AUTOMATIC EXTRACTION OF THE SEMANTIC CONTENT OF MUSIC LYRICS

VIRGILE BOULANGER

Master ATIAM internship report

Supervisor : Dr. Jochen Steffens

Audio-Kommunikation Group - TU Berlin

August 2017

ABSTRACT

In this report, we study the analysis of song lyrics in the context of developing a recommendation system based on the emotions conveyed by music. We thus consider several techniques for extracting the semantic content of textual data by studying two approaches : lexicon-based approach which detect keywords that deals with particular emotion (eg. love, beautiful), and machine-learning approaches (both supervised and unsupervised) which can classify text in emotion category, or detect topics in a song lyric.

For that we consider two different emotional models : the categorical model in which emotions are segregated into categories (eg. anger, sadness, happiness) and the dimensional model in which emotions a represented with two coordinates : valence (pleasantness) and arousal (intensity). The final recommendation system use a multi-modal approach by adding semantic features to a set of standard low-level audio features.

Keywords : Lyric analysis, music emotion detection, natural language processing, semantics, machine learning.

Dans ce rapport, nous étudions l'analyse de paroles de chanson afin de développer un système de recommendation musicale basé sur les émotions transmises par la musique. Nous considérons donc plusieurs techniques pour extraire le contenu sémantique de données textuelles en étudiant deux principales approches : l'approche par dictionnaire d'une part, qui permet de détecter des mot-clés qui ont un lien particulier avec certaines émotions (amour, beauté...), et l'approche par apprentissage automatique d'autre part (supervisé et non supervisé) avec laquelle nous pouvons classer les textes selon différentes catégories d'émotion, et également détecter automatiquement les thèmes principaux des paroles.

Pour cela nous considererons deux modèles pour la représentation d'émotion : le modèle en catégories dans lequel les émotions sont regroupé en différents types (par exemple: colère, tristesse, joie), et le modèle dimensionnel qui permet de représenter les émotions suivant deux variables (valence et intensité). Le système final de recommendation utilise une approache multi modale en ajoutant des caractéristiques sémantiques à un ensemble de caractéristiques audio.

Mots-clés : Analyse de parole, detection d'émotion, traitement du language naturel, sémantique, apprentissage automatique.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Jochen Steffens for his kindness during this five months internship. Thank you for having me in the Audio-Kommunikation group.

Many thanks to Athanasios Lykartsis for his support concerning the scientific questions and the insights he gave me for this work. Thanks for organizing the successful MIR Meetups in Berlin that helped me to connect more with people in the MIR community in Berlin and abroad.

Thanks to all the people of Audio-Kommunikation Group of TU Berlin who have the chance to talk with. This laboratory gather a lot of talented and enthusiastic people.

Thanks to Manuel Anglada for his friendship and support.

CONTENTS

1 INTRODUCTION 11
1.1 The Audio-Kommunikation Group . . . 11
1.2 Context . . . 11
1.3 Objectives . . . 12
1.4 Organization of the report . . . 12
2 STATE OF ART 13
2.1 Music and emotion : overview . . . 13
2.1.1 The concept of emotion . . . 13
2.1.2 Emotion models . . . 14
2.1.3 Emotion in music . . . 16
2.2 Emotion detection from text . . . 17
2.2.1 Keyword spotting approach . . . 17
2.2.2 Machine learning approach . . . 19
2.3 Multimodal music classification . . . 21
3 API STUDY 23
3.1 Building the lyric dataset . . . 23
3.2 ground-truth sample and exploratory factor analysis . . . 23
3.3 Synesketch API . . . 25
3.3.1 Operation of the API . . . 25
3.3.2 Results . . . 25
3.4 IBM Watson Tone Analyzer . . . 26
3.4.1 Operation of the API . . . 26
3.4.2 Results . . . 27
3.5 Conclusion . . . 28
4 MACHINE LEARNING APPROACH 29
4.1 Purpose . . . 29
4.2 Emotion-annotated datasets . . . 29
4.2.1 Fairy tales dataset . . . 29
4.2.2 ISEAR dataset . . . 29
4.3 Extracted features & preprocessing . . . 30
4.3.1 n-grams . . . 30
4.3.2 Word2Vec vectors . . . 30
4.4 Experiments . . . 31
4.5 Topic detection with NMF . . . 33
4.5.1 Non-negative Matrix factorization . . . 33
4.5.2 Procedure . . . 34
4.6 Conclusion & outlook . . . 35
5 KEYWORD-SPOTTING APPROACH 37
5.1 Purpose . . . 37
5.2 JEmAS Framework . . . 37
5.3 Procedure . . . 37
5.4 Results & conclusion . . . 38

6 CONCLUSION & OUTLOOK 41
7 APPENDIX A - GMBI LIST AND FACTOR LOADINGS FROM FCTOR ANALYSIS 43

BIBLIOGRAPHY 45

LIST OF FIGURES

Figure 1 Robert Plutchik's wheel of emotions 14
Figure 2 Russel's valence-arousal model 15
Figure 3 Keyword-spotting algorithm typical stages 18
Figure 4 Contextual relationships between words learned by word embedding systems 20
Figure 5 Sentiment analysis framework using an hybrid approach (taken from [16]) 21
Figure 6 Overall model of emotion classification system (taken from [24]) 22
Figure 7 NMF of the term-document matrix (taken from [25]) . . 34
Figure 8 Computation of the songs' affiliation to the clusters (taken from [25]) 35
Figure 9 Scatter plot of the song lyrics in the VAD space 38
Figure 10 Correlation scores from the experiment and from our method 38

LIST OF TABLES

Table 1 MIREX music mood classification task clusters 15
Table 2 Common features for content-based audio mood analysis 17
Table 3 List of variables gathered from the listening experiment 24
Table 4 Pearson's correlation between GMBI factors scores and Synesketch scores 26
Table 5 Pearson's correlation between GMBI factors scores and Tone Analyzer emotional scores 27
Table 6 Pearson's correlation between GMBI factors scores and Tone Analyzer social tones scores 27
Table 7 Comparison of different classifier on Fairy Tales dataset with word2Vec features 31
Table 8 Confusion matrix for the training set 32
Table 9 Confusion matrix for the test set 32
Table 10 Classification results using different set of features 33

ACRONYMS

API Application Programming Interface

VAD Valence-Arousal-Dominance

ANEW Affective Norm for English Words

SVM Support Vector Machine

GMBI Global Music Branding Inventory

BOW Bag-of-Words

NMF Non-negative Matrix Factorization

1 INTRODUCTION

1.1 THE AUDIO-KOMMUNIKATION GROUP

This internship took place within the Audio-Kommunikation Group of TU Berlin, which is the largest scientific university in Berlin, and one of the most well know in Germany. The Audio-kommunikation group is dedicated to the communication of music and speech in acoustical or electro-acoustical systems. Research and teaching topics deals with audio recording and reproduction technologies, 3D audio with binaural technology and sound field synthesis. The laboratory also work on the composition and realization of electro-acoustic music as well as empirical approaches to study the reception of media content. Research projects at Audio-Kommunikation group are interdisciplinary studies reaching from humanities, cultural studies, across social sciences and psychology to computer science and engineering. The department is particularly know for running the world's largest wave field synthesis installation which contains more than 2700 loudspeakers.

To name a few, current research projects includes "Survey Music and media" : a project that aims to analyze the dominating patterns of audio media usage in Germany, or "Acoustical Investigations of Theatrical Spaces in the Early Modern Era", as well as "Sound Field Synthesis for Line Source Arrays".

