	LIWC	82-15-**81**
M42(SemBF)	Features based on gazetteers	20-16-**81.5**
M43(SemBF)	GI	182-87-**82**

Table 3.8. Classification by Valence Meridians: Best F-Measure results for model.

These results show the importance of the semantic features in general, since the semantic models (M41, M42 and M43) are significantly better than the classifiers of the other types of features (at $p < 0.05$). Features related with the positivity or negativity of the words such as *VinDAL* or *posemo* (positive words) have an important role to these results

Table 3.9 shows the combinations by feature sets and the combination of the combinations respectively.

In comparison to the previous studies (quadrants and arousal), these results are better in general. We can see this in the BOW experiments (baseline-84.2%) where we achieved a performance close to the best combination (C4V). The best results are also in general achieved with less features as we can see in C3V and C4V.

The mixed classifier (89.2%) is significantly better than the best classifiers by type of feature: C1V, C2V, C3V and C4V (at $p < 0.05$).

Model ID	#Features	Selected Features	F-Measure (%)
C1V (CBF)	1095	750	**84.2**
C2V (StyBF)	14	11	**72.2**
C3V (StruBF)	4	4	**56.4**
C4V (SemBF)	39	6	**85.9**
Mixed C1V+C2V+C3V+C4V	859	594	**89.2**

Table 3.9. Classification by Valence Meridians: Combination of the best models by category.

Binary Classification

As a complement to the multiclass problem seen previously, we also evaluated a binary classification (BC) approach for each one of the emotion categories (e.g., quadrant 1). Negative examples of a category are lyrics that were not tagged with that category but were tagged with the other categories. For example (Table 3.10) the BC in the quadrant 1 uses 88 examples, 44 positive examples and 44 negative examples. The latter 44 examples are equally distributed by the other quadrants.

Sets of Emotions	**#lyrics**	**F-Measure (%)**
Quadrant 1	88	88.6
Quadrant 2	82	91.5
Quadrant 3	102	90.2
Quadrant 4	88	88.6

Table 3.10. F-Measure values for BC.

The results in Table 3.10 were reached using 396, 442, 290 and 696 features, respectively for the four sets of emotions (quadrants).

The good results of this classifiers, namely for quadrant 2, indicate that the prediction models can capture the most important features of these quadrants.

The analysis of the most important features by quadrant will be the starting point for the identification of the best features by sets of emotions or quadrants.

3.4.3 New Features: Comparison to Baseline

Considering CBF as the baseline in this area, we thought it would be important to assess the performance of the models created when we add to the baseline the new proposed features. The new proposed features are contained in three categories: StyBF (feature set M22), StruBF (feature set M31) e SemBF (feature set M42). Next, we created new models adding to C1* each one of the previous feature sets in the following way: C1*+M22; C1*+M31; C1*+M42; C1*+M22+M31+M42. In C1*, 'C1' denotes a feature set that contains the combination of the best Content-Based Features – baseline and '1' denotes CBF, as mentioned above; "*" denotes expansion notation, indicating the different experiments conducted: Q denotes classification by quadrants, A by arousal hemispheres and V by valence meridians. These models were created for each of the 3 classification problems seen in the previous section: Classification by quadrants (see Table 3.11); classification by arousal (see Table 3.12); classification by valence (see Table 3.13).

Model ID	Selected Features	F-Measure (%)
C1Q+M22	384	68.9
C1Q+M31	466	68.4
C1Q+M42	576	74.5
C1Q+M22+M31+M42	388	79.8

Table 3.11. Classification by quadrants (baseline + new features).

The baseline model (C1Q) alone reached 68.2% with 812 features selected (Table 3.5). We improve the results with all the combinations but only the models C1Q+M42 and C1Q+M22+M31+M42 are significantly better than the baseline model (at $p < 0.05$). However the

model C1Q+M22+M31+M42 is significantly better (at $p < 0.05$) than the model C1Q+M42. This shows that the inclusion of StruBF and StyBF have improved overall results.

Model ID	Selected Features	F-Measure (%)
C1A+M22	652	80.6
C1A+M31	373	80.4
C1A+M42	690	83.3
C1A+M22+M31+M42	1307	83.3

Table 3.12. Classification by arousal (baseline + new features).

The baseline model (C1A) alone reached an F-measure of 79.6% with 1098 features selected (Table 3.7). We improve the results with all the combinations but only the models C1A+M42 and C1A+M22+M31+M42 are significantly better than the baseline model (at $p < 0.05$). This shows the importance of the semantic features.

Model ID	Selected Features	F-Measure (%)
C1V+M22	679	83.7
C1V+M31	659	82.8
C1V+M42	493	85.8
C1V+M22+M31+M42	88	86.5

Table 3.13. Classification by valence (baseline + new features).

The baseline model (C1V) alone reached an F-measure of 84.2% with 750 features selected (Table 3.9). The models C1V+M42 and C1V+M22+M31+M42 are significantly better than the baseline model (at $p < 0.05$), however C1V+M22+M31+M42 is not significantly better than C1V+M42. This suggests the importance of the SemBF for this task in comparison to the other new features.

In general, the new StyBF and StruBF are not good enough to improve the baseline score, however we got the nearly similar results with much less features: for classification by quadrants we decrease the number of features of the model from 812 (baseline) to 384 (StyBF) and 466 (StruBF). The same happens for arousal classification (1098 features - baseline to 652 - StyBF and 373 – StruBF) and for valence classification (750 features – baseline to 679 – StyBF and 659 – StruBF).

However, the model with all the features is always better (except for arousal classification) than the model with only baseline and SemBF. This shows a relative importance of the novel StyBF and StruBF. It is important to highlight that M22 (StyBF) has only 3 features and M31 (StruBF) has 12 features.

The new SemBF (model M42) seems important because it can improve clearly the score of the baseline. Particularly in the last problem (classification by valence) it requires a much less features (750 down to 88).

3.4.4 Best Features by Classification Problem

We determined, in (Section 3.4.2 Classification Analysis), the classification models with best performance for the several classification problems. These models were built through the interaction of a set of features (from the total of features after feature selection). Some of these features are possibly strong to predict a class when they are alone but others are strong only when combined with other features.

Our purpose in this section is to identify the most important features, when they act alone, for the description and discrimination of the problem's classes.

We will determine the best feature for:

- Arousal (Hemispheres) description – the classes used are negative arousal (AN) and positive arousal (AP)
- Valence (Meridians) description - negative valence (VN) and positive valence (VP)
- Arousal when valence is positive – negative arousal (AN) and positive arousal (AP),

which means quadrant 1 vs quadrant 4

- Arousal when valence is negative – negative arousal (AN) and positive arousal (AP), which means quadrant 2 vs quadrant 3
- Valence when arousal is positive – negative valence (VN) and positive valence (VP), which means quadrant 1 vs quadrant 2
- Valence when arousal is negative – negative valence (VN) and positive valence (VP), which means quadrant 3 vs quadrant 4

In all the situations we identify the 5 features that, after analysis, seem the best features. This analysis starts from the rankings (top 20) of the best features extracted from the models of the (Section 3.4.2 Classification Analysis), with ReliefF. Next, to validate ReliefF's ranking, we compute the probability density functions (pdf) (Montgomery et al., 1998) for each of the classes of the previous problems. Through the analysis of these pdfs we take some conclusions about the description of the classes and identify some of their main characteristics.

The images below show the pdfs of 2 of the 5 best features for the problem of valence description when the arousal is positive (distinguish between 1st quadrant and 2nd quadrant) (Figure 3.5). The features are *M44-Anger_Weight_Synesketch* (a) and *M42-DinANEW* (b).

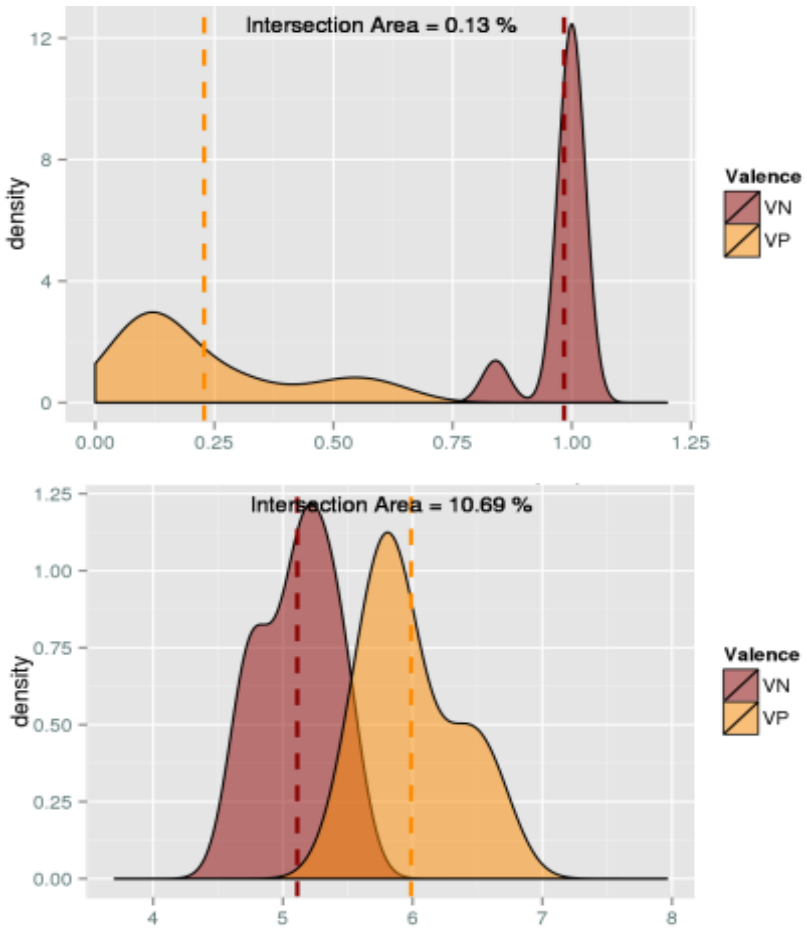

Figure 3.5. PDF of the features a) Anger_Weight_Synesketch and b) DinANEW for the problem of valence description when arousal is positive.

As we can see, the feature in the top image is more important for discriminating between the 1st and 2nd quadrants than the feature in the second image, because the density functions (f) are more separated. We use one measure that indicates this separation: *Intersection_Area*, which represents the intersection area (in percentage) between the two functions.

$$Intersection\_Area = \frac{\int f_A \cap \int f_B}{\int f_A \cup \int f_B} \quad (2)$$

In equation 2, A and B are the compared classes (VN and VP in the example of the Figure 3.5)

and f_A and f_B are respectively the pdfs for A and B.

For this measure, lower values indicate more separation between the curves.

Both features are important to describe the quadrants. The first, taken from the Synesketch framework measures the weight of anger in the lyrics and, as we can see, it has higher values for the 2nd quadrant as expected, since anger is a typical emotion from the 2nd quadrant. The 2nd feature represents the average dominance of the ANEW's words in the lyrics and, although some overlap, it shows that predominantly higher values indicate the 1st quadrant and lower values indicate the 2nd quadrant.

Based on above metric, the top-5 best features were identified for each problem, i.e., the features that separate better the different problems.

Best Features for Arousal Description

As we can see (Table 3.14), the two best features to discriminate between arousal hemispheres are new features proposed by us. *FCL* represents the number of words started by a capital letter and it describes better the class AP than the class AN, i.e., lyrics with *FCL* greater than a specific value correspond normally to lyrics from the class AP. For low values there is a mix between the 2 classes. The same happens to *#Slang*, *#Title*, *WC* (word count - LIWC), *active* (words with active orientation – GI) and *vb* (number of verbs in the base form). The feature *negate* (number of negations – LIWC) has an opposite behavior, i.e., mix between classes for lower values and the class AN from a specific point. The features not listed above, *sad* (words of the negative emotion sadness – LIWC), *angry* (angry weight in ConcepNet) and *numb* (words indicating the assessment of quantity, including the use of numbers – GI) have a similar pattern of behavior as the feature *negate*, while the novel features *CH* (number of repetitions of the chorus) and *TotalVorCH* (number of repetitions of verses or chorus) have similar pattern of behavior as the feature *FCL*.

Feature	Intersection Area
M22-FCL	24.6%
M22-#Slang	29%
M43- active	33.1%
M21- vb	34.2%
M31-#Title	37.4%

Table 3.14. Best features for Arousal description (classes AN, AP).

Best Features for Valence Description

The best features and not only the 5 on Table 3.15, are essentially semantic features. The feature *VinDAL* can describe both classes: lower values are more associated to the class VN and higher values to the class VP. The feature *DinANEW* has a similar pattern but not so good. The features *VinGAZQ1Q2Q3Q4*, *negemo* (words associated with negative emotions - LIWC), *negativ* (words of negative outlook – GI) and *VinANEW* are better for discrimination of the VN class. For the VP class they are not so good. The feature *posemo* (number of positive words – LIWC) for example describes better the VP class.

Feature	Intersection Area
M41- posemo	18.5%
M43- negativ	24.8%
M42-VinDAL	25.6%
M42-VinGAZQ1Q2Q3Q4	25.8%
M42- VinANEW	26.1%

Table 3.15. Best features for Valence description (classes VN, VP).

Best Features for Arousal Description when Valence is Positive

As can be seen in Table 3.16, the features *#GAZQ1*, *FCL*, *iav* (verbs giving an interpretative explanation of an action – GI), *motion* (measures dimension motion – LIWC), *vb* (verbs in base form, *vbn* (verbs in past participle), *active*, *you* (pronouns indicating another person is being addressed directly – GI) and *#Slang* are good for discrimination of the 1st quadrant (higher values associated to the class AP).

The features *angry_CN*, *numb* and *article* (number of articles – LIWC) are good for discrimination of the 4th quadrant. The feature *AinGAZQ1Q2Q3Q4* is good for both quadrants.

Feature	Intersection Area
M42-#GAZQ1	4.6%
M43- active	12.5%
M21- vbn	17.6%
M43- you	17.8%
M21- vb	18.7%

Table 3.16. Best features for Arousal (V+) (classes AN, AP).

Best Features for Arousal Description when Valence is Negative

These features are summarized in Table 3.17. The features *Anger_Weight_Synesketch* and *Disgust_Weight_Synesketch* (weight of the emotion disgust) are good to discriminate between the quadrants 2 and 3 (higher values are associated as it was predictable to instances from the quadrant 2), although in the latter we have more overlap between the classes than in the prior. The features *vbp* (verb, non-3rd person singular present) and *anger* can discriminate the class AP (higher values) but for lower values we have a mix between the classes. Other features with similar behavior are *FCL*, *#Slang*, *negativ* (negative words - GI), *cc* (number of coordinating conjunctions) and *#Title*. *AinGAZQ2* and *past* can discriminate the 3rd quadrant, i.e., the class AN. Finally the feature *article*

(the number of definite, e.g., the, and indefinite, e.g., a, an, articles in the text) can discriminate both quadrants (tendency for 3rd quadrant with lower values and 2nd quadrant with higher values).

Feature	Intersection Area
M44-Anger_ Weight_Synesketch	7.9%
M42- AinGAZQ2	16.2%
M21-vbp	17.8%
M41-anger	21.1%
M21- cc	25.4%

Table 3.17. Best features for Arousal (V-) (classes AN, AP).

Best Features for Valence Description when Arousal is Positive

The feature *Anger_Weight_Synesketch* is clearly discriminative to separate the quadrants 2 and 3 (see Table 3.18 and Figure 3.5). The novel semantic features *VinANEW*, *VinGAZQ1Q2Q3Q4*, *VinDAL* and *DinANEW* have a similar pattern behavior to the first feature but with a little overlap between the functions. The features *negemo* (negative emotion words – LIWC), *swear* (swear words – LIWC), *negative* (words of negative outlook – GI) and *hostile* (words indicating an attitude or concern with hostility or aggressiveness – GI) are good for the discrimination of the 2nd quadrant (higher values).

Feature	Intersection Area
M44-Anger_ Weight_Synesketch	0.1%
M42- VinANEW	4.4%
M42- VinGAZQ1Q2Q3Q4	7.2%
M42- VinDAL	7.7%
M42- DinANEW	10.7%

Table 3.18. Best features for Valence (A+) (classes VN, VP).

Best Features for Valence Description when Arousal is Negative

The best features for valence discrimination when arousal is negative are presented in Table 3.19.

Between the quadrants 3 and 4, the features *vbd*, *I*, *self* and *motion* are better for the 3rd quadrant discrimination, while the features *#GAZQ4*, *article*, *cc* and *posemo* are better for 4th quadrant discrimination.

Feature	Intersection Area
M41- posemo	15.6%
M43- self	24.9%
M21-vbd	27%
M42-#GAZQ4	28.4%
M41- motion	29.2%

Table 3.19. Best features for Valence (A-) (classes VN, VP).

Best Features by Quadrant

Until now we have identified features important to discriminate, for example, between two quadrants. Next, we will evaluate if these features can discriminate completely the four quadrants, i.e., one quadrant against the other three.

To evaluate the quality of the discrimination of a specific feature concerning a quadrant Qz, we have established a metric based on two measures:

- Discrimination support (support of a function is the set of points where the function is not zero-valued (Folland, 1999)), which corresponds to the difference between the total support of the two pdf (Qz and Qothers) and the support of the Qothers pdf. The result is the support of the Qz pdf except the support of the intersection area and is in percentage of the total support. The higher this metric the better (3);

$$Discri\,min\,ation\_sup\,port = \frac{len\left(sup\left(f_{Q_Z} \cup f_{Q_{others}}\right)\right) - len\left(sup\left(f_{Q_{others}}\right)\right)}{len\left(sup\left(f_{Q_Z} \cup f_{Q_{others}}\right)\right)} \quad (3)$$

In (3), *len(sup(f))* stands for the length of the support of function f and f_{Q_Z} and $f_{Q_{others}}$ are respectively the pdfs for Qz and Qothers.

- Discrimination area, which corresponds to the difference between the area of the Qz's pdf and the intersection area between the two pdf. The result is in percentage of the Qz's pdf total area. The higher this metric the better (4).

$$Discri\,min\,ation\_area = \frac{\int f_{Q_Z} - \left(\int f_{Q_Z} \cap \int f_{Q_{others}}\right)}{\int f_{Q_Z}} \quad (4)$$

In this analysis (Table 3.20), we have experimentally defined a minimum threshold of 30% for the Discrimination_Support. To do the ranking of the best features, we use the metric Discrimination_support and in case of a draw, we use the metric Discrimination_Area.

Feature	**Disc_Support / Disc_Area (%)**	**Quadrant**
M42_#GAZQ1	75.4 / 66.3	Q1
M43_socrel	62.4 / 29.5	Q1
M43_solve	60.8 / 25.8	Q1
M41_humans	59.1 / 28.6	Q1
M43_passive	48.1 / 29.2	Q1
M31- #Title	41.1 / 36.2	Q1
M21- vbp	40.3 / 32.8	Q1
M44_Happy_CN	39.7 / 19.9	Q1
M44_CN-A	30.1 / 22.1	Q1
M41-anger	84.9 / 74	Q2

M21-vbg	56 / 30.6	Q2
M43_negativ	52.7 / 51.4	Q2
M22- #Slang	52.7 / 33.5	Q2
M41- negemo	50.2 / 52	Q2
M21-nn	49.7 / 31.5	Q2
M41-WC	49.3 / 32.1	Q2
M43_wittot	46.5 / 23.5	Q2
M22- FCL	46.1 / 36.6	Q2
M21-dt	45.7 / 31.2	Q2
M43-hostile	45.2 / 45.6	Q2
M21-cc	45.1 / 30.5	Q2
M21-prp	40 / 36	Q2
M42-#GAZQ3	63.3 / 41.3	Q3
M41-negate	38.9 / 33.8	Q3
M41-cogmech	32.9 / 19.9	Q3
M42-VinGAZQ1Q2Q3Q4	32.4 / 10.5	Q3
M42-#GAZQ4	56.1 / 36.8	Q4
M41-Dic	47.2 / 17.8	Q4
M41-hear	46 / 19.5	Q4
M31-totalVorCH	40.7 / 27.8	Q4
M42- DinDAL	39.3 / 20.9	Q4

Table 3.20. Type of discrimination of the features by quadrant.

Among the features that best represent each quadrant, we have features from the state of the art, such as features, from LIWC (M41) – *humans* (references to humans), *anger* (affect words), *negemo* (negative emotion words), *WC* (word count), *negate* (negations), *cogmech* (cognitive

processes), *Dic* (dictionary words) and *hear* (hearing perceptual process); from GI (M43) – *socrel* (words for socially-defined interpersonal processes), *solve* (words referring to the mental processes associated with problem solving), *passive* (words indicating a passive orientation), *negativ* (negative words) and *hostile* (words indicating an attitude or concern with hostility or aggressiveness); from ConcepNet (M44) - *happy_CN* (happy weight), *CN_A* (arousal weight); from POS Tags (M21) – *vbp* (verb, non-3rd person singular present), *vbg* (verb, gerund or present participle), *nn* (noun, singular or mass), *dt* (determiner), *cc* (coordinating conjunction) and *prp* (personal pronoun). We have also novel features, such as, StyBF (M22) – *#Slang* and *FCL*; StruBF (M31) - *#Title* and *TotalVorCH*; SemBF (M42) - *#GAZQ1*, *#GAZQ3*, *VinGAZQ1Q2Q3Q4*, *#GAZQ4* and *DinDAL*.

Some of the more salient characteristics of each of the quadrants:

- Q1: typically lyrics associated to songs with positive emotions and high activation. Songs from this quadrant are often associated to specific musical genres, such as, dance, pop and by the importance of the features we point out the features related with repetitions of the chorus and title in the lyric.

- Q2: we point out stylistic features such as *#Slang* and *FCL* that indict high activation with predominance of negative emotions or features that are related with negative valence such as *negativ* (negative words), *hostile* (hostile words) and *swear* (swear words). This kind of features influence more Q2 than Q3 (although Q3 have also negative valence) because Q2 is more influenced by specific vocabulary such as the vocabulary in that features, while Q3 is more influenced by negative ideas, so we think that it is more difficult the perception of emotions in the 3rd quadrant.

- Q3: we point out the importance of the verbal tense (past) in comparison with the other quadrants which have the predominance of the present tense. On the contrary, Q2 have also some tendency to the gerund tense and the Q1 to the present simple. We highlight also in comparison with the other quadrants more use of the 1st singular person (I).

- Q4: Features related with activation, as we have seen for the quadrants 1 and 2, have low weight for this quadrant. We point out the importance of a specific vocabulary as we have in *#GAZQ4*.

Generally, semantic features are more important to discriminate the valence (e.g. *VinDAL*, *VinANEW*). Features important for sentiment analysis such as *posemo* (positive words) or *ngtv* (negative words) are also important for valence discrimination.

On the other hand, stylistic features related with the activation of the written text such as *#Slang* or *FCL* are important for arousal discrimination. Features related with the weight of emotions in the written text are also important (e.g. *Anger_Weight_Synesketch*, *Disgust_Weight_Synesketch*).

3.4.5 Interpretability

After we have made a study to understand the best features to describe and discriminate each set of emotions, we are going to extract some rules/knowledge that allow to understand how these features and emotions are related. With this study we intend to attain two possible goals: i) find out relations between features and emotions (e.g., if feature A is low and feature B is high then the song lyrics belong to quadrant 2); ii) find out relations among features (e.g., song lyrics with feature A high also have feature B low).

Relations Between Features and Quadrants

In this analysis we use the Apriori algorithm (Agrawal et al., 1993).

First, we pre-processed the employed features through the detection of features with a nearly uniform distribution, i.e., the feature values depart at most 10% from the feature mean value. We did not consider these kind of features. Here, we employed all the features selected in Mixed C1Q + C2Q + C3Q + C4Q model (see Table 3.5), except for the ones excluded as described. In total, we employed 144 features.

Then we defined the following premises.

- Consideration of only rules up to 2 antecedents. It was applied an algorithm to eliminate redundancy, considering the more generic rules to avoid complex rules;
- Due to the fact that n-grams features are sparse, we did not consider rules with part of

the antecedent of type n-gram = Very Low. It means probably that the feature does not exist;

- Features were discretized in 5 classes using equal-frequency discretization: very low (VL), low (L), medium (M), high (H), very high (VH). Rules containing non-uniform distributed features were ignored.

We considered two measures to assess the quality of the rules: confidence and support. The ideal rule has simultaneously high representativity (support) and high confidence degree.

Table 3.21 shows up the best rules for quadrants. We defined a threshold of support = 7.2% (15 lyrics) and confidence = 60%.

We think these rules are in general self-explanatory and understandable, however we will explain some of them not so explicit.

We can see for Q1 the importance of the feature *#GAZQ1* together with the feature from GI, *afftot* (words in the affect domain), both with VH values. We can also highlight for this quadrant the relation between a VL weight for *sadness* and a VH value for the feature *positiv* (words of positive outlook) and the relation between a VH number of title's repetitions in the lyric and a VL weight for the emotion angry.

We can point out for quadrant 2 the importance of the features *anger* from LIWC and Synesketch, *negemo_GI* (negative emotion), *#GAZQ2*, *VinANEW*, *hostile* (words indicating an attitude or concern with hostility or aggressiveness), *powcon* (words for ways of conflicting) and some combinations among them.

For quadrant 3, we can point out the relation between a VH value for the emotion sadness and a VL value for the number of swear words in the lyrics.

For quadrant 4 we can point out the relation between the features *anger* and *weak* (words implying weakness) both with VL values.

These results confirm the results reached in the previous section, where we identified the most important features for each quadrant.

#	Rule	Support / Confidence (%)
1	#GAZQ1=VH ==> Q=Q1	13.8 / 80
2	#GAZQ1=VH and afftot_GI=VH => Q1	8.8 / 72
3	sad_LIWC=VL and positiv_GI=VH => Q1	7.7 / 82
4	#Title=VH and angry_CN=VL => Q1	7.2 / 72
5	VinANEW=VL => Q2	20 / 61
6	hostile_GI=VH and Sadness_Weight_Synesketch=VH => Q2	14.4 / 69
7	Anger_Weight_Synesketch=VH and Valence_Synesketch=VL => Q2	12.7 / 76
8	anger_LIWC=H => Q2	11.1 / 85
9	negemo_GI=VH => Q2	11.1 / 67
10	#GAZQ2=VH => Q2	10.5 / 100
11	Anger_Weight_Synesketch=VH and negemo_LIWC=VH => Q2	8.8 / 94
12	anger_LIWC=VH => Q2	8.8 / 100
13	VinGAZQ2=VH => Q2	8.3 / 83
14	hostile_GI=VH and powcon_GI=VH => Q2	8.3 / 78
15	sad_LIWC=VH and swear_LIWC=VL => Q3	8.8 / 72
16	dt=VL and article_LIWC=VL => Q3	8.3 / 71
17	dt=VL and Valence_Synesketch=VL => Q3	8.3 / 71
18	anger_LIWC=VL and weak_GI=VL => Q4	10 / 72
19	swear_LIWC=VL and #GAZQ4=VH => Q4	9.4 / 73
20	#Slang=VL and #GAZQ2=VL => Q4	8.8 / 76
21	prp=VL and #GAZQ2=VL => Q4	8.8 / 72

Table 3.21. Rules from classification association mining.