My work was under the supervision of Dr. Jochen Steffens whose research interests includes functions of listening to music, noise assessment, multimodal interactions, musical taste, psychoacoustics and product sound design.

1.2 CONTEXT

This internship is part of the research project 'ABCDJ' (Artist-to-Business-to-Business-to-Consumer Audio Branding System). This project seeks to provide new tools for European creative agencies in the field of audio branding in order to actively include creators of music, (independent) labels as well as respective multipliers in the audio branding value chains. The research contribution of TU Berlin investigates (in cooperation with the University of York) the correlations between musical contents and their acoustic parameters on one hand and brand attributes (such as 'sportiness', 'elegance', 'reliability') on the other hand. Based on the research findings, this project aims to develop tools that enable brands and branding agencies to identify brand-fitting music titles from large music archives in order to automatically create playlists that may be used for marketing and point of sale branding activities such as selection of in-store music. A further fundamental research aim behind this application-oriented approach is to identify music-internal and external parameters which are constitutive for the perceived 'semantics' of music. To answer this question, a

pan-European listening experiment was carried out, resulting in a unique comprehensive empirical data basis. Using machine learning methods, a statistical model will be developed from this in order to predict semantic connotations of musical pieces based on musical and acoustical features of music archive titles, the latter being generated by Music Information Retrieval methods. Furthermore, the statistical model will incorporate the cultural and sociodemographic background of listeners.

1.3 OBJECTIVES

My contribution to this research project aims to investigate the text features that can be extracted from the lyric content of the songs. The main goal of my work was to answer the questions : How can text features can be extracted from lyrics ? How to build consistent lyric-based predictors and how to take them into account in the modeling approach. In a first step we did a benchmark of existing text mining tools in order to quantify the contribution of the text features in the future model. This benchmark will take into account online APIs as well as frameworks developed for scientific purposes. We will take a closer look on what are the different kind of informations that can be extracted from the text and how relevant these data can be when considering them as features for a classification or regression system. The main idea behind this work is to propose different techniques and axes of research for detecting emotional content of lyrics.

1.4 ORGANIZATION OF THE REPORT

This report is organized as follows: foremost we will look into the state of the art in the different fields related to our topic (Chapter 2) : we will describe an overview of the studies on music and emotion, the techniques used for emotion detection in text and the recent research concerning multi-modal mood classification of music. In a second part we will present a non exhaustive API study for emotion detection (Chapter 3). We thus consider two different APIs (Synesketch & IBM Tone Analyzer) that helped us to compute emotion scores from the lyrics. Then in chapter 4 we will introduce the so-called *Machine Learning approach* to process text, by presenting the main features extracted from text as well as different classifiers for emotion classification of text. We also introduce a topic modeling approach to detect song topics in order to links them to emotional semantics. In the last part (Chapter 5), we present the *Keyword-spotting approach* by introducing the JeMAS framework which is used to compute emotion scores. Finally, we will conclude this thesis by interpreting the results and discussing the outlooks of this work.

2

STATE OF ART

2.1 MUSIC AND EMOTION : OVERVIEW

In this section we introduce the concept of emotion and draw a parallel with the concept of mood which is more common in the Music information retrieval context. We then explain why this concept tend to be highly subjective, especially when conducting music listening experiments. Then, we introduce the two main emotion models : the categorical model and the dimensional model namely. Finally, we focus on the study of music and emotion in its psychological aspects.

2.1.1 *The concept of emotion*

EMOTION DEFINITION Referring to the Larousse dictionary, emotion is a transient emotional reaction of relatively high intensity, usually caused by stimulation from the environment. Another definition by Cabanac [8] explains that emotion is any conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure. Though, there is no known consensus on a definition as this concept is complex and not easy to define. A statement from Fehr and Russell [14] says : "Everybody knows what an emotion is, until you ask them a definition". The term emotion is often used interchangeably with mood. Indeed, these concepts are linked together according to Paul Ekman [13]. The term of mood refers to a more continuous emotional state and it can last longer than an emotion which is usually spontaneous. In the context of the study of music, it should be more appropriate to speak of emotions conveyed instead of moods.

SUBJECTIVITY The main issue people are facing when studying emotional behavior is subjectivity. Indeed, there is numerous factors that can change the perceived emotion from a person to another, such as past experiences and memories, cultural background, gender, age, personality or social factors. Regarding music, perceived emotion is intrinsically subjective because of musical preferences. In a recent study, McDermott [33] shown that there is cultural variation in music perception. This is obvious when referring to different ethnic groups (e.g. comparing western and indian cultural backgrounds), but more fine-grained factors such as age, gender or personality are factors that induce different emotion perception [27].

2.1.2 *Emotion models*

The nature of emotional phenomena can be interpreted in different ways. Thus, affective computing researchers have built two main model for representing and studying emotion. Each of these two representation are common in the psychology research.

CATEGORICAL MODEL This emotion model is representing emotions into discrete categories. Thus it assumes that people experience emotions that are distinct from each other [45]. The first psychologist who introduced such an emotion model is Paul Ekman [12] when he classified emotions into six categories, the so-called basic emotions : anger, fear, disgust, sadness, happiness, surprise. Though this model was widely used for psychological research purposes it was developed initially for interpret facial expression so it it not always accurate for other cases. Furthermore, it is a simple model that can lack of exactness.

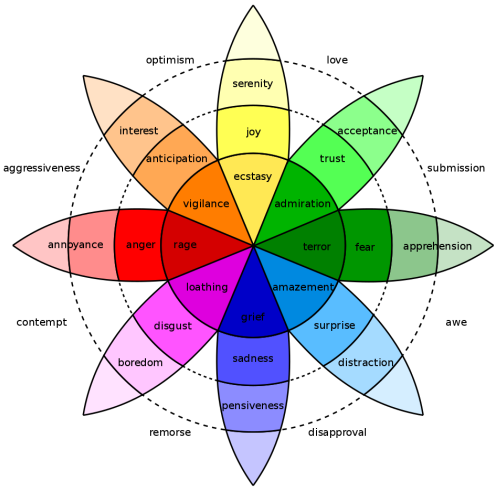

Figure 1: Robert Plutchik's wheel of emotions

In parallel, Robert Plutchik introduced in 1980 [37] the wheel of emotion (Fig. 1), assessing that there is eight different basic emotions and decline them with nuances. This eight emotions are described with bipolarity : joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation.

The categorical model of emotion has been applied in other ways. For instance in the Music Information Retrieval research community, where five mood clusters have been created for MIREX music mood classification as seen on [20].
According to [24], these five categories were constructed following a clustering on a co-occurrence matrix of mood tags from the All Music Guide[1]

1 All Music Guide : http://www.allmusic.com

CLUSTER	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 1: MIREX music mood classification task clusters

DIMENSIONAL MODEL This other main paradigm suggest that mood can be measured by simple multi-dimensional metrics. In this type of models, emotions are organized along two or three axes, and it is commonly assumed that this representation tend to be more accurate and entails lower ambiguity than the categorical model. Seminal work from Russel [39] and Thayer [43] established one of the most well know model in this category : the valence-Arousal (V-A) space (see Fig. 2) where emotions are located in a two dimensional space : arousal which is the intensity of an emotion (Y-axis) and valence (X-axis) which correspond to its pleasantness. Multiple studies concerning music emotion recognition have used this representation [30], [23].