Relations Among Features

The same premises concerning discretization were applied as in the prior section.

We have considered rules with a minimum representativity (support) of 10% and a minimum confidence measure of 95%. After that all the rules were analyzed and redundant rules were removed.

The results show (Table 3.22) only the more representative rules and are in consonance with what we suspected after the analysis made in the last sections.

We will analyze briefly the scope of the prior rules.

(Rule 1) The feature *GI_passive* (words indicating a passive orientation) has, for the class VH, almost all the songs in the quadrants 1 and 2. The same happens for the features *vb* (verb in base form) and *prp* (personal pronouns). We would say that this rule reveals an association among the features namely for positive activation.

(Rule 2) *GI_intrj* (includes exclamations as well as casual and slang references, words categorized "yes" and "no" such as "amen" or "nope", as well as other words like "damn" and "farewell") and *GI_active* (words implying an active orientation) both with values very high imply a VH value for the feature *GI_iav* (verbs giving an interpretative explanation of an action, such as "encourage, mislead, flatter"). This rule is predominantly true for the quadrant 2.

(Rule 3) the features *#Slang* and *you* (pronouns indicating another person is being addressed directly) have higher values for quadrant 2 and this implicate and higher number of *prp* in the written style. This is typical from genres like hip-hop.

(Rule 4) Almost all the samples with a value VL for the feature *VinANEW* are in the quadrants 2 (more) and 3 (less). *Fear_Weight_Synesketch* has a VH value essentially in the quadrant 2. *Sadness_Weight_Synesketch* has higher values for quadrants 3 and 2, so probably this rule is applied more on songs of quadrant 2.

(Rule 5) We can see the association among the features *#Slang*, *FCL*, *dav* (verbs of an action or feature of an action, such as run, walk, write, read) and *WC* (word count), all of them with high values and we know that this rule is more associated with the 2nd quadrant.

#	Association rules	Support / Confidence (%)
1	GI_passive=VH and vb=VH => prp=VH	20 / 100
2	GI_intrj=VH and GI_active=VH => GI_iav=VH	19 / 100
3	#Slang=VH and GI_you=VH => prp=VH	18 / 100
4	VinANEW=VL and Fear_W_Syn=VH => Sadness_W_Syn=VH	18 / 100
5	#Slang=VH and FCL=VH and dav=VH => WC=VH	18 / 100
6	strong=VH and GI_active=VH => iav=VH	22 / 95
7	#Slang=VL and prp=VL => WC=VL	21 / 95
8	#Slang=VL and FCL=VL => WC=VL	21 / 95
9	vb=VH and GI_you=VH => prp=VH	21 / 95
10	#Slang=VH and jj=VH => WC=VH	19 /95
11	VinGAZQ1Q2Q3Q4=VL and Fear_W_Syn=VH => Sadness_W_Syn=VH	19 / 95
12	#Slang=VL and active=VL => strong=VL	19 / 95
13	FCL=VH and active=VH => iav=VH	19 / 95

Table 3.22. Rules from association mining.

(Rule 6) This rule is more associated to the quadrants 1 and 2. High values for the features *strong* (words implying strength), *active* and *iav*.

(Rules 7 and 8) Almost all the songs with *#Slang*, *prp*, *FCL* and *WC* equal to VL, belong to the quadrants 3 and 4.

(Rule 9) The feature *vb* has higher values for quadrant Q2 followed by quadrant Q1 while feature *GI_you* has higher values for quadrant Q2 followed by the quadrant 3. *Prp* with VH values is predominantly in the quadrant 2, so this rule is probably more associated to the quadrant 2.

(Rule 10) These features, *#Slang*, *jj* (number of adjectives) and *WC* have VH values essentially for the quadrants 1 and 2.

(Rule 11) This rule is probably more applied in the quadrants 2 or 3, since the feature *VinGAZQ1Q2Q3Q4* has predominantly lower values for quadrants 2 and 3, while *Fear_Weight_ Synesketch* has higher values in the same quadrants.

(Rule 12) The three features have VL values essentially for the quadrants 3 and 4.

(Rule 13) The three features have VH values essentially for the quadrants 1 and 2.

3.4.6 Misclassified Lyrics

Having into account the results achieved in this chapter, we are going to enumerate some misclassified lyrics and try to suggest why they are misclassified.

The lyric "Dance with my father" (below) from Luther Vandross is about very beautiful memories of a daughter about her father of when he was alive. The overall feeling conveyed by the lyric is sadness, because he passed away and she would give everything to have him with her.

This lyric was classified by the annotators with an average of -1.5 for valence and -1 for arousal in a range between -4 and 4. The lyric was in some models wrongly classified in Q1.

Back when I was a child
Before life removed all the innocence
My father would lift me high
And dance with my mother and me
And then
Spin me around 'till I fell asleep
Then up the stairs he would carry me
And I knew for sure
I was loved

If I could get another chance
Another walk
Another dance with him
I'd play a song that would never ever end
How I'd love love love
To dance with my father again

When I and my mother
Would disagree
To get my way I would run
From her to him
He'd make me laugh just to comfort me
yeah yeah
Then finally make me do
Just what my mama said
Later that night when I was asleep
He left a dollar under my sheet
Never dreamed that he
Would be gone from me

If I could steal one final glance
One final step
One final dance with him
I'd play a song that would never ever end
Cause I'd love love love to
Dance with my father again

…

A possible hypothesis for this wrong classification can be related to the fact that the descriptions in the lyric are all positive and with a lot of tenderness, i.e., the description corresponds only to good moments. The feeling of loss (the bad parts of the loss) is not explicitly written in the lyric.

We use the same reasoning to suggest why the lyric “Tears in Heaven” from Eric Clapton classified by the annotators in Q3 was incorrectly classified in Q4. The lyric is about the death of the

4 years old Eric Clapton's son.

The lyric "La Isla Bonita" de Madonna is annotated in the dataset in Q1 (valence equal to 2.5 and arousal equal to 2), but the predicted quadrant is Q4. According to Madonna in an interview to New York Times, this song is a tribute to the beauty and mystery of Latin American people (Kutner and Leigh, 2005). Although the arousal equal to 2, we think that the lyric is very tranquil and inspires relaxation, so we understand this classification in Q4. Another possible justification is the big number of keywords (CBF – unigrams) present in the lyric and normally associated to a states of calm and relaxation (e.g., breeze, nature, sea, sun, sky).

A last example, the song "Animal" from Pearl Jam was annotated in Q2, but the system classified it in Q4. According the site http://songmeanings.com the song is ambiguous. Some people say that the song is about animal rights, others say that the song is about the pressure of the Media against Pearl Jam. Sentences like "torture from you to me, abducted from the street" show possible reasons to the classification in Q2. The classification in Q4 suggests that the system have considered that the big frequency of certain words as numerals (e.g., one, two, three) or the word "animal" in its literal sense are possible reasons to the classification in Q4, since, they are frequent in lyrics of Q4.

In a general way, we have observed that lyrics written in a poetic way are normally more ambiguous and so they may have more interpretations.

Chapter 4

MUSIC-LYRICS EMOTION VARIATION DETECTION

My music and my lyrics are essentially emotional postcards

Sarah McLachlan

After analyzing the lyrics as a whole through classification and regressions approaches to find out the best prediction models and the best features regarding the quadrants and emotions of the Russell's model, we aim to study how the emotions vary along the lyric.

To attain this goal, we built a Sentence Emotion Recognition Model (SERM) using a keyword-based approach.

In this work, we consider the terms sentence and verse interchangeably for two reasons:

- Lyrics do not often have punctuation marks or clear delimiters of the sentences. Hence, it is difficult to apply algorithms to split text into sentences as we could apply easily with a more formal text (e.g., journalistic);
- Composers express normally their ideas in the lyrics through verses. Even when, for instance,

the verse is composed of two phrases, they generally complement each other to convey a unique idea. Other verses can convey more than one idea, but for normalization issues and to ensure the consistency of our classification model, independently of the type of lyric, we regard the verse as the basic unit for emotion recognition.

The chapter is structured as described in the following paragraphs.

Section 4.1 Sentence Dataset Construction (DT2)

This section shows the process of creation of the training and the validation datasets, including the data collection, annotation and validation stages.

Section 4.2 Sentence Emotion Recognition Model (SERM)

In this section we present the full process of creation of our keyword-based system.

Section 4.3 Results and Discussion

In this step, we present the process we use to optimize the parameters of SERM. We end up with results and discussion of SERM applied to the validation dataset.

4.1 Sentence Dataset Construction (DT2)

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.

4.1.1 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 (Table 4.1):

Genres	**#Songs**
Pop/Rock	6
Pop	18
Rock	8
Heavy-Metal	3
Folk	2
R&B	1
Hip-Hop	4
Country	2

Table 4.1. Distribution of genres by the songs in DT2.

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

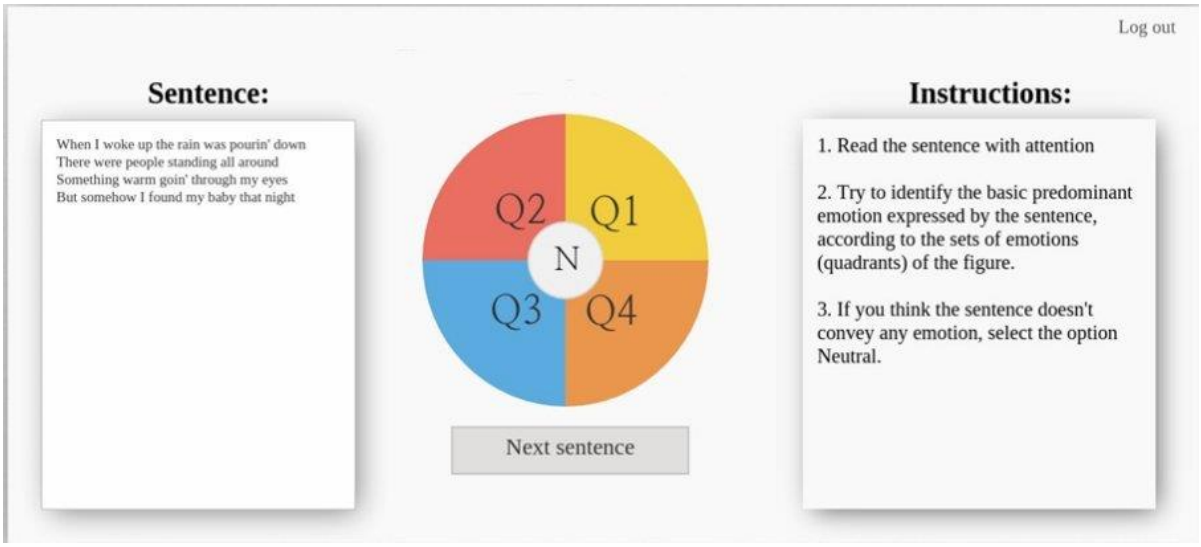

Figure 4.1. Main screen of the annotation platform.

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 4.2);
3. If you think the sentence does not convey any emotion, select the option Neutral.

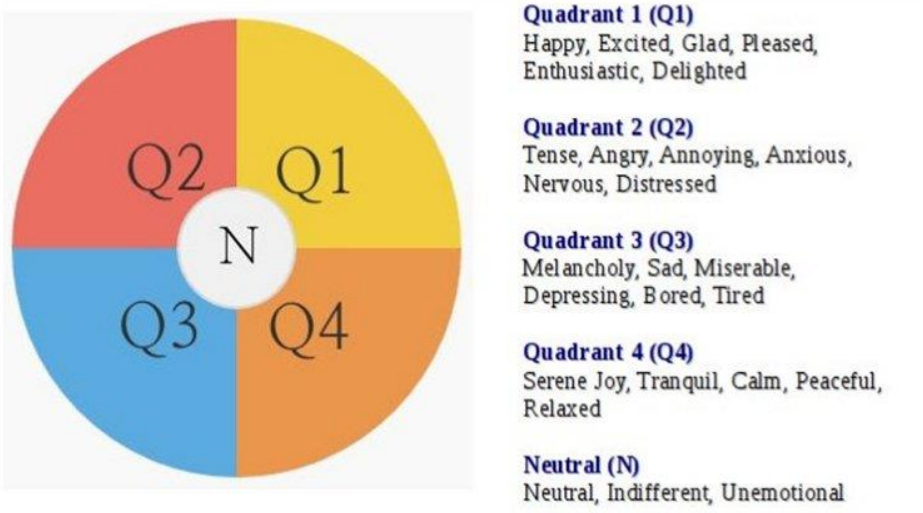

Figure 4.2. Predominant emotions by quadrant.

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 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 4.2):

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

Table 4.2. Distribution of the sentences by quadrant.

As can be observed in Table 4.2, 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.

4.1.2 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 4.3. As can be seen, this training is nearly balanced.

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

Table 4.3. Distribution of the sentences by quadrant.

4.2 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 4.3 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 Section 3.2.4 Semantic-Based Features

(SemBF)).

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 et al., 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., 2006b) 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 (Table 4.4).

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

Table 4.4. Example of modifiers in sentences.

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 4.5 shows some examples of adverbs classified according to its intensity.

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

Table 4.5. Examples to the weight of the word "happy" in sentences with adverb modifiers.

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).