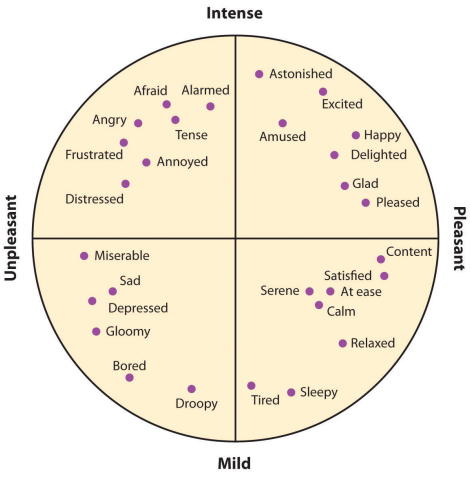

Figure 2: Russel's valence-arousal model

This model can be completed with a third dimension called dominance, which is a sense of control to act when feeling the emotion. However, this last dimension is subject to controversial comments and disagreements [4]. In some studies [46], researchers used this model but divide it into four discrete categories, called quadrants. But even with the emotion plane, the categorical taxonomy of emotion classes is still inherently ambiguous. Each emotion class represents an area in the emotion plane, and the emotion states within each area may vary a lot. For example, the first quadrant of the emotion plan (high

arousal, high valence) contains emotions such as excited, happy, and pleased, which are different in nature.

2.1.3 *Emotion in music*

The study of emotion in music aims to understand the psychological relationship between human affect and music. Though it is a branch of music psychology with numerous areas of study (emotional reactions, listener's characteristics, components of musical composition which induce emotion...), it can be considered as a computational problem when we try to analyze automatically the emotional content of music.

CONVEYING EMOTION THROUGH MUSIC The ability to perceive emotion in music is said to develop early in childhood, and improve significantly throughout development [11]. The capacity to perceive emotion in music is also subject to cultural influences, and both similarities and differences in emotion perception have been observed in cross-cultural studies. Empirical research has looked at which emotions can be conveyed as well as what structural factors in music help contribute to the perceived emotional expression.

In [40], Scherer et al. have argued that the emotion experienced from a piece of music is a function of structural features, performance features, listener features and contextual features of the piece. In our work, we focus on the structural features of a piece in order to detect perceived emotion automatically.

STRUCTURAL FEATURES Structural features of a piece of music includes the acoustic structures such as duration, amplitude, or pitch but also the foundational structures such as melody, tempo, rhythm or even harmony, as explained in [40]. As said by Gabrielsson et al. in [15], there are a number of specific musical features that are highly associated with particular emotions. For example, tempo is typically considered as one of the most important : a fast tempo can induce happiness, excitement or anger, as a slow tempo can express sadness or serenity. A number of other factors such as mode (scale), loudness and melody also influence the emotional valence of a piece.

AUTOMATIC EMOTION RECOGNITION These structural features are often used for automatic mood classification. Indeed study like [29] used among others features like tempo, mode, key strength, loudness or timbre descriptors such as MFCC (Mel-Frequency Cepstral Coefficient). In Table 2 we sum up the most common features that are used for automatic mood recognition.

TYPE	FEATURES
Timbre	Mel-frequency cepstral coefficients (MFCCs), spectral shape
Harmony	Roughness, harmonic changes, key clarity, maharanis
Rhythm	Rhythm strength, regularity, tempo, beat histograms
Register	Chromagram, chroma centroid and deviation
Articulation	Event density, attack slope, attack time
Dynamics	RMS energy

Table 2: Common features for content-based audio mood analysis

2.2 EMOTION DETECTION FROM TEXT

Nowadays, the task of emotion detection from text is becoming increasingly studied and is deserving a lot of attention in the scientific community. Indeed, with the growing popularity of social networks as well as the exponential number of reviews published on the Internet there is a need for companies to analyze people thoughts on different products or services. Still, in comparison to the other areas of emotion detection such as audio, speech and facial emotion detection, there is a lot of room for research in text emotion detection. We can distinguish two main approaches in current text emotion recognition systems : the keyword-spotting approach and the learning based approach.

2.2.1 *Keyword spotting approach*

The keyword spotting approach (or lexicon-based approach) consists to detect the presence of keywords which have high absolute valence or which are strongly correlated to a specific emotion class.

LEXICONS Typically, these systems use one or several lexical resources. The most common lexical resource is WordNet, a lexical database introduced by Miller in 1995 [35]. In 2004, Strapparava et al. [42] extended this english word database by creating WordNet-Affect, which is an extension of WordNet that includes emotional informations. This lexicon is well adapted for experiments using the categorical model of emotions. In [41], they used this extension in order to implement a simple algorithm that check the presence of affective words. Another well known lexical resource was developed by Bradley in 1999 [6]. The Affective Norms for English Words (ANEW) consists of a large set of words labeled with valence, arousal and dominance values. This lexicon is used for studies using the dimensional model of emotion.

TECHNIQUE The idea behind these algorithms is to consider a text as an input and generate an emotion class or score as an output (See Fig.3). In a first step, the text is preprocessed in order to analyze it at a sentence level or

at a word level. Especially, the text document is converted into tokens using a parser. In some cases, text preprocessing include *stopwords* removal (also called function words). These words include mainly determinants, pronouns and other particles which convey no emotions. After this step, the algorithm consider a list of words or a list of sentences instead of a long string. Next, Within these tokens, the emotional words are identified (Keyword detection). When a keyword is detected, an emotional score is associated to the sentence or the document. This step is followed by applying some heuristic rules, which in most of the case consists of negation check (that would flip the valence of a sentence / word), or taking into account specific punctuation (for instance exclamation point at the end of a sentence) or intensifying words (emotional keyword preceded by words such as "very", "extremely"...) [26]. The final score / class of a sentence or a document is obtained by an overall averaging.

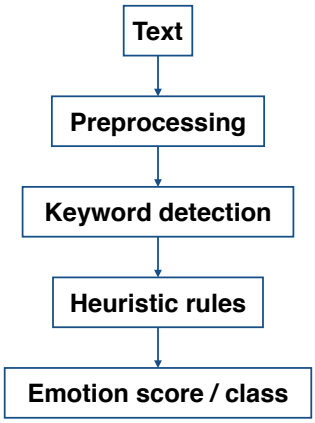

Figure 3: Keyword-spotting algorithm typical stages

LIMITATIONS Referring to the study of Hirat and Mittal [17], there are some limitations in the Keyword-spotting approach :

- Ambiguous definition of keywords. Some emotional keywords can have different meaning according to the context in which it is used. The common lexicons that are used don't take the different meanings into account and thus it can lead to wrong interpretations.
- Keywords are the only source of emotional content. The algorithms are based only on the presence of emotional keywords in the text. But sometimes a sentence with none of these keywords can still convey emotions that are passed by the main idea of the sentence.
- Lack of linguistic information. Syntactic and semantic structures of sentences can radically change their meanings and thus the emotion conveyed. For instance : "I laughed at him" and "He laughed at me" express opposed feelings.

2.2.2 *Machine learning approach*

This approach is based on training classifier or regressors trough extracted semantic features of emotion-annotated datasets of documents. That lead to building models that can predict the emotional content of any document. When using supervised learning methods, this approach requires large annotated datasets for training. That is one problem researchers in NLP face nowadays : the lack of annotated data.