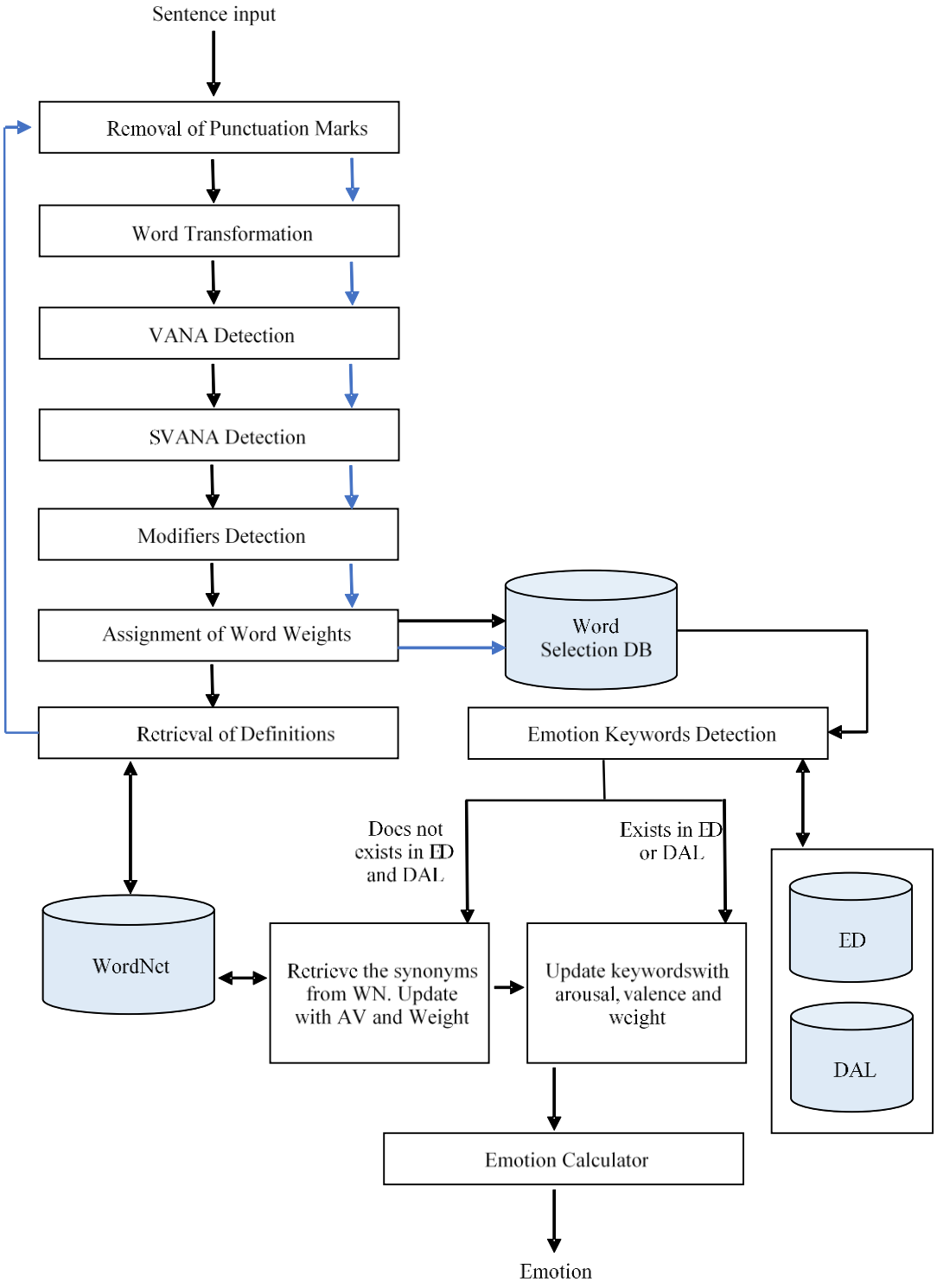

Figure 4.3. Architecture of the Sentence Emotion Recognition Model (SERM).

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 4.6 the weight of the object “happy” for sentences using the previous adverbs as 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

Table 4.6. Examples to the weight of the word “happy” in sentences with adverb 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).

Retrieval of Definitions

The system retrieves the definition of the selected words, 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, SVANA 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 (Estimator)

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.

4.3 Results and Discussion

4.3.1 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 (Table 4.9).

Table 4.7 shows the best values for each weight.

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

Table 4.7. Statistics for the best training model.

Table 4.8 shows the confusion matrix using these parameters.

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%**	

Table 4.8. Statistics for the best training model.

We can see that this combination of weights achieved a performance of 70.3% (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 chapter 3.

We can see the comparison of results of the 10 best models in Table 4.9.

WL					Statistics		
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%

Table 4.9. Statistics for the best 10 training models.

4.3.2 Classification of Sentences

We applied SERM with the selected parameters to our sentence validation dataset. The achieved results are summarized in Table 4.10.

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%**	

Table 4.10. Statistics for the validation model.

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

In Table 4.10, 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 46.8%) 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 (59.3%). This is due to the sentences from Q1 and Q3 being erroneously classified in Q4 (see example below).

At the end of (Section 4.1.1 Validation Set), we illustrated the annotation results for 4

sentences of the dataset. Table 4.11 and the text below show the predicted classes for these sentences and possible explanations for the errors.

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

Table 4.11. Classification with SERM of several sentences.

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 behavior 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 (see Section 3.4.2 Classification Analysis).

Sentence	Actual	Pred.
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, so much	Q1	Q2
You're the fear, I don't care 'Cause I've never been so high Follow me to the dark Let me take you past our satellites You can see the world you brought to life, to life	Q2	Q2
So love me like you do, lo-lo-love me like you do Love me like you do, lo-lo-love me like you do Touch me like you do, to-to-touch me like you do What are you waiting for?	Q1	Q1
Fading in, fading out On the edge of paradise Every inch of your skin is a holy gray I've got to find Only you can set my heart on fire, on fire	Q1	Q4
Yeah, I'll let you set the pace 'Cause I'm not thinking straight My head spinning around I can't see clear no more What are you waiting for?	Q2	Q2
Love me like you do, lo-lo-love me like you do Love me like you do, lo-lo-love me like you do Touch me like you do, to-to-touch me like you do What are you waiting for?	Q1	Q1
Love me like you do, lo-lo-love me like you do (like you do) Love me like you do, lo-lo-love me like you do (yeah) Touch me like you do, to-to-touch me like you do What are you waiting for?	Q1	Q1
I'll let you set the pace 'Cause I'm not thinking straight My head spinning around I can't see clear no more What are you waiting for?	Q2	Q2
Love me like you do, lo-lo-love me like you do (like you do) Love me like you do, lo-lo-love me like you do (yeah) Touch me like you do, to-to-touch me like you do What are you waiting for?	Q1	Q1
Love me like you do, lo-lo-love me like you do (like you do) Love me like you do, lo-lo-love me like you do (yeah) Touch me like you do, to-to-touch me like you do What are you waiting for?	Q1	Q1

Table 4.12. Using SERM to classify the song "Love me like you do" from Ellie Goulding.

To test our model in a real scenario of a whole song, we will show the application in the song *Love me like you do* from Ellie Goulding (Table 4.12). This song has 10 sentences, 7 of them were annotated in Q1 and 3 in Q2. The developed classifier predicts correctly 80% of the sentences.

For example, the first sentence was annotated in Q1 for the annotators, but since the sentence has several negative words more associated to Q2 (e.g., *pain*, *blood*, *night*), because it has some antithesis, the classifier is confused with that and classified it in Q2.

4.4 Comparing SERM with a Supervised ML Classifier

After the application of this KBA to the sentences it is important to validate the results through comparison to other works or approaches. To the best of our knowledge, we are not aware of any study that performs MER to a dataset of this kind, so we decided then to use a supervised machine learning approach to classify the sentences in emotions.

In Chapter 3 we have used a ML approach to classify lyrics in emotions using the same emotion model. In that study we have two datasets of lyrics. The first one has 180 lyrics from several genres and eras and was annotated manually by 39 annotators. According to quadrants, the songs are distributed in the following way: Q1 – 44 lyrics; Q2 – 41 lyrics; Q3 – 51 lyrics; Q4 – 44 lyrics.

The second dataset has 771 lyrics and was annotated from AllMusic's tags and then validated by three people. Here, we followed the same procedure employed by Laurier et al. (Laurier et al., 2008): a song is validated into a specific quadrant if at least one of the annotators agreed with AllMusic's annotation (Last.FM in their case). This resulted into a dataset with 771 lyrics (211 for Q1, 205 for Q2, 205 for Q3, 150 for Q4).

We extracted all features from the different categories described in Chapter 3.

In this study, for classification, we use SVM, since, based on previous evaluations, this technique performed generally better than other methods. A polynomial kernel was employed and a grid parameter search was performed to tune the parameters of the algorithm. Feature selection and ranking with the ReliefF algorithm were also performed in each feature set, in order to reduce the number of features. Results were validated with repeated stratified 10-fold cross validation (with 20

repetitions) and the average obtained performance is reported.

Using this process we have trained our large dataset (771 lyrics dataset) and we decided to join together the two datasets and therefore to train a still larger dataset with 951 lyrics (180+771). The best models after feature selection contained, respectively, for the two scenarios, 150 and 120 selected features.

We then applied the previous training sets to our test sets of sentences. First we applied to all the sentences available which are 368 sentences (129+239) and then to have a direct comparison with our KBA, we applied to the 239-sentences dataset.

We can observe in the following table the results of the application of the training sets to our testing sets.

Train Set #lyrics / #feats	**Test Set (#sentences)**	**Precision**	**Recall**	**F-Measure**
771 / 150	368	71.3	56.3	55.4
(771 + 180) / 120	368	74.1	57.3	57.2
771 / 150	239	69.8	51.5	51
(771 + 180) / 120	**239**	**73.7**	**52.3**	**52.7**

Table 4.13. Supervised ML approach: best training-testing scenarios.

For the same situation, the results of F-Measure are slightly better when the training set is bigger (e.g., for the same testing set, the results are always better when the training set is the 951-lyrics dataset). On the other hand, the results are also always better when the testing set is the 368-lyrics dataset).

The best result that we achieved, for the 239-sentences dataset, was an F-Measure of 52.7%. Comparing this result to our KBA, which achieved 67.4% (F-Measure), we conclude that the results are much better in our KBA. This suggests the importance of this approach for classification of sentences or smaller pieces of text.

Chapter 5

BIMODAL ANALYSIS

A right balance between music and lyrics is important. Music complements lyrics

Kailash Kher

This chapter presents our bimodal analysis (based on audio and lyrics) and includes also the process of creation of the audio dataset (DT1-A) corresponding to the lyrics dataset created in Lyrics Classification and Regression (DT1-L).

The chapter is structured as described in the following paragraphs.

Section 5.1 Audio Dataset Construction (DT1-A)

This section presents the process of data collection, annotation and validation of the audio dataset (DT1-A).

Section 5.2 Bimodal Dataset Construction

Here we explain the process of construction of the bimodal dataset, namely, how we join DT1-L and DT1-A.

Section 5.3 Feature Extraction

We present in this section the process of feature extraction (lyrics and audio) for the bimodal dataset.

Section 5.4 Results and Discussion

In this section, we present the results achieved for the three types of experiments performed, concerning bimodal analysis: classification by quadrant categories, by arousal hemispheres and by valence meridians.

Section 5.5 Other Experiments

We present in this section other bimodal experiments made at the beginning of this research. These experiments were based on a dataset (DT3) from AllMusic and in that time we used only features from the state of the art.

5.1 Audio Dataset Construction (DT1-A)

To accomplish the goal of making bimodal analysis, we asked the annotators to classify audio samples following the same principles they used to annotate lyrics (Section 3.1 Lyrics-Dataset Construction (DT1-L)).

Each annotator classified only one of the dimensions (audio or lyrics) for each song, never both simultaneously.

5.1.1 Data Collection

We start from the same 200 songs, whose selection criteria were described in (Section 3.1.1 Data Collection).

Next, for each song, we used the AllMusic API to search for audio clips of 30 seconds provided by the platform. When the song was not present in AllMusic, we collected the song manually and then we extracted the 30-sec clip: a representative part of the song normally including the chorus. All the clips were converted to mp3 with a sampling rate of 22050 Hz.

5.1.2 Annotation and Validation

The annotation of the dataset was performed by 39 people with different backgrounds. During the process, we recommended the following annotation methodology:

1. Hear the audio clip (try to ignore the meaning of the lyric);
2. Identify the basic predominant emotion expressed by the audio (if the user thought that there was more than one emotion, he/she should pick the predominant);
3. Assign values (between -4 and 4) to valence and arousal;
4. Fine tune the values assigned in 3) through ranking of the samples.

To further improve the quality of the annotations, the users were also recommended not to search for information about the audio neither the song on the Internet or another place and to avoid tiredness by taking a break and continuing later.

We obtained an average of 6 annotations per audio clip. Then, the arousal and valence of each song were obtained by the average of the annotations of all the subjects.

To improve the consistency of the ground truth, the standard deviation (SD) of the annotations made by different subjects for the same song was evaluated. Songs with an SD above 1.2 were excluded from the original set. As a result, 38 songs were discarded, leading to a final dataset containing 162 audio clips. This leads to a 95% confidence interval (Montgomery et al., 1998) of about ±0.5. We believe this is acceptable in our -4.0 to 4.0 annotation range. We can see in Figure 5.1 the distribution of the standard deviations in the validated songs.

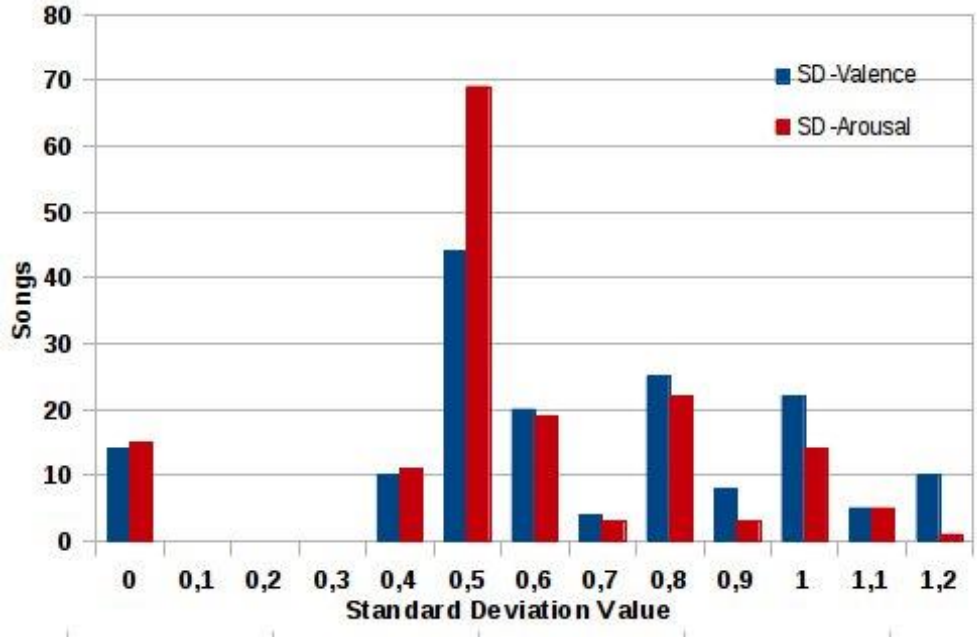

Figure 5.1. Audio: Distribution of the Standard Deviations in the Validated Songs.

Finally the consistency of the ground truth was evaluated using Krippendorff's alpha (Krippendorff, 2004), a measure of inter-coder agreement. This measure achieved, in the range -4 up to 4, 0.69 and 0.72 respectively for the dimensions valence and arousal. This is considered a substantial agreement among the annotators (Landis and Koch, 1977).

As with as the lyrics dataset, the size of the audio dataset is not too large, however we think it is acceptable for experiments and is similar to other manually annotated datasets (e.g., (Yang et al., 2008) has 195 songs).

The following two figures (Figure 5.2 and Figure 5.3) show the histogram for arousal and valence dimensions as well as the distribution of the 162 selected songs for the 4 quadrants.

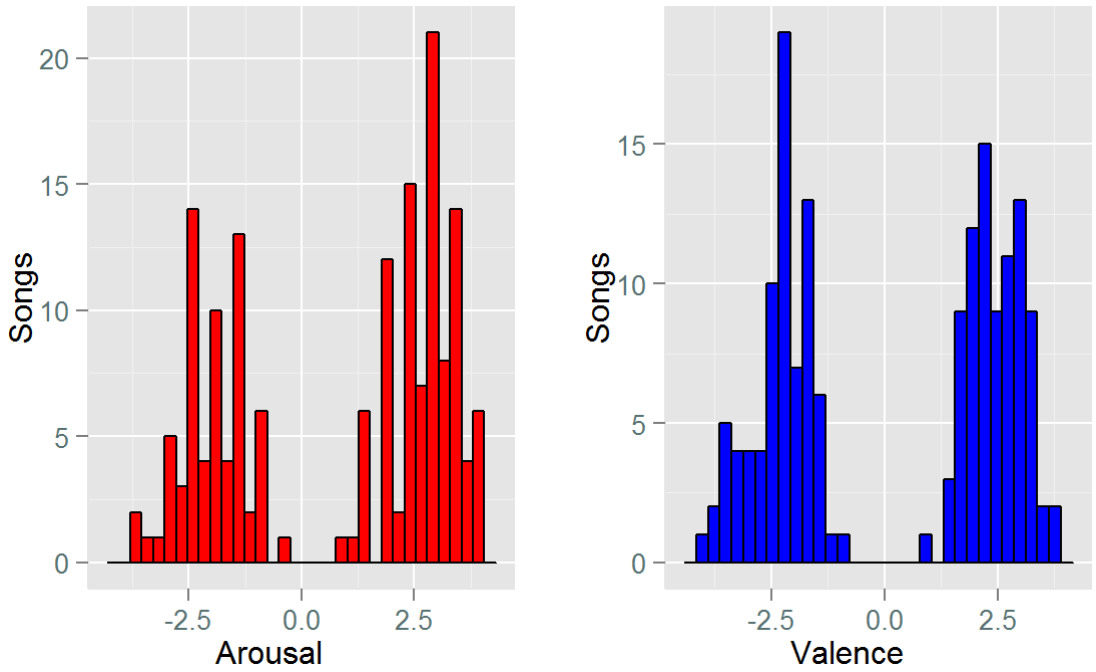

Figure 5.2. Audio: Arousal and valence histogram values.

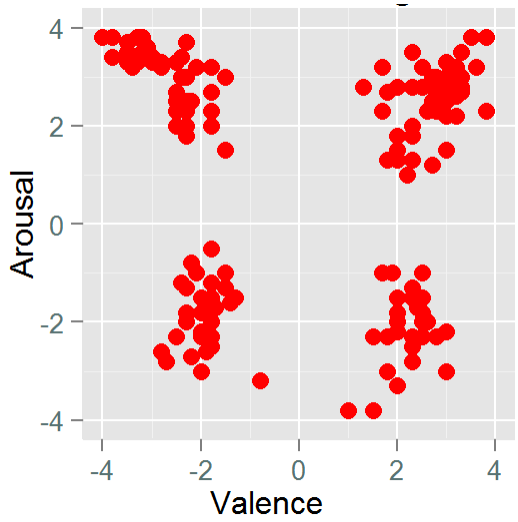

Figure 5.3. Audio: Distribution of the songs for the 4 quadrants.

Finally, each song is labeled as belonging to one of the four possible quadrants (Table 5.1),

Quadrant	Number of Songs
1	52
2	45
3	31
4	34

Table 5.1. Audio: Number of Songs per Quadrant.

5.2 Bimodal Dataset Construction

From the datasets constructed for the lyrics (Section 3.1 Lyrics-Dataset Construction (DT1-L)) and for the audio (Section 5.1 Audio Dataset Construction (DT1-A)), we created a bimodal dataset.

We consider that a song (audio + lyrics) is a valid song to integrate this bimodal dataset, if the song belongs simultaneously to the audio and lyrics dataset and in both datasets the sample belongs to the same quadrant, i.e., we can only consider songs in which the classification (quadrant) for the audio sample is equal to the classification for the lyric sample.

Quadrant	Number of Songs
1	37
2	37
3	30
4	29

Table 5.2. Bimodal Dataset: Number of Songs per Quadrant.

So we start from a dataset of lyrics containing 180 samples and a dataset of audio containing

162 samples, obtaining a bimodal dataset that contains 133 songs. Table 5.2 shows their distributing across the 4 quadrants of the Russell's model.

Table 5.3 and Table 5.4 show respectively the way they are distributed for the 2 hemispheres and parallels.

Hemispheres	Number of Songs
North (AP)	74
South (AN)	59

Table 5.3. Bimodal Dataset: Number of Songs per Hemisphere.

Parallels	Number of Songs
East (VP)	66
West (VN)	67

Table 5.4. Bimodal Dataset: Number of Songs per Parallel.

5.3 Feature Extraction

5.3.1 Audio Feature Extraction

In musical theory, the basic musical concepts and characteristics are commonly grouped under broader distinct elements such as rhythm, melody, timbre and others. In this work, we organize the available audio features under these same elements. A total of 1701 features (Table 5.5) were extracted using known state of the art audio frameworks. This part of the work is described in (Malheiro et al., 2016).

Categories	#Features
Dynamics	196
Expressive Techniques	27
Harmony	245
Melody	120
Musical Form	14
Rhythm	70
Tone Color	1029
Total of Features	**1701**

Table 5.5. Number of Features per Audio Category.

We can see a short definition of each one of the categories:

- **Dynamics.** All musical aspects relating to the relative loudness (or quietness) of music. Important aspects include the relative softness and loudness of sound, change of loudness (contrast), and the emphasis on individual sounds (accent). Some audio features related with dynamics are average silence ratio, loudness, low energy rate or root-mean-squared energy;

- **Expressive Techniques.** Used to create the musical detail that articulates a style or interpretation of a style and refers to the way a performer plays a piece of music. Features related with expressive techniques are vibrato rate, vibrato extent and coverage;

- **Harmony.** Related to the verticalization of pitch. It can be seen has the combination of pitches into chords (several notes played simultaneous), or it may be produced by

two or more sources playing together. Some features that try to capture harmony are inharmonicity, key and key clarity, and modality estimation;

- **Melody.** Defined as a horizontal succession of pitches. Some of the main components of melody are pitch (definite or indefinite), range, register and melodic contour, movement and arrangement. Features such as pitch estimation, salience, range and shape class descriptors are used to capture melody information;

- **Musical Form.** The term musical form (or musical structure) refers to the overall structure of a piece of music, and describes the layout of a composition as divided into sections. These sections are usually identified by changes in rhythm and texture, such as "verse" and "chorus", the foundation of popular music. Few of the used features are related with musical form. Some of those are similarity matrix and novelty curve;

- **Rhythm.** The element of "time" in music, the patterns of long and short sounds and silences found in music. The composer creates rhythm through patterns of long and short sounds and silences in the music. Some rhythm related features are events density, tempo estimation and pulse / rhythmic clarity;

- **Tone Color.** Also known as timbre, refers to the properties of sound that allows listeners to identify the sound source, as well as combination of sounds. It is influenced by three key factors: the material of the instrument or voice; the techniques employed in producing the sound; and the layers of sound and the effects the sound has on the music. Most of the extracted features are related with timbre. Some examples are the spectral moments (centroid, skewness, kurtosis), as well as MFCCs.

There are several frameworks to extract features from the audio. In this work we use the following frameworks (Table 5.6).

Frameworks	#Features
MIR Toolbox 1.6.1 (Lartillot and Toiviainen, 2007)	370
Marsyas (Tzanetakis, 2007)	778
PsySound 3 (Cabrera et al., 2008)	455
Melodic Audio Features (Salamon et al., 2012), (Rocha, 2011)	98
Total	**1701**

Table 5.6. Frameworks used for Audio Feature Extraction.

5.3.2 Lyrics Feature Extraction

The lyric features used for bimodal analysis were all the features described in (Section 3.2. Feature Extraction), namely all the Stylistic-Based, Song Structure-Based and Semantic-Based Features. We have performed tests using Content-Based Features but the results were not better and yet the dimensionality was much bigger, so we decided not to include this kind of features. Hence, a total of 1232 lyrics features resulted.

5.4 Results and Discussion

We conduct three types of experiments concerning bimodal analysis: i) by quadrant (4 categories – Q1, Q2, Q3 and Q4); ii) by arousal hemispheres (2 categories – AP and AN); iii) by valence meridians (2 categories – VP and VN).

We use Support Vector Machines (SVM) (Boser et al., 1992) algorithm, since, based on previous evaluations, this technique performed generally better than other methods. That is why it is the reference in this kind of works (e.g., Hu e Laurier). The classification results were validated with repeated stratified 10-fold cross validation (Duda et al., 2000) (with 20 repetitions) and the average obtained performance was reported.

For each experiment, we constructed first, both for audio and lyric dimensions, the best possible classifiers. We apply, for each one of the dimensions, feature selection and ranking using the ReliefF algorithm (Robnik-Šikonja and Kononenko, 2003) in order to reduce the number of features. Next we combine the best features of audio and lyrics and construct, using the same prior terms, the best bimodal classifier.

We will show in the next sections the results of the experiments.

5.4.1 Bimodal Analysis for Quadrants

We can see in Table 5.7 the performance of the best model for lyrics, audio and for the combination of the best lyric and audio features. The fields *#Features, Selected Features and F-measure(%)* represents respectively the total of features, the number of selected features and the results accomplished via the F-measure metric after feature selection. In the last line, the total number of bimodal features is the sum of selected lyrics and audio features.

Classification by Quadrants	#Features	Selected Features	F-measure (%)
Lyrics	1232	647	79.3
Audio	1701	418	72.6
Bimodal	1065	1057	88.4

Table 5.7. Classification by Quadrants: Performance (F-Measure) of the Classifiers.

As can be seen, the best lyrics-based model achieved better performance than the best audio-based model (79.3% vs 72.6%). This is not the more frequent pattern in the state of the art, where usually the best results are achieved with the audio. This happens for example in (Laurier et al., 2008). (Hu et al., 2009b) is the only research, as far as we know, where lyrics performance supplants audio performance, but only for some few moods or emotions. This suggests that our new lyric features (Section 3.2. Feature Extraction) have an important role for these results.

As we can see, both dimensions are important, since bimodal analysis improves significantly

($p<0.05$ Wilcoxon Test) the results of the lyrics classifier (from 79.3% to 88.4%). Furthermore, the best bimodal classifier, after feature selection, contains almost all the features from the best classifiers of lyrics and audio (1057 features in 1065 possible features). This suggests the importance of the features from both dimensions.

The following tables (Table 5.8 – Table 5.10) show, for the best bimodal, best lyrics and best audio model, the corresponding confusion matrices (in percentage) and the precision, recall and F-measure values for each quadrant, as well as the overall results (first line).

Q1	Q2	Q3	Q4	← classified as	Precision 88.5%	Recall 88.4%	F-measure 88.4%
24.5	3.1	0	0.2	Q1	89.3	88.1	88.7
2.3	24.5	0.1	0.9	Q2	88.7	88	88.3
0.1	0	20	2.5	Q3	91.3	88.8	90
0.6	0	1.8	19.4	Q4	84.3	89	86.6

Table 5.8. Quadrants – Best Bimodal Model: Confusion Matrix and statistic Measures.