FEATURE EXTRACTION In text analysis, the most used features are content-based features or bag-of-words (BOW). These features turn the text to analyze into a set of bags that are n-grams which are contiguous sequences of n items from a given sequence of text. In most of the case, n-grams are extracted up to n=3 (unigrams, bigrams and trigrams). Illustrating, in the sentence below we show the unigrams, bigrams and trigrams representations :

"Bringing back sweet memories "
Unigrams : Bringing; back; sweet; memories
Bigrams : Bringing back; back sweet; sweet memories
Trigrams : Bringing back sweet; back sweet memories

When extracting unigrams from text, this is equivalent to create a dictionary of the words contained in the text and counting each time a word appear (term frequency). When we increase the order of n-grams (n>1), this allows to catch more relevant contextual and linguistic information. Typically, when creating a term-document matrix by using n-grams features, a term frequency - inverse document frequency (TF-IDF) weighting is applied to the n-gram feature matrix. The tf-idf value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Thus, this feature represents a document d as a vector of terms t weighted by the following function :

$$\mathrm{tfidf}(d,t) = \mathrm{tf}(t) * \mathrm{idf}(d,t) \tag{1}$$

where:

- $\mathrm{tf}(t)$ is the term frequency of term t
- $\mathrm{idf}(d,t) = \log(\frac{n}{\mathrm{df}(d,t)}) + 1$
- n is the number of documents
- $\mathrm{df}(d,t)$ is the number of documents d that contain term t

Another type of feature which is often used is the Part-Of-Speech tags (POS tags). Part-of-speech tagging is a lexical categorization or grammatical tagging of words according to their definition and the textual context they appear in.

Different part-of-speech categories are for example nouns, verbs, articles or adjectives. The POS tagging is typically followed by a BOW analysis. In [32] Mayer et al. are using POS tags, among other features, on lyrics for genre classification task.

In some studies like [19], people are using text stylistic features. They include for instance interjections ("yeah", "woo"), punctuation marks, and statistics over the type of word used, number of unique words, number of lines. These kind of feature can bring relevant information for text classification task as for instance genre recognition.

More recently, the development of *word embedding* systems as Word2Vec [34] and GloVe [36] led to huge improvements in some NLP related tasks. Word embedding aims to quantify and categorize semantic similarities between linguistic items based on their distributional properties in large samples of language data. When trained on large datasets, these algorithms produce accurate vector space models of words (ie. vector representation of words) that catch very fine-grained contextual information (see figure 4). In particular, this facilitates the semantic analysis of the words. This new representation has the typical characteristic that words appearing in similar contexts have corresponding vectors which are relatively close. For example, one might expect that the words "dog" and "cat" are represented by vectors relatively close to each other in the vector space where these vectors are defined.

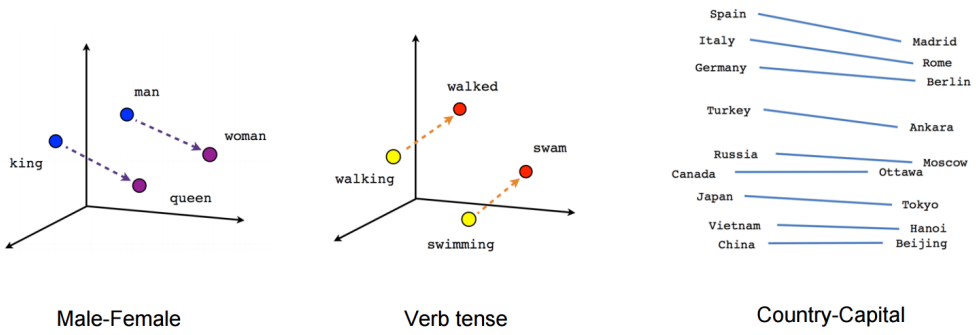

Figure 4: Contextual relationships between words learned by word embedding systems

EMOTION CLASSIFICATION In the literature, common classifiers used for text classification in a supervised way includes Random Forest, Naïves Bayes and SVM (Support Vector Machine). Naïve Bayes and Random Forest often serves as a baseline, while SVM seems to achieve top performance. For emotion detection using the categorical model, the problem is a mutliclass classification, while when dealing with the dimensional model, it turns into a regression problem [7]. In [1], they used an annotated corpus with an extended set of Ekman's basic emotions. In [41], Strapparava et al. applied Naïves Bayes classifier trained on a blog corpora annotated with Ekman's emotions. More recent works ([3], [38]) used emotion classifier based on multi-class SVM. In

[31], R. Malheiro et al. did a comparative study and various tests with SVM, K-Nearest Neighbors, C4.5 and Naïves Bayes classifiers on audio and lyrical features. The best results where always reached with SVM. In order to improve the performance of the classifiers, feature selection is usually performed to reduce the number of feature. Indeed, when considering large dataset of text, n-grams features are many.

HYBRID APPROACHES This approach is a combination of the previous methods. They thus can improve results from training combination of classifiers and adding knowledge-rich linguistic information from different ressources and dictionnaries [5].In a recent work by Giatsoglou et al. [16], they implemented a sentiment analysis framework by using word embedding-based features as well as sentiment lexicon data in order to build hybrid vector of document. These vector are then used for training an SVM classifier. The figure 5 shows the operation of their framework.

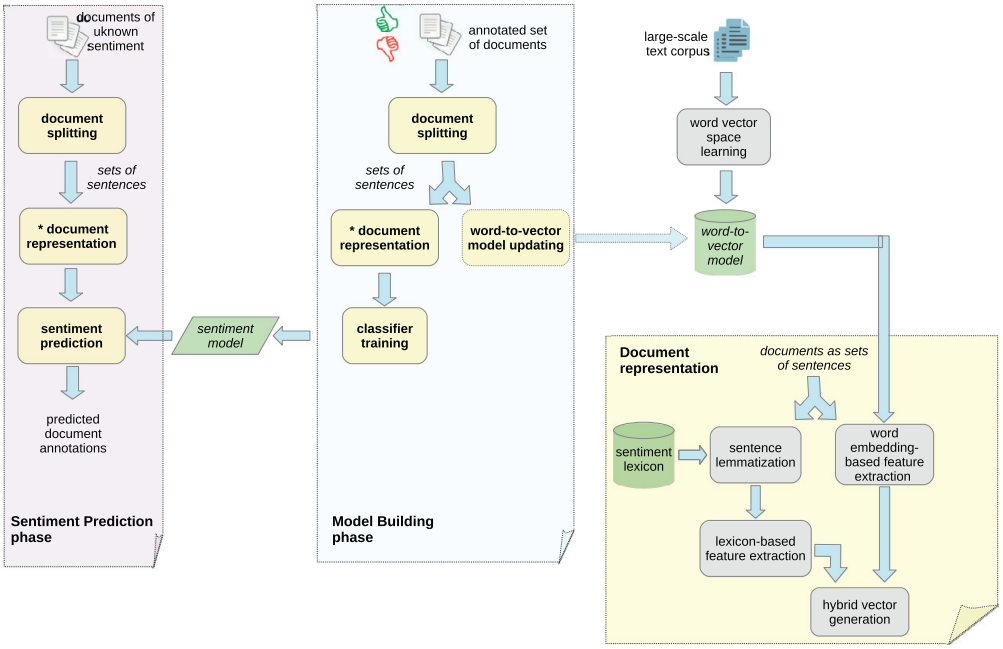

Figure 5: Sentiment analysis framework using an hybrid approach (taken from [16])

2.3 MULTIMODAL MUSIC CLASSIFICATION

In the past few years, music information retrieval has been very active and had produced automatic classification methods in order to deal with the amount of digital music available. The task of emotion / mood classification is relatively recent. It has been showed that audio-based techniques can achieve satisfying results [29]. More recently the study of other type of data (social tags coming from streaming websites, reviews, lyrics, video) gain a lot more attention, since they can yield to useful information for genre recognition, artist similarity,

or mood classification tasks [24]. In [22], Juslin showed that 29% of people mentioned the lyrics as a factor of how music express emotions, thus it is relevant to study the lyrics in that context. Multimodal systems aims is to focus on the complementarity of two (or more) different sources of information to automatically classify songs (see Fig. 6). In our study, we focus on lyrics and audio.