Q1	Q2	Q3	Q4	← classified as	Precision 79.7%	Recall 79.2%	F-measure 79.3%
20.3	2.1	2.8	2.4	Q1	82.6	73.1	77.6
0.1	23.7	2.9	1.2	Q2	88.7	84.7	86.7
1.9	0.1	18.4	2.3	Q3	70.8	81.3	75.7
2.3	0.8	1.8	16.9	Q4	73.9	77.6	75.7

Table 5.9. Quadrants – Best Lyrics Model: Confusion Matrix and statistic Measures.

Q1	Q2	Q3	Q4	← classified as	Precision 72.8%	Recall 72.6%	F-measure 72.6%
20.7	6.4	0.2	0.6	Q1	74	74.3	74.2
6.4	20.4	0.2	0.8	Q2	75.9	73.4	74.6
0.1	0	15.4	7	Q3	74.4	68.7	71.4
0.7	0.1	5	16	Q4	65.7	73.4	69.4

Table 5.10. Quadrants – Best Audio Model: Confusion Matrix and statistic Measures.

The analysis of the previous tables allows us to conclude, as other authors concluded (Shaukat and Chen, 2008), (Vallverdu and Casacuberta, 2009), that audio is more important for arousal discrimination, while in valence discrimination the lyrics are more important. In fact, in Table 5.10, we can see that there are more audio clips incorrectly classified between quadrants of the same hemispheres (e.g., Q1 and Q2; Q3 and Q4) than between quadrants of the same meridians (e.g., Q1 and Q4; Q2 and Q3). We can observe that for example 7% of the audio songs from Q3 were incorrectly classified in Q4, 5% from Q4 were wrongly classified in Q3, 6.4% from Q1 were incorrectly classified in Q2 and 6.4% from Q2 were incorrectly classified in Q1.

We can point out also that in the case of the audio, the quadrants of the hemisphere north (Q1 and Q2) have better performance (F-Measure) than the quadrants 3 and 4.

For lyrics (Table 5.9), these kind of relations are not so obvious at least with the classification by quadrants.

5.4.2 Bimodal Analysis for Arousal

Table 5.11 shows the performance of the best models for lyrics, audio and for the combination of the best lyric and audio features.

Classification by Arousal Hemispheres	#Features	Selected Features	F-measure (%)
Lyrics	1232	94	88
Audio	1701	578	97.9
Bimodal	672	613	98

Table 5.11. Classification by Arousal Hemispheres: Performance (F-Measure) of the Classifiers.

We can point out the excellent behavior of the classifiers for the task of arousal discrimination. Bimodal analysis has a performance (F-measure) of 98%. This is due to the fact that the results achieved by audio are almost the same (97.9%). It is also worth to point out the high performance of the lyrics classifier, which achieves 88% with only 94 features.

The following tables (Table 5.12 – Table 5.14) show, for the best bimodal, best lyrics and best audio model for arousal discrimination, the corresponding confusion matrices and some statistics as we have explained before.

AN	AP	← classified as	Precision 98%	Recall 98%	F-measure 98%
43.4	1	AN	97.9	97.7	97.8
0.9	54.7	AP	98.2	98.3	98.2

Table 5.12. Arousal: best bimodal model.

The discrimination of arousal hemispheres in bimodal analysis is seen by the fact that only 1.9% of the songs were incorrectly classified: 1% was wrongly classified in the class AP and 0.9% was wrongly classified in the class AN.

AN	AP	← classified as	Precision 88%	Recall 88%	F-measure 88%
37.3	7.1	AN	88.4	84	86.1
4.9	50.7	AP	87.7	91.2	89.4

Table 5.13. Arousal: best lyrics model.

AN	AP	← classified as	Precision 97.9%	Recall 97.9%	F-measure 97.9%
43.1	1.2	AN	98	97.2	97.6
0.9	54.8	AP	97.8	98.4	98.1

Table 5.14. Arousal: best audio model.

The best audio model is very similar to the bimodal model. This shows the clear importance of the audio for the problem of arousal discrimination. Although the general performance of the lyrics model is high (88%, F-Measure), it is not as good as the audio performance.

5.4.3 Bimodal Analysis for Valence

Table 5.15 shows the performance of the best models for lyrics, audio and for the combination of the best lyric and audio features.

Classification by Valence Parallels	**#Features**	**Selected Features**	**F-measure (%)**
Lyrics	1232	413	87.3
Audio	1701	659	71.5
Bimodal	1072	30	90.8

Table 5.15. Classification by Valence Meridians: Performance (F-Measure) of the Classifiers.

The results observed in Table 5.15 confirm the importance of the lyrics for valence

discrimination (87.3% F-Measure) with 413 features. Still more interesting is the fact that bimodal analysis improves performance to 90.8% with only 30 features. Analyzing these features, we observe that the first 10 are all lyric features, most of them based in valence value. For example, the first 5 features, by this order, are *VinANEW*, *VinGAZQ1Q2Q3Q4*, *negemo*, *Sadness_Weight_Synesketch*, *Anger_Weight_Synesketch*. In the 30 features, 7 are novel semantic features proposed by us and in total we have 18 features from lyrics and 12 from audio.

The following tables (Table 5.16 – Table 5.18) show some statistics about the best bimodal, best lyrics and best audio model for valence discrimination.

VP	VN	← classified as	Precision 90.9%	Recall 90.8%	F-measure 90.8%
46	3.7	VP	89.3	92.7	90.9
5.5	44.8	VN	92.5	89	90.7

Table 5.16. Valence: best bimodal model.

VP	VN	← classified as	Precision 87.3%	Recall 87.3%	F-measure 87.3%
43.3	6.4	VP	87.3	87.2	87.2
6.3	44	VN	87.4	87.5	87.4

Table 5.17. Valence: best lyrics model.

VP	VN	← classified as	Precision 71.5	Recall 71.5	F-measure 71.5
36.2	13.5	VP	70.6	72.8	71.7
15	35.3	VN	72.4	70.1	71.2

Table 5.18. Valence: best audio model.

Through the observation of the previous table, we conclude that lyrics have an important role

in valence discrimination and when we combine lyrics and audio, the results are always better. This confirms the general idea about the importance of both dimensions for the analysis of songs in a real scenario.

5.5 Other Experiments

Before taking the decision of mannually creating the datasets for lyrics (DT1-L) and audio (DT1-A), we made some experiments using the same methods as other state of the art authors. We collected a dataset from AllMusic and used only state of the art features. The results of these experiments are published in (Malheiro et al., 2013) and (Panda et al., 2013). The following is a detailed explanation of these experiments.

We started from a dataset of 764 samples (audio+lyrics) and performed feature extraction using several natural language processing techniques. Our goal was to build classifiers for the different featuresets, comparing different algorithms and using feature selection. The best results (44.2% F-measure) were attained with SVMs. We also performed a bimodal analysis that combined the best feature sets of audio and lyrics.The combination of the best audio and lyrics features achieved better results than the best feature set from audio only (63.9% F-Measure against 62.4% F-Measure).

5.5.1 Dataset from AllMusicGuide (DT3)

We created a bimodal dataset, based on the AllMusic knowledge base and organized in a similar way as the MIREX taxonomy. It contains five clusters with several emotional categories each: cluster 1: passionate, rousing, confident, boisterous, rowdy; cluster 2: rollicking, cheerful, fun, sweet, amiable/good natured; cluster 3: literate, poignant, wistful, bittersweet, autumnal, brooding; cluster 4: humorous, silly, campy, quirky, whimsical, witty, wry; cluster 5: aggressive, fiery, tense/anxious, intense, volatile, visceral (Table 5.19).

Cluster 1	Passionate, Rousing, Confident, Boisterous, Rowdy
Cluster 2	Rollicking, Cheerful, Fun, Sweet, Amiable/Good Natured
Cluster 3	Literate, Poignant, Wistful, Bittersweet, Autumnal, Brooding
Cluster 4	Humorous, Silly, Campy, Quirky, Whimsical, Witty, Wry
Cluster 5	Aggressive, Fiery, Tense/anxious, Intense, Volatile, Visceral

Table 5.19. MIREX Mood Dataset: The five clusters and respective subcategories.

The first step consisted in accessing automatically the AllMusic API to obtain a list of songs with the MIREX emotion tags and other meta-information, such as song identifier, artists and title. To this end, a script was created to fetch existing audio samples from the same site, mostly being 30-second mp3 files.

The next step was to create the emotion annotations. To do so, the songs containing the same emotion tags present in the MIREX clusters were selected. Since each song may have more than one tag, the tags of each song were grouped by cluster and the resulting song annotation was based in the most significant cluster, i.e., the one with more tags (for instance, a song with one tag from cluster 1 and three tags from cluster 5 is marked as cluster 5). A total of 903 MIREX-like audio clips, nearly balanced across clusters, were acquired.

We used a dataset of 903 audio excerpts organized into five clusters, similarly to the MIREX campaign. This dataset and user annotated clusters were gathered from the AllMusic database. Next, we developed tools to automatically search for lyrics files of the same songs using the Google API. In this process, three sites were used for lyrical information (lyrics.com, ChartLyrics and MaxiLyrics). After removal of some deficient files, the interception of the 903 original audio clips with the lyrics resulted in a dataset containing 764 lyrics and audio excerpts (Table 5.20).

Cluster	Number of Songs
1	135
2	138
3	192
4	173
5	126

Table 5.20. Songs distribution across clusters.

5.5.2 Feature Extraction

We have used 2 types of features: features based on existing frameworks like Jlyrics[34], Synesketch[35] and ConceptNet [36] (FF) and BOW features. We considered BOW features with several transformations: stemming, stopwords removal, with none or with both of the previous operations.

For each operation, we compared two types of representations for the features: Boolean and TFIDF. For each one of the previous combinations, we calculate unigrams, bigrams and trigrams, creating a total of 24 feature sets (Figure 5.4).

[34] http://jmir.sourceforge.net/jLyrics.html

[35] http://synesketch.krcadinac.com/blog/

[36] http://web.media.mit.edu/~hugo/conceptnet/

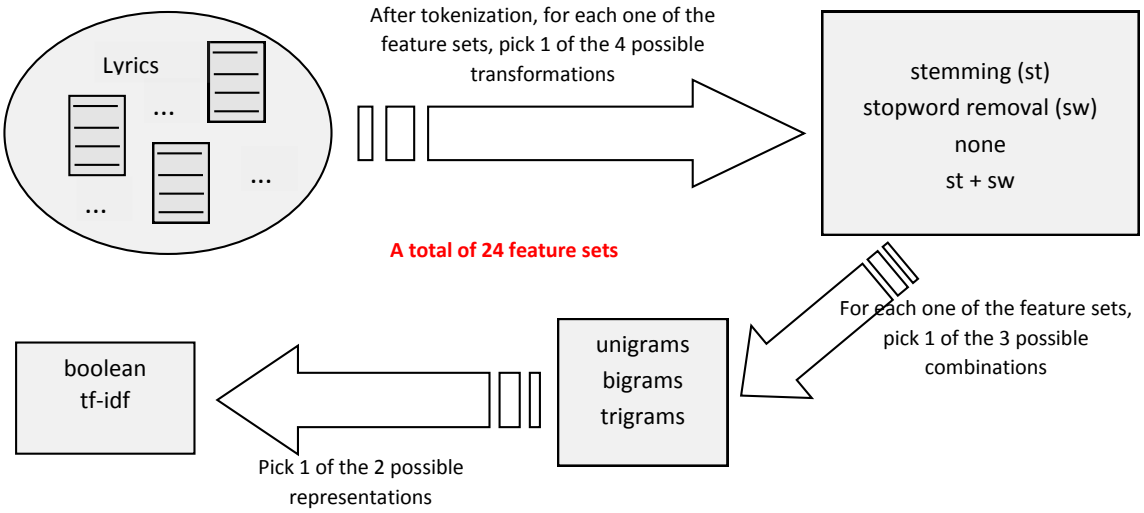

Figure 5.4. Process of Feature Sets Construction.

The best feature sets with unigrams, bigrams and trigrams are combined as follows: unigrams+bigrams (combination of unigrams and bigrams) (UB) and unigrams+bigrams+trigrams (UBT). We have also evaluated UB and UBT combined to the best features extracted from FF. At the end, we evaluated the feature sets UB+FF+Audio and UBT+FF+Audio, where (Audio is the best set of audio features, as reported in (Rocha et al., 2013)). Various tests were run with the following supervised learning algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), C4.5 and Naïve Bayes (NB). In addition to classification, feature selection and ranking with the ReliefF algorithm (Robnik-Šikonja and Kononenko, 2003) were also performed in order to reduce the number of features and improve the results. For both feature selection and classification, results were validated with repeated stratified 10-fold cross validation (with 20 repetitions), reporting the average obtained accuracy.

5.5.3 Experimental Results

Several experiments were performed to evaluate the importance of the various subsets of features and the effect of their combination in emotion classification. In these experiments we performed feature selection to identify the best features in each dataset. In Table 5.21, we present the best results

achieved for the evaluated classifiers in each feature set: UB, UBT, FF and Audio.

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

Table 5.21. F-Measure results for Classification Task.

The best results were always reached with SVM classifiers. Concerning to lyrical features, content-based features (BOW) achieved better results than FF features (predominantely based on the structure of the lyric). These results reinforce the importance of content-based features, as we can see in other studies like (Hu, 2010). The results in datasets containing unigrams, bigrams and trigrams are always better than the ones attained in datasets with unigrams and bigrams.

The results achieved with the combination of features from audio and lyrics are slightly better than the reference (audio). These results support our initial hypothesis that the combination of audio+lyric features helps to improve the performance attained by each one of them separately. The best results (63.9% F-Measure) were obtained in a feature set of 12 features (after feature selection) (11 from audio and 1 from lyrics). This feature from lyrics is a unigram (the token achieved after stemming – *babi*). The next 3 more important features from lyrics were also unigrams: *gonna, love, night*. We can see the description of the best 11 features from audio in (Rocha et al., 2013).