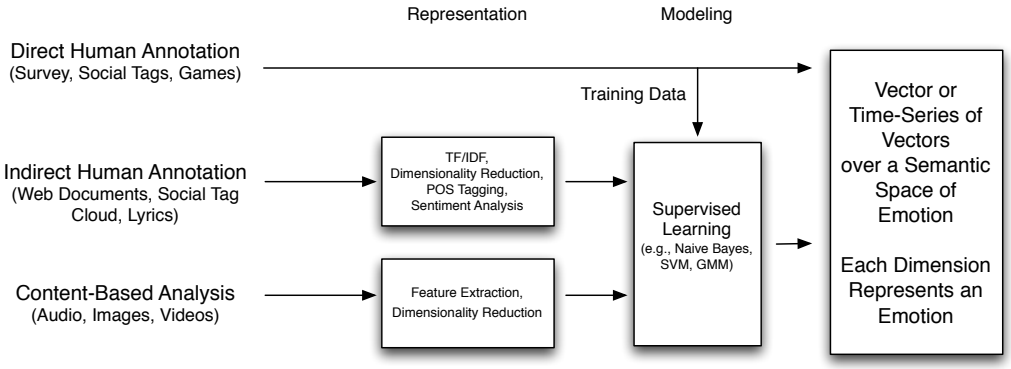

Figure 6: Overall model of emotion classification system (taken from [24])

In [28], Laurier et al. did several expreriments for emotion classification on lyrics : using k-NN (k-Nearest Neighbors algorithm) with a TF-IDF distance or latent semantic analysis with SVM, logistic and random forest classifier. When mixing the audio features and the lyrics, their results prove that audio and lyrics information combined led to better music mood classification system.
In [46], Yang et al. exploited audio features (such as MFCCs, spectral centroid, spectral moment & roughness) and lyrical features (unigrams, bigrams, LSA latent vectors) for performing a 4-class classification system using Support vector Machine. They introduced three different methods for fusing audio and text cues. Each of them led to better accuracy results (the best approach improves from 46.6% to 57.1%).
In [21], they also used a combination of audio and lyrical features to perform a mood classification using the k-nearest neighbors algorithm. The audio features included BPM, Mode, loudness,danceability and energy. The lyrical features consisted on valence and arousal scores from ANEW and WordNet ressources. They achieved and average accuracy of 83.4% on 795 songs divided into 9 classes.

3

API STUDY

In this chapter, we will focus on the study of two APIs, (Application Programming Interfaces) that helped us to compute emotional score from text files containing the lyrics of the song we study. In a first part, we will explain how did we build the dataset of lyrics by mining the data on the web. Secondly, we will introduce the ground-truth sample coming from a listening experiment. Then, we present *Synesketch* API, an open source library for sentence-based emotion recognition. Finally, we will present the IBM *Tone Analyzer* API. We correlated the results from these two APIs with the data from the experiment in order to conclude if the lyrics have an impact on the emotional content of the songs.

3.1 BUILDING THE LYRIC DATASET

First of all, the first task was to constitute the lyric dataset. From a list of 183 songs, the goal was to retrieve automatically the lyrics from the Internet by looking at several well-known website that gather lyrical resources. To do so, we use the *Lyricfetcher*[1] Python package which allows us to search lyrics from different web sources (genius[2], lyricswikia[3], metrolyrics[4] and azlyrics[5])
We then wrote a Python script using this package that allowed us to gather 89 lyric files out of the 183 tracks of our list. This was due to a high number of instrumental tracks, no corresponding lyrics on the different web sources, as well as a number of non-English tracks that we not considered, since the APIs we use handle only English language. A text processing part was also included in the script in order to get rid of some present non-lyrics information like "Chorus" or "solo".

3.2 GROUND-TRUTH SAMPLE AND EXPLORATORY FACTOR ANALYSIS

GROUND-TRUTH SAMPLE For this analysis, we use an empirical ground-truth based on an online study. This study was conducted in UK, Spain and Germany on 3485 participants. The data gathered in this study includes socio-demographic variables such as country, gender, age group and education, but most importantly it contains music-related ratings on 51 adjectives. These adjectives were established by music branding experts and thus constitutes the Global Music Branding Inventory (GMBI) (See appendice A). Listeners in the

1 https://github.com/bharatkalluri/lyricfetcher
2 https://genius.com/
3 http://lyrics.wikia.com/
4 http://www.metrolyrics.com/
5 https://www.azlyrics.com/

study were able to rate the degree of fit of these attributes to the music as perceived. The data also includes direct manifest ratings of "liking" and "knowing" of the music titles listened. During this experiment, every participant had listened to randomly selected four 30 seconds excerpts of the 183 titles of our tracklist. The structure of the ground-truth sample is described in Table 3.

VARIABLE	MEASUREMENT LEVEL
Survey country	1-UK 2-Spain 3-Germany
Education	1-ISCED 0-3 2-ISCED 3-4 3-ISCED 5-6
Age group	1- 18-24 2-35-51 3- 52-68
Gender	0 - male 1 - Female
Degree of GMBI fit	Range 1-6 Very bad fit - very good fit
Degree of liking	Range 1-6 Very bad fit - very good fit
Degree of knowing	Range 1-6 Very bad fit - very good fit

Table 3: List of variables gathered from the listening experiment

For our study we will consider only the ratings of the degree of GMBI fit. Indeed, this data brings relevant information about the emotional content of the song as perceived by the participants.

FACTOR ANALYSIS The factor analysis which was performed by Steffen Lepa and Jochen Steffens aims to use latent variable factors to model the dependent variables in the prediction model for musical expression. Indeed, we exploited the innate semantic redundancy of language when dealing with music semantic description. This factor analysis was performed to determine the number of independent latent semantic dimensions within the rating items. The final goal of this approach is to reduce the number of variable (GMBI items) to the four of them that explain the most the ratings within the experimental study. The four items that resulted are "Easy-going", "Joyful", "Authentic" and "Progressive". In appendix A, we show the different factors loadings between each GMBI item and this four factor.

EVALUATION OF THE APIS In order to evaluate the performance of each APIs, we calculated the Pearson's correlation coefficient between the factor scores values of each track (taken from the ground-truth dataset) and the each emotion scores given by the API. The Pearson's correlation coefficient between two random variables X and Y is defined as :

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

Where $Cov(X,Y)$ is the covariance between the variables X and Y and σ their standard deviation. We can write this equation as :

$$Cor(X,Y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{3}$$

where :

- x_i, y_i are the single samples indexed with i
- $\bar{x} = \frac{1}{n}\Sigma x_i$ is the sample mean

By performing the correlations between variables, we can conclude if the API can automatically explain the data gathered in our experiment. Thus it is a good indicator if a specific technique for textual emotion recognition algorithm can be implemented in our system.

3.3 SYNESKETCH API

3.3.1 *Operation of the API*

The Synesketch API was proposed in 2013 by Krcadinac et al [26]. They introduced a recognition algorithm that classifies a text according Ekman's six basic emotions : happiness, sadness, anger, fear, disgust and surprise. The algorithm estimates emotional weights for each category in the form of a numerical vector. It computes these scores at a sentence level and then does an overall average emotional score of the document. The method is based on the keyword-spotting approach that use lexicons of words and expressions related to emotions, and also includes several heuristic rules.