Chapter 6

CONCLUSIONS AND PERSPECTIVES

Singing is a way of releasing an emotion that you sometimes can't portray when you're acting. And music moves your soul, so music is the source of the most intense emotions you can feel

Amanda Seyfried

At the beginning of our research work, most of the studies in MER were more based on audio content analysis of music. Even if some of them have also performed bimodal analysis (audio + lyrics), they extracted normally the lyrics state of the art features, namely the BOW features. We can see this in works such as (Laurier et al., 2008) and (Yang et al., 2008). As far as we know (Hu, 2010) is the only researcher that went a step further and included also stylistic features.

Our goal was to make a deep study about the importance of the lyric features on the performance of the MER system. However, this required a manually annotated dataset from the lyrics, because we are not aware of any public dataset (we contacted a few authors, but did not receive any positive answers) using the same emotion model and, for platforms like AllMusic, details on how they annotate the songs are not explicit. Hence, we decided to create it, and during our work we have proposed some novel features such as new stylistic, semantic and structural-based features. To further

validate these experiments, we built a validation set comprising 771 lyrics extracted from the AllMusic platform, and validated by three volunteers.

We considered three different classification problems which were: classification in quadrants (4 classes or 4 sets of emotions); classification by arousal hemispheres (2 classes); and classification by valence parallels (2 classes). For these problems, we studied the importance (for the description and discrimination of classes) of the features when they acted together in models and when they acted alone. This was made not only for our new proposed features, but also for the other features we have used, such as features from platforms like LIWC and GI. Compared to the baseline features (e.g., BOW), the novel features have significantly improved the classification results.

To have a deeper understanding of the emotions conveyed by the lyric, we complemented the previous black-box systems with rule-based systems. We conducted experiments to understand the relations between features and emotions (quadrants), not only for our new proposed features, but also for all the other features from the state of the art that we have used, namely CBF and features from known frameworks such as LIWC, GI, Synesketch and ConceptNet. This analysis have shown good results for some of the novel features in specific situations, such as StyBF (e.g. #Slang and FCL), StruBF (e.g. #Title), and SemBF in general. To the best of our knowledge, this feature analysis was absent from the state of the art and so this is also a relevant contribution. To understand how this relation works, we have identified interpretable rules that show the relation between features and emotions and the relations among features.

As our dataset was dimensional, we performed experiments with regression models and the results, in comparison with similar studies for audio, were much better for the dimension valence and very close from audio for the dimension arousal.

To understand the importance of the lyrics in a real scenario, we performed bimodal analysis. For that, we created a manually annotated audio dataset (the annotators did not annotate simultaneously audio and lyrics for the same song) and we considered only the songs with audio and lyrics annotated in the same quadrant. Unlike most state of the art studies, classification with only lyrics achieved better results than classification with only audio. This can be possibly explained by the fact that in classification by lyrics we have used all the features we considered before, while in audio we considered most of the state of the art features, but not new features. These results, together

with the results achieved for regression, confirms the importance of the lyrics in the MER process.

Most of the studies referenced in the state of the art assign a global emotion to each song, but we know that the lyric is composed of several segments (e.g., title, chorus), which might convey different emotions. From this idea we created SERM (Sentence Emotion Recognition Model) to detect emotions in sentences/verses.

In short, some of our main contributions were:

- Ground truth dataset, manually annotated through the audio and the lyrics;
- Validation dataset of 771 songs annotated through AllMusic
- Ground truth dataset of manually annotated sentences;
- New features and/or features adapted from other domains;
- A set of rules that relate lyric features and emotions and features each other.

In the future, we will continue with the proposal of new features, particularly at the stylistic and semantic level. We will use the knowledge acquired from the relations between features and emotions and from the relations among features to propose new more precise features.

To improve our emotion gazetteers we will extend the current ones through the dictionary from Warriner (Warriner et al., 2013) which contains 13915 English words annotated with the dimensions arousal, valence and dominance.

Since our goal is to build a system to classify sentences in one of the four possible quadrants, we have ignored in our work the sentences annotated as neutral sentences. In the future we intend to expand our model to detect previously if a sentence is emotional or non-emotional.

In the study of emotion variation detection along the lyric, we want to understand the importance of the different structures (e.g. chorus) along the lyric, to know for example if the emotions conveyed by the chorus are in general the same as the emotions conveyed by the whole lyric. Additionally, we intend to make music emotion variation detection in a bimodal scenario,

including audio and lyrics. This implies an audio-lyrics alignment.

Finally, we aim to address the problem of lyrics transcription from the singing voice. This a very challenging task, that is so far in a very embryonic stage (Mesaros, 2013).

REFERENCES

Abbasi, A., Chen, H., Thoms, S. and Fu, T. (2008). “Affect analysis of Web forums and Blogs using correlation ensembles”. IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, pp. 1168-1180.

Abbasi, A., Chen, H. and Salem, A. (2008). “Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums”. ACM Transactions on Information Systems, Vol. 26, No. 3, pp. 12:1-12:34.

Abeles, H. and Chung, J. (1996). “Responses to Music”. pp. 285–342. IMR Press, San Antonio, TX.

Agrawal, A. and An, A. (2012). “Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations”. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, pp 346-353.

Agrawal, R., Imieliński, T. and Swami, A. (1993). “Mining association rules between sets of items in large databases”. ACM SIGMOD Record, vol. 22, pp. 207–216.

Airoldi, E., Bai, X. and Padman, R. (2006). “Markov blankets and meta-heuristic search: Sentiment extraction from unstructured text”. Lecture Notes in Computer Science, 3932 (Advances in Web Mining and Web Usage Analysis), pp. 167–187.

Alm, C., Roth, D. and Sproat, R. (2005). “Emotions from text: machine learning for text-based emotion prediction”. In: Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 579-586.

Altman, N. (1992). “An introduction to kernel and nearest-neighbor nonparametric regression”. The American Statistician, Vol. 46, No. 3, pp. 175–185.

Aman, S. (2007). “Recognizing Emotions in Text”. MSc Thesis, University of Ottawa, Canada.

Aman, S. and Szpakowicz, S. (2007). “Identifying Expressions of Emotion in Text”. In V. Matousek, P. Mautner (eds): In 10th International Conference on Text, Speech and Dialogue TSD 2007, Lecture Notes in Computer Science 4629, Springer, pp. 196-205.

Argamon, S., Saric, M. and Stein, S. (2003). “Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results”. In: ACM International Conference on Knowledge Discovery and Data Mining, pp. 475-480.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N. and Levitan, S. (2007). “Stylistic text classification using functional lexical features”. Journal of the American Society for Information Science and Technology, Vol. 58, No. 6, pp. 802-822.

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

Binali, H., Wu, C. and Potdar, V. (2010). “Computational Approaches for Emotion Detection in Text”. In: 4th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2010), pp. 172-177.

Boser, B., Guyon, I. and Vapnik, V. (1992). “A training algorithm for optimal margin classifiers”. In: 5th Annual Workshop on Computational Learning Theory, pp. 144–152.

Bradley, M. and 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.

Brandt, A. (2016). “Musical Form”. Retrieved from http://cnx.org/content/ m11629/1.13/ (Accessed in March, 2016).

Cabrera, D., Ferguson, S. and Schubert, E. (2008). “PsySound3: an integrated environment for the analysis of sound recordings”. Acoustics 2008: In: Australian Acoustical Society Conference, Geelong, Australia.

Chaffar, S. and Inkpen, D. (2011). “Using a Heterogeneous Dataset for Emotion Analysis in Text”. Advances in Artificial Intelligence, pp. 62-67.

Chopade, C. (2015). “Text based Emotion Recognition”. International Journal of Science and Research (IJSR), Vol. 4, No. 6, pp. 409-414.

Chuang, Z-J. and Wu, C-H. (2004). “Multi-Modal Emotion Recognition from Speech and Text”. Vol. 9, No. 2, pp. 45-62.

Chunling, M., Prendinger, H. and 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.

Cohen, W. and Hirsh, H. (1998). “Joins that generalize: text classification using WHIRL”. In: 4th International Conference on Knowledge Discovery and Data Mining (New York, NY, 1998), pp. 169–173.

Cooke, D. (1959). “The Language of Music”. London Oxford University Press.

Das, S. and Chen, M. (2001). “Yahoo! for Amazon: Extracting market sentiment from stock message boards”. In: Asia Pacific Finance Association Annual Conference (APFA).

Das, S. and Chen, M. (2007). “Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web”, Management Science, Vol. 53, No. 9, pp. 1375–1388.

Dave, K., Lawrence, S. and Pennock, D. (2003). “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews”. In: 12th International Conference on World Wide Web, pp. 519–528.

del-Hoyo, R, Hupont, I., Lacueva, F., and Abadia, D. (2009). “Hybrid Text Affect Sensing System for Emotional Language Analysis”. In: International Workshop on Affective-Aware Virtual Agents and Social Robots, pp. 1-4.

Downie, J. (2008). “The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research”. Acoustical Science and Technology, Vol. 29, No. 4, pp. 247–255.

Duda, R., Hart, P. and Stork, D. (2000). “Pattern Recognition”. New York, John Wiley & Sons, Inc.

Ekman, P. (1982). “Emotion in the Human Face”. Cambridge University Press, 2nd Ed.

Ekman, P. (1992). “An argument for basic emotions”. Cognition and Emotion, Vol. 6, No. 3/4, pp. 169-200.

Ekman, P. (2003). “Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional Life”. Times Books.

Farnsworth, P. (1954). “A study of the Hevner adjective list. J. Aesthetics Art Criticism”. Vol. 13, pp. 97–103.

Fehr, B. and Russel, J. (1984). “Concept of Emotion viewed from a prototype perspective”. Journal of Experimental Psychology, Washington, pp. 464-486.

Feng, Y., Zhuang, Y. and Pan, Y. (2003). “Music Information Retrieval by Detecting Mood via Computational Media Aesthetics”. In: IEEE/WIC International Conference on Web Intelligence, pp. 235-241.

Feng, Y., Zhuang, Y. and Pan, Y. (2003). “Popular music retrieval by detecting mood”. In: International Conference on Information Retrieval, pp.375–376.

Folland, G. (1999). “Real Analysis. Modern Techniques and their Applications”. 2nd Ed. New York: John Wiley, 1999.

Fontaine, J., Scherer, K. and Soriano, C. (2013). “Components of Emotional Meaning”. A Sourcebook. Oxford University Press.

Gabrielsson, A. and Juslin, P. (1996). “Emotional expression in music performance: Between the performer's intention and the listener's experience”. Psychology of Music, Vol. 24, pp. 68-91.Gabrielsson, A. and Lindstrom, E. (2001). “The influence of musical structure on emotional expression“. In: Music and Emotion: Theory and Research, P. N. Juslin and J. A. Sloboda Eds., Oxford University Press, Oxford, UK.

Gabrielsson, A. (2002). “Emotion Perceived and Emotion Felt: Same or Different?”. Musicae Scientiae (special issue), European Society for the Cognitive Sciences of Music (ESCOM), pp.

123-147.

Gamon, M. (2004). “Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis”. In: 20th International Conference on Computational Linguistics (COLING).

Hancock, J., Landrigan, C. and Silver, C. (2007). “Expressing emotions in text-based communication”. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 929-932.

Henderson, I. (1957). "Ancient Greek Music". In The New Oxford History of Music, vol.1: Ancient and Oriental Music, edited by Egon Wellesz, pp. 336–403. Oxford: Oxford University Press.

Hevner, K. (1936). “Experimental studies of the elements of expression in music”. American Journal of Psychology, Vol. 48, pp. 246-268.

Hirat, R. and Mittal, N. (2015). “A Survey on Emotion Detection Techniques using Text in Blogposts”. International Bulletin of Mathematical Research, Vol. 2, No. 1, pp. 180-187.

Holbrook, M. and Schindler, R. (1989). “Some exploratory findings on the development of musical tastes”. J. Consumer Research, Vol. 16, pp. 119–124.

Hu, X. (2010). “Improving Music Mood Classification using Lyrics, Audio and Social Tags”. PhD Thesis, University of Illinois at Urbana-Champaign.

Hu, X. and Downie, J. (2007). “Exploring mood metadata: Relationships with genre, artist and usage metadata”. In: 8th International Conference on Music Information Retrieval (ISMIR’07).

Hu, X. and Downie, J. (2010). “When Lyrics Outperform Audio For Music Mood Classification: A Feature Analysis”. 11th International Society for Music Information Retrieval Conference (ISMIR 2010).

Hu, X. and Downie, J. (2010). “Improving mood classification in music digital libraries by combining lyrics and audio”. In: 10th annual joint conference on Digital libraries, pp. 159-168.

Hu, X., Downie, J., Laurier, C., Bay, M. and Ehmann, A. (2008). “The 2007 MIREX Audio Music

Classification task: lessons learned". In: 9th International Conference on Music Information Retrieval (ISMIR'08), pp. 462-467.

Hu, Y., Chen, X. and Yang, D. (2009). "Lyric-Based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method" 10th International Society for Music Information Retrieval Conference, pp. 123-128.

Hu, X., Downie, S. and Ehman, A. (2009). "Lyric text mining in music mood classification". In: 10th International Society for Music Information Retrieval Conference (ISMIR), pp. 411–416.

Huron, D. (2000). "Perceptual and cognitive applications in music information retrieval". D. Byrd & J. S. Downie, Eds. Cognition, Vol. 10, No. 1, pp. 83–92. University of Massachusetts at Amherst.

Jargreaves, D. and North, A. (1997). "The Social Psychology of Music". Oxford University Press, Oxford, UK.

Juslin, P. (2013). "From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions". Phys Life Rev, Vol. 10, No. 3, pp. 235-266.

Juslin, P. and Laukka, P. (2004). "Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening". Journal of New Music Research, Vol. 33, No. 3, pp. 217–238.

Juslin, P. and Sloboda, J. (2001). "Music and Emotion: theory and research". Oxford University Press.

Juslin, P., Karlsson, J., Lindstrom, E., Friberg, A. and Schoonderwaldt, E. (2006). "Play it again with feeling: Computer feedback in musical communication of emotions". Journal of Experimental Psychology: Applied, Vol. 12, No. 1, pp. 79-95.

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

Keerthi, S. and Lin, C. (2003). "Asymptotic behaviors of support vector machines with Gaussian

kernel". Neural Computation, Vol. 15, No 7, pp. 1667–1689.

Kelsey Museum of Archaeology. University of Michigan (2003). "Music in Ancient Egypt". [ONLINE] Available at: http://www.umich.edu/~kelseydb/ Exhibits/MIRE/Introduction/ AncientEgypt/AncientEgypt.html (Accessed 21 June 2016).

Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J. and Turnbull, D. (2010). "Music emotion recognition: A state of the art review". In: 11th International Society of Music Information Retrieval (ISMIR), pp. 255 -266.

Korhonen, M., Clausi, D. and Jernigan, M. (2006). "Modeling emotional content of music using system identification". IEEE Transactions Systems Man and Cybernetics, Vol. 36, No. 3, pp. 588-599.

Krippendorff, K. (2004). "Content Analysis: An Introduction to its Methodology". 2nd Ed., chapter 11. Sage, Thousand Oaks, CA.

Kutner, J. and Leigh, S. (2005). "1000 UK Number One Hits". Kindle Edition.

Landis, J. and Koch, G. (1977). "The measurement of observer agreement for categorical data". Biometrics, Vol. 33, pp. 159–174.

Lartillot, O. and Toiviainen, P. (2007). "A Matlab Toolbox for Musical Feature Extraction from Audio", In: 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux.

Laurier, C., Grivolla, J. and Herrera, P. (2008). "Multimodal music mood classification using audio and lyrics". In: International Conference on Machine Learning and Applications.

Laurier, C., Sordo, M., Serra, J. and Herrera, P. (2009). "Music mood representation from social tags". In: International Society for Music Information Conference, Kobe, Japan.

Laurier, C. (2011). "Automatic Classification of Musical Mood by Content-Based Analysis". PhD Thesis, University of Pompeu Fabra, Barcelona.

Li, T. and Ogihara, M. (2004). "Semi-Supervised Learning from Different Information Sources". Knowledge and Information Systems, Vol. 7, No. 3, pp. 289-309.

Li, H., Pang, N. and Guo, S. (2007). “Research on Textual Emotion Recognition Incorporating Personality Factor”. In: International Conference on Robotics and Biomimetics, Sanya, China.

Lu, C., Hong, J. and Lara, S. (2006). “Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques”. 3rd Taiwanese-French Conference on Information Technology (TFIT 2006).

Lu, L., Liu, D. and Zhang, H. (2006). “Automatic mood detection and tracking of music audio signals“. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 1, pp. 5-18.

Malheiro, R., Panda, R., Gomes, P. and Paiva, R. (2013). "Music Emotion Recognition from Lyrics: A Comparative Study", In: 6th International Workshop on Machine Learning and Music, Prague.

Malheiro, R., Panda, R. and Paiva, R. (2016). Bimodal Music Emotion Recognition: Features and Dataset”. Technical Report, University of Coimbra.

Mandel, M., Poliner, G. and Ellis, D. (2006). “Support Vector Machine Active Learning for Music Retrieval”. Multimedia Systems, Vol. 12, No. 1, pp. 3-13.

Martinazo, B. (2010). “Um Método de Identificação de Emoções em Textos Curtos para o Português do Brasil”. MSc Thesis. Pontifícia Universidade Católica do Paraná.

Matsumoto, S., Takamura, H. and Okumura, M. 2005. “Sentiment classification using word sub-sequences and dependency sub-trees”. In: PAKDD’05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining.

Mayer, R., Neumayer, R. and Rauber, A. (2008). “Rhyme and Style Features for Musical Genre Categorisation by Song Lyrics”. In: International Conference on Music Information Retrieval.

McKay, C. (2002). “Emotion and music: Inherent responses and the importance of empirical cross-cultural research”. Course Paper. McGill University.

Mesaros, A. (2013). “Singing voice identification and lyrics transcription for music information retrieval invited paper”. In: 7th Conference on Speech Technology and Human Computer

Dialogue (SpeD), pp. 1-10, Cluj-Napoca, Romania.

Meyer, L. (1956). “Emotion and Meaning in Music”. Chicago: University of Chicago Press.

Meyers, O. (2007). “A mood-based music classification and exploration system”. MSc thesis, Massachusetts Institute of Technology.

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

Mishne, G. (2005). “Experiments with mood classification in Blog posts”. In: 1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005.

Montgomery, D., Runger, G. and Hubele, N. (1998). “Engineering Statistics”. Wiley.

Mullen, T. and Collier, N. (2004). “Sentiment analysis using support vector machines with diverse information sources”. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 412–418, Poster paper.

Ng, V., Dasgupta, S. and Arifin, S. (2006). “Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews”. In: COLING/ACL 2006 Main Conference, pp. 611-618.

Panda, R., Malheiro, R., Rocha, B., Oliveira, A. and Paiva, R., (2013). "Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis". In 10th International Symposium on Computer Music Multidisciplinary Research (CMMR’2013), Marseille, France.

Pang, B. and Lee, L. (2008). “Opinion mining and sentiment analysis”. Foundations and Trends in Information Retrieval, Vol. 2, No. 1-2, pp. 1-135.

Pratt, C. (1950). “Music as the language of emotion”. The Library of Congress.

Read, J. (2004). “Recognising affect in text using Pointwise-Mutual Information”. MSc Thesis, University of Sussex.

Robnik-Šikonja, M. and Kononenko, I. (2003). “Theoretical and Empirical Analysis of ReliefF and

RreliefF". Machine Learning, Vol. 53, No. 1-2, pp. 23–69.

Rocha, B., Panda, R. and Paiva, R. (2013). " Music Emotion Recognition: The Importance of Melodic Features". In: 5th International Workshop on Machine Learning and Music, Prague, Czech Republic.

Rocha, B. (2011). "Genre classification based on predominant melodic pitch contours". MSc thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Russell, J. (1980). "A circumspect model of affect". Journal of Psychology and Social Psychology, Vol. 39, No. 6, pp. 1161-1178.

Russell, S. and Norvig, P. (2003). "Artificial Intelligence: A Modern Approach". 2nd Ed. Prentice Hall.

Salamon, J., Rocha, B. and Gómez, E. (2012). "Musical Genre Classification using Melody Features Extracted from Polyphonic Music Signals". In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan.

Scherer, K., and Zentner, M. (2001). "Emotional effects of music: production rules". In: P. N.]uslin and J. A Sloboda (eds), Music and emotion. Theory and research, pp. 361-392). New York: Oxford University Press.

Schimmack, U. and Reisenzein, R. (2002). "Experiencing activation: energetic arousal and tense arousal are not mixtures of valence and activation". Emotion (Washington, D.C.), Vol. 2, No. 4, pp. 412-7. American Psychological Association.

Schmidt, E., Turnbull, D. and Kim, Y. (2010). "Feature selection for content-based, time-varying musical emotion regression". In: ACM International Conference on Multimedia Information Retrieval, pp. 267-274.

Schubert, E. (1999). "Measurement and time series analysis of emotion in music". PhD dissertation, School of Music Education, University of New South Wales, Sydney, Australia.

Schubert, E. (2003). "Update of the Hevner adjective checklist". Perceptual Motor Skills, Vol. 96,

pp. 1117–1122.

Sebastiani, F. (2002). “Machine learning in automated text categorization”. ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47.

Sen, A. and Srivastava, M. (1990). “Regression Analysis: Theory, Methods, and Applications”. New York, Springer.

Seol, Y., Kim, D-J. and Kim, H-W. (2008). “Emotion Recognition from Text Using Knowledge-based ANN”. In: 23rd International Technical Conference on Circuits/Systems, Computers and Communications, pp. 1569-1572.

Shaheen, S., El-Hajj, W., Hajj, H. and Elbassuoni, S. (2014). “Emotion Recognition from Text Based on Automatically Generated Rules”. In: IEEE International Conference on Data Mining Workshop, pp. 383-392.

Shaukat, A. and Chen, K. (2008). “Towards automatic emotional state categorisation from speech signals.” In: Interspeech’08, Brisbane Australia, pp. 2771–2774.

Sloboda, J. (2001). “Psychological perspectives on music and emotion”. Chap. 4, pp. 71–105. Oxford: Oxford University Press.

Sloboda, J. and Juslin, P. (2001). “Psychological perspectives on music and emotion”. In: P. N. Juslin and Sloboda (eds), Music and emotion. Theory and research, pp. 71-104), New York: Oxford University Press.

Smola, A. and Schölkopf, B. (2004). “A tutorial on support vector regression”. Statistics and Computing.

Solomatine, D. and Shrestha, D. (2004). “AdaBoost.RT: A boosting algorithm for regression problems”. In: Proceedings. IEEE International Joint Conference Neural Networks, pp. 1163–1168.

Stone, P., Dunphy, D., Smith, M. and Ogilvie, D. (1966). “The general inquirer: A computer approach to content analysis”. Cambridge, MA: The MIT Press.

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

Strapparava, C. and Mihalcea, R. (2008). “Learning to identify emotions in text”. In: 2008 ACM symposium on Applied computing, New York, pp. 1556-1560).

Subasic, P. and Huettner, A. (2001). “Affect analysis of text using fuzzy semantic typing”. IEEE Transactions on Fuzzy Systems, Vol. 9, No. 4, pp. 483–496.

Tang, H., Tan, S. and Cheng, X. (2009). “A survey on sentiment detection of reviews”. Expert Systems with Applications: An International Journal, Vol. 36, No. 7, pp. 10760-10773.

Tao, J. and Tan, T. (2004). “Emotional Chinese Talking Head System”. In: 6th International Conference on Multimodal interfaces, New-York, pp. 273-280.

Taylor, A., Marcus, M. and Santorini, B. (2003). “The Penn Treebank: an overview”. Series Text, Speech and Language Technology, Chap 1, Vol. 20, pp. 5-22.

Tellegen, A., Watson, D. and Clark, L. (1999). “On the Dimensional and Hierarchical Structure of Affect”. Psychological Science, Vol. 10, No. 4, pp. 297-303. SAGE Publications.

Teng, Z., Ren, F. and Kuroiwa, S. (2006). “Recognition of emotion with SVMs”. D.S. Huang, K. Li and G.W. Irwin (Eds): International Conference on Intelligent Computing (ICIC 2006), Lecture Notes in Computer Science, Vol. 4114, pp. 701-710.

Thayer, R. (1989). “The Biopsychology of Mood and Arousal”. New York: Oxford University Press.

Turnbull, D., Liu, R., Arrington, L. and Anckriet, G. (2007). “A game-based approach for collecting semantic annotations of music”. In: International Conference on Music Information Retrieval.

Tzanetakis, G. (2007). “Marsyas a case study in implementing Music Information Retrieval Systems”. Intelligent Music Information Systems Tools and Methodologies. (Eds) Shen, Shepherd, Cui, Liu, Information Science Reference.

Vallverdu, J. and Casacuberta, D. (2009). “Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence”. IGI

Global, 1st Edition.

Vignoli, F. (2004). “Digital Music Interaction concepts: a user study“. In: 5th International Conference on Music Information Retrieval (ISMIR’04).

Warriner, A., Kuperman, V. and Brysbaert, M. (2013). “Norms of valence, arousal, and dominance for 13,915 English lemmas”. Behavior Research Methods, Vol. 45, pp. 1191-1207.

Whissell, C. (1989). “Dictionary of Affect in Language”. In: Plutchik and Kellerman (Eds.) Emotion: Theory, Research and Experience, Vol. 4, pp. 113‑131, Academic Press, NY.

Wilson, T., Wiebe, J. and Hwa, R. (2006). “Recognizing strong and weak opinion clauses”. Computational Intelligence, Vol. 22, No. 2, pp. 73-99.

Yang, Y., Liu, C. and Chen, H. (2006). “Music emotion classification: A fuzzy approach”. In: ACM International Conference on Multimedia, pp. 81-84.

Yang, Y., Lin, Y., Su, Y. and Chen, H. (2008). “A Regression Approach to Music Emotion Recognition”. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 2, pp. 448-457.

Yang, Y. and Chen, H. (2011). “Music Emotion Recognition”. Multimedia Computing, Communication and Intelligence Series, CRC Press, Taylor & Francis Group.

Yang Y. and Chen H. (2012). "Machine recognition of music emotion: a review“. In: ACM Transactions on Intelligent Systems and Technology (TIST). Vol. 3, No. 3.

Yang, D. and Lee, W. (2009). “Music emotion identification from lyrics”. In: IEEE International Symposium on Multimedia, pp. 624-629.

Yang, C., Lin, K. and Chen, H. (2007). “Emotion Classification Using Web Blog Corpora”. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 275-278.

Yu, B. (2008). “An Evaluation of Text Classification Methods for Literary Study”. Literary and Linguistic Computing, Vol. 23, No. 3, pp. 327-343.

Zaanen, M. and Kanters, P. (2010). “Automatic Mood Classification using tf*idf based on Lyrics”. In: J. Stephen Downie and Remco C. Veltkamp, editors, 11th International Society for Music Information and Retrieval Conference.