3.3.2 *Results*

In order to compute the emotion scores with synesketch API, we had to modify the framework's code so that we calculate automatically all score for each songs and gather them into text file. Once this was done, we then calculate the correlations between the factor analysis data (factor scores for "Easy-going",

	JOY	SADNESS	ANGER	FEAR	DISGUST	SURPRISE
Easy-going	0,076	**0,374**	0,114	0,138	0,157	0,083
Joyful	0,094	-0,27	-0,104	-0,033	-0,142	0,028
Authentic	**0,211**	0,012	0,042	0,045	0,161	0,006
Progressive	**-0,234**	-0,127	0,084	-0,025	0,001	-0,077

Table 4: Pearson's correlation between GMBI factors scores and Synesketch scores

"Joyful", "Authentic" and "Progressive" for each song) and the emotion scores (scores for "happiness", "sadness", "anger", "fear", "disgust" and "surprise"). The results of the correlations are presented in Table 4. Here, the correlation was calculate for the 89 songs for which we have the lyrics.

INTERPRETATION The results we have from Synesketch correlate with some GMBI factors (See Table 4). The significant correlation are shown in bold. In statistics, Significant correlation value means that it is unlikely that it was obtained by a mere chance. The significance threshold is set to 0.05, Which means that the observed result has less than a 5% chance of being obtained by chance. It is quite complicated to interpret these correlations as the GMBI factors are to easy to explain semantically. We can see that *easy-going* correlates well with *Sadness* in a positive way, that means that for a relatively high amount of songs, when the people have given high score for *easy-going*, the synesketch API detected *sadness* in the lyrics. We can explain this by saying that people rated *easy-going* when they heard a melancholic melody, which also can express sadness. The two other significant values of correlation are more complex to explain. We have to look at the loading matrix (See appendix A) and notice that *progressive* is related to the GMBI *young* and *urban*. So people could have rated these GMBI in the case of rap songs for example, in which the lyrics are not expressing happiness.

3.4 IBM WATSON TONE ANALYZER

3.4.1 *Operation of the API*

Tone Analyzer is a commercial service from IBM that uses linguistic analysis to detect emotions (joy, fear, sadness, anger, disgust), social tones (analytical, confident, tentative), and language style (openness, conscientiousness, extroversion, agreeableness) cues found in text. An emotional range score also give a measure of how the text convey emotional intensity. *Tone Analyzer* service is based on the theory of psycholinguistics, intended for exploring the relationship between linguistic behavior and psychological theories. The service uses linguistic analysis and the correlation between the linguistic features of written text and emotional, social, and language tones to develop scores for

each of these tone dimensions. To derive emotion scores from text, IBM use a stacked generalization-based ensemble framework; stacked generalization uses a high-level model to combine lower-level models to achieve greater predictive accuracy. Features such as n-grams (unigrams, bigrams, and trigrams), punctuation, emoticons, curse words, greetings (such as "hello," "hi," and "thanks"), and sentiment polarity are fed into machine-learning algorithms to classify emotion categories.
Though this service is intended to be use for commercial purpose (customer relations for instance), we found it interesting to use this API in order to analyses lyrics because the model build by IBM tend to be more accurate.

3.4.2 *Results*

In order to compute the results from the *Tone Analyzer* API, we developed a script that automatically "call" the API for each lyrics text files, and then get the scores and process the results in order to get them in a table. We then performed the Pearson's correlation between the GMBI factors and the different tone scores given by the API. The results are shown in Table 5 and Table 6. Significant values at a 0.05 threshold are shown in bold.

	Anger	Disgust	Fear	Joy	Sadness
Easy-going	-0,188	-0,001	0,09	0,2	**0,203**
Joyful	0,024	-0,107	0,028	0,133	-0,136
Authentic	**-0,257**	0,037	0,081	0,006	0,005
Progressive	**0,341**	0,013	0,021	-0,049	-0,054

Table 5: Pearson's correlation between GMBI factors scores and Tone Analyzer emotional scores

	Opennes	Conscientiousness	Extraversion	Agreeableness
Easy-going	0,007	0,035	**-0,232**	0,156
Joyful	-0,180	0,061	**0,289**	0,141
Authentic	0,008	-0,179	-0,085	-0,064
Progressive	-0,022	0,196	0,023	0,028

Table 6: Pearson's correlation between GMBI factors scores and Tone Analyzer social tones scores

INTERPRETATION The results from*Tone Analyzer API* shows some significant (at a 0.05 threshold) correlations with the GMBI factors. We can see that

"anger" is negatively correlated with "authentic" and positively correlated with "progressive". From this last result we can draw a parallel with the Synesketch results that showed negative correlation between "Joy" and "progressive". We can also notice that "sadness" is correlated with "easy-going", a result that appear as well with Synesketch.
Concerning the social tones scores, the two significant correlation are quite logical : "extraversion" is negatively correlated with "easy-going" and positively correlated with "joyful".

3.5 CONCLUSION

In this chapter we used two different APIs, *Synesketch* and *IBM Tone Analyzer*, in order to compute emotion or tones scores from our lyrics. For that we developed scripts that allowed us to perform this task automatically and get the output data in the right form. To evaluate the accuracy of each APIs, we performed correlations between the four GMBI factors and the APIs scores. We concluded that we can notice some significant correlations which are sometimes quite difficult to interpret. We didn't get high correlation values for keywords which are semantically the same, for instance joyful and joy which point out the fact that our ground-truth can be biased due the the different categories of people such as age, gender, nationality but most of all, the fact that subjectivity is the main problem when dealing with emotion analysis of music.

4

MACHINE LEARNING APPROACH

In this chapter, we study the so-called machine learning approach for detecting emotion into text documents. For that, we consider the categorical model for representing emotions, so we make the hypothesis that a one lyric document belongs to one emotion class. In the first section we explain the procedure we adopted to apply this technique to our problem. This approach requires emotion-annotated datasets of text documents that we present in section 2. Then, we present the extracted features for training the classifiers, as well as different preprocessing steps in order to get better results. In section 4, we introduce the classifiers we experimented. Then we explain the different experiment that was performed on the song lyrics for emotion classification and the results we obtained from these experiments. Finally, we study the Non-negative Matrix Factorization technique in the context of topic modeling of text data.

4.1 PURPOSE

The purpose of those experiments was to build accurate models for text emotion classification. The procedure is to train different classifiers on emotion-annotated dataset in order to get the maximum accuracy. Once an acceptable model is build we can perform emotion classification on lyrics.

4.2 EMOTION-ANNOTATED DATASETS

4.2.1 *Fairy tales dataset*

This dataset was build by Cecilia Ovesdotter Alm for her PhD study : "Affect in Text and Speech" [2]. This dataset include 176 fairy tales in which sentences were extracted, and each of them was annotated with an emotion class. This dataset contains approximatively 1000 annotated sentences. The emotion considered are : "Angry-Disgusted", "Fearful", "Happy", "Sad" and "Surprised".

4.2.2 *ISEAR dataset*

ISEAR stands for International Survey On Emotion Antecedents And Reactions. This dataset was build by conducting a survey on student respondents, both psychologists and non-psychologists, who were asked to report situations in which they had experienced all of 7 different emotions (joy, fear, anger, sadness, disgust, shame, and guilt). In each case, the questions covered the way they had appraised the situation and how they reacted. The final data set thus

contained reports on seven emotions each by close to 3000 respondents.This dataset contains 7666 emotion-annotated sentences.

4.3 EXTRACTED FEATURES & PREPROCESSING

4.3.1 *n-grams*

We previously introduced these features in section 2.2.2. For our purpose, we extract unigrams, bigrams and trigrams (n 3). Going further in the rang would increase too much the dimension of the feature matrix. We use the Scikit-learn[1] Python package for extracting the n-grams from the text. The function `CountVectorizer` allows us to extract word n-grams, choose the range, remove stopwords and define a maximum of word to extract for creating the dictionary. The output of this function is a *term-document* matrix where the number of columns is the defined maximum of features (words) and the number of row is the number of sentences to analyze.
Once we have a term-document matrix, we can apply the TF-IDF transform (see section 2.2.2). The goal of using tf-idf instead of the raw frequencies of words of a token in a given document is to scale down the impact of words that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus. We use the `TfidfTransformer` function that computes for a term t the tf-idf function defined as :

$$\mathrm{tfidf}(d,t) = \mathrm{tf}(t) * \mathrm{idf}(d,t) \tag{4}$$

where:

- $\mathrm{idf}(d,t) = \log[n/\mathrm{df}(d,t)] + 1$
- n is the number of documents
- $\mathrm{df}(d,t)$ is the number of documents d that contain term t

4.3.2 *Word2Vec vectors*

As introduced in section 2.2.2, Word2Vec is a group of neural networks model that produce word embeddings. These word vectors are trained to reconstruct the linguistic context of words. Thus, these features are very informative for our purpose. For training these models we needed a large database of text. We used instead a pre-trained model that includes word vectors for a vocabulary of 3 million words and phrases that was trained on roughly 100 billion words from a Google News dataset. The vector length is 300 features. Since each word is represented by a vector, we did a basic averaging of all word vectors within a sentence to get a vector that this sentence. Thus, we got a feature matrix

1 http://scikit-learn.org/stable/

with 300 columns (Word2Vec features) and the number of row corresponding to the number of sentences in the dataset. In order to improve the classification accuracy of our classifiers, we normalize each feature vector by its L2 norm.

4.4 EXPERIMENTS

CLASSIFIER COMPARISON We did several experiments on the two annotated datasets, with the different features and classifiers . First, we used the Fairy Tales dataset with Word2Vec features and tested three different classifiers : Random Forest [18], Logistic Regression [10] and Support Vector Machines [9] with both linear and radial basis function kernel. In order to evaluate the accuracy of our trained models, we computed the precision, recall and F-measure for each class. These indicators are defined by :

$$Recall = \frac{number\ of\ correct\ classified\ documents\ in\ a\ class}{number\ of\ documents\ belonging\ to\ the\ class} \tag{5}$$

$$Precision = \frac{number\ of\ correct\ classified\ documents\ in\ a\ class}{total\ number\ of\ documents\ classified\ in\ the\ class} \tag{6}$$

$$F - score = 2 * \frac{precision * recall}{precision + recall} \tag{7}$$

In Table 7, we show the classifier comparison results. The emotion classes are shown as : {Angry-Disgusted, Fearful, Happy, Sad, Surprised}.

	F-score for each class	Classifier score
Random Forest (50 estimators)	[0.3035, 0.145, 0.416, 0.294, 0]	0.3125
Logistic regression (C=50 , 100 iterations)	[0.445, 0.353, 0.378, 0.266, 0.205]	0.358
SVM Linear (C=1)	[0.461, 0.353, 0.371, 0.266, 0.074]	0.341
SVM RBF (C=50, gamma = 0,4)	[0.473, 0.441, 0.403, 0.279, 0.176]	0.392

Table 7: Comparison of different classifier on Fairy Tales dataset with word2Vec features

The first thing we notice is that the F-score can vary a lot between a class from another. This id due to the unbalanced data we used. Indeed, some classes like *Angry-Disgusted* are overrepresented. We thus need to build manually balanced classes for better results. Secondly, we notice that the best classifier is the Support Vector machine with a radial basis function kernel. For the next experiments we will only use this classifier but tune its parameters C and gamma by

Real \ Predict	Angry	Fearful	Happy	Sad	Surprised
Angry	55	0	1	0	0
Fearful	1	45	3	6	1
Happy	0	0	55	1	0
Sad	4	0	2	47	3
Surprised	1	2	4	0	49

Table 8: Confusion matrix for the training set

Real \ Predict	Angry	Fearful	Happy	Sad	Surprised
Angry	33	4	9	0	2
Fearful	15	12	5	5	9
Happy	13	1	16	2	10
Sad	4	2	3	5	3
Surprised	7	5	4	0	7

Table 9: Confusion matrix for the test set

performing a *grid search* in order to get the maximum accuracy. Tables 8 and 9 shows the confusion matrix for a SVM RBF classifier with tuned parameters (C=10, gamma = 0,001) and with balanced classes.

We can see that the scores increased with balanced classes. Here the training score is 0.89 and the test score is 0.42.

ISEAR DATASET We performed other experiments on this dataset because it contains more data (7666 annotated sentences - 7 emotion classes : joy, fear, anger, sadness, disgust, shame, guilt) and thus lead to better results. We did these experiment on n-grams features with Tf-Idf weighting as well as Word2Vec. We also tried the so-called "bagging" technique which consists in combining features in one unique vector. The results of these experiments are shown in table 10. The scores shown between brackets stands for the F-score of each emotion class : joy, fear, anger, sadness, disgut, shame and guilt.

INTERPRETATION Firstly, these results are showing that the more there is features (ie. words) taken into account, the more the classification performs well. Indeed, with 5000 words instead of 1000, the documents are better represented as the vocabulary increase. We also noticed that taking bigrams into account improve the score a bit, because this allows to take some expressions into account, or couple of words that have a particular meaning. Against all expectation, the Word2Vec features don't outperform the TF-IDF feature. This is maybe due to the lack of large dataset for training. Finally, the bagging technique improve the classification results by few percent as expected.

FEATURES	SVM RBF (C=10, GAMMA=0,001) L2 NORMALIZATION
Word count Unigrams 5000 features	Train 0,986 Test 0,469 [0.617, 0.640, 0.350, 0.419, 0.412, 0.393, 0.430]
Word count Unigrams + bigrams 5000 features	Train 0,985 Test 0,486 [0.586, 0.649, 0.363, 0.494, 0.42, 0.435, 0.473]
TF-IDF Unigrams 1000 features	Train 0,972 Test 0,418 [0.495, 0.603, 0.329, 0.429, 0.364, 0.338, 0.378]
TF-IDF Unigrams + bigrams 5000 features	Train 0,981 Test 0,502 [0.523, 0.619, 0.365, 0.511, 0.441, 0.452, 0.497]
Word2Vec features	Train 0,798 Test 0,396 [0.391, 0.443, 0.258, 0.334, 0.384, 0.314, 0.294]
W2V + TFIDF combined 1300 features	Train 0,964 Test 0,429 [0.442, 0.418, 0.331, 0.365, 0.403, 0.392, 0.438]

Table 10: Classification results using different set of features

The best score was obtained with TF-IDF features, by extracting 5000 unigrams and bi-grams for building the vocabulary. We obtained 50% of good classification, which is far from we hoped in the beginning, but it turns out it is quite satisfactory given the lack of data.

4.5 TOPIC DETECTION WITH NMF

In this section we study Non-negative Matrix factorization which is an unsupervised machine learning technique that allows to detect topics in text documents. This study was inspired by the work of Kleedorfer et al. in [25].

4.5.1 *Non-negative Matrix factorization*

Non-negative matrix factorization is an algorithm that relies on linear algebra, it allows a matrix to be factorized into two other matrices and with the property that all three matrices have non-negative elements. For our purpose, we apply this algorithm on a term-document matrix in order to create term clusters in the matrix, that will represent our lyrics topics.

4.5.2 *Procedure*

We start by reading the lyrics files and creating the Term-document matrix, as same as done previously for classification purpose. We choose 1000 words to be taken into account in the vocabulary. This step include a stopwords removal from the text. Indeed those words with no specific meaning would not be able to semantically represent the topics.
Once we have our term document matrix, with dimensions corresponding to 1000 words (columns) and 89 lyrics (rows), we apply a tf-idf weighting schema in order to amplify the weights of the terms that are more typical for a song and lower the weights of the other terms.
Next, we perform the NMF to the term-document matrix. The most prominent parameter of the NMF is the number of cluster (or topics) k. With matrices dimensions as indices, we can write :

$$T_{documents*terms} = W_{documents*k} * H_{k*terms} \tag{8}$$

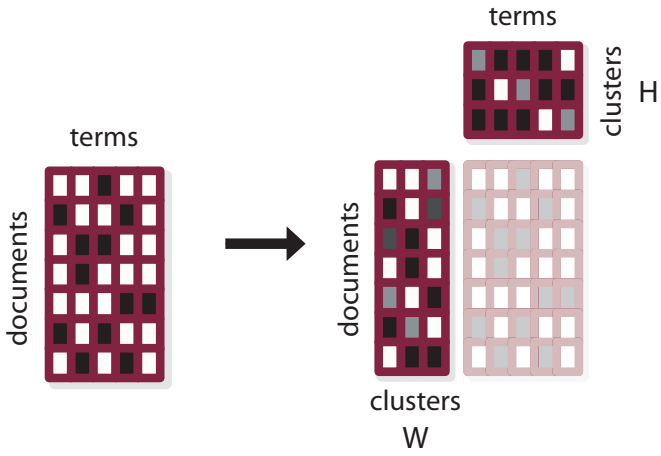

Figure 7: NMF of the term-document matrix (taken from [25])

Once the NMF is performed, we can look closely to the matrix H, which contains the weight of each term in each cluster. We thus can display for each cluster the top words. Here, we choose $k = 10$ clusters, and we display the top 5 words which have the bigger weight in the matrice H :

Topic 0: window rain way stand won
Topic 1: love say keeps fall heart
Topic 2: oh la high girl named
Topic 3: baby leave don come situation
Topic 4: think dreams stop matter try
Topic 5: town boys hard times brother
Topic 6: gonna change try lovin come
Topic 7: child yes folks understand just
Topic 8: hey mr dancing going play

Topic 9: dance just child let dark

We can see that terms in some cluster seems to be relatively coherent (topic 1 for instance) but some other are quite difficult to interpret.
Once we have the topic clusters, we have to know which song is affiliated to which topic, and to know the weights associated to each cluster. Thus, we can associate for each song a score for each semantic meaning of its lyrics. To do so, we use the original term-document matrix and select only the terms that appeared in the matrix H. The we multiply the resulting matrix by the transposed matrix H (See Fig.8). Hence, for each cluster and each document, each term belonging to the document is multiplied by the weight of the term in the cluster and the sum over these products is regarded as the weight of the cluster for the document.

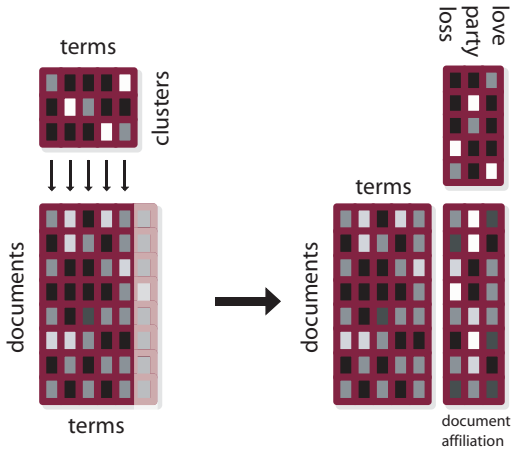

Figure 8: Computation of the songs' affiliation to the clusters (taken from [25])

This technique allows us to define for each song some score associated to keywords in a topic cluster. In order to evaluate this technique, we can for instance perform correlation between these score and the experimental study, as done in the previous chapter.

4.6 CONCLUSION & OUTLOOK

In this chapter we studied two ways for attributing automatically semantic keywords to song lyrics.
First, we studied supervised machine learning algorithms in order to classify text into several emotion classes. We used two different emotion-annotated dataset for training classifiers. We noted that the classifier that achieve the best score is an SVM with a radial basis function. When dealing with text data we extract term frequency count, which is the most common feature for NLP tasks and allows to transform the text in a vector space model. We noticed that the accuracy of a classifier increase with the number of features (which are words). Also, using n-grams at a higher rank allows to catch more semantically relevant sequence of words and thus lead to better classification too. We

considered the recently developed models Word2Vec, that allows to learn word vector which catch more meaningful information and are able to capture the context of a given word. We didn't use this powerful tool at its maximum : we took a pre-trained Word2vec model that was trained on a dataset of news. A better thing would have been to train ourself the Word2Vec model with a very large lyrics dataset, which we unfortunately did not have. A model trained on lyrics would have caught more meaningful relationships and contextual information between words in a lyrical context, which is really different from news text.

The best score we obtained is 50,2% of classification accuracy. Thus, we unfortunately didn't consider the model good enough to classify our lyrics and attribute each song an emotion class to it.

Secondly, we studied an unsupervised approach. We apply the non-negative matrix factorization in a topic modeling context. We extracted a term-document matrix from our lyrics and applied a tf-idf weighting schema, then we performed the NMF in order to get term clusters that corresponds to song topics. With this approach we thus can have score for different semantic meaning for each songs. These topics can be, in some case, connected to specific emotions manually or with the help of lexicon that maps word to emotional valence and arousal values.

5 KEYWORD-SPOTTING APPROACH

5.1 PURPOSE

In this section, we use the keyword-spotting approach to study the emotional content of our lyrics. We use the JEmAs framework that allows us to compute emotional score from our lyrics dataset. But, instead of computing these score for emotion categories (as done in chapter 3), this framework employs the dimensional model of emotion. Thus, our goal is to place each song in the valence-arousal-dominance space. In order to perform correlations with the ground-truth data we gathered from the study, we aim to calculate scores based on the distance of a song in the valence-arousal space and a specific GMBI item. The coordinates of each GMBI item would be given by a lexicon.

5.2 JEMAS FRAMEWORK

JEmAS is an open source tool developped by Sven Buechel and Udo Hahn [7]. It is measuring the emotional content of a textual document of arbitrary length. It employs a simple bag-of-words and lexicon-based approach and follows the psychological Valence-Arousal-Dominance model of emotion so that an emotion will be represented as three-dimensional vector of numerical values. The elements of this emotion vector refer to Valence (the degree of pleasantness or unpleasantness of an emotion), Arousal (degree of calmness or excitement), and Dominance (the degree of perceived control ranging from submissive to dominant).
This framework is based on a lexicon developed by Warriner et al. [44] which is a replica and an extension of ANEW lexicon. This lexicon was developed in a crowdsourcing campaign. It contains more than ten times the entries of ANEW with 13915 words associated to valence, arousal and dominance values.
The JemAs frameworks employs common keyword-spotting techniques (as described in Chapter 1) in order to determine the VAD values of a document.

5.3 PROCEDURE

We used the JEmAS framework to compute the Valence, Arousal, and dominance scores for each of our songs. Thus, we can represent our lyrics in the VAD space (See fig.8).

From the lexicon provided by Warriner et al. , we can also place the 51 GMBI items in the VAD space. We thus can get scores for each GMBI items by simply calculate an euclidean distance between the points in this space.
With these scores we can perform correlations with the data from the study in which participants rated each song a score from 1 to 6 for each GMBI item.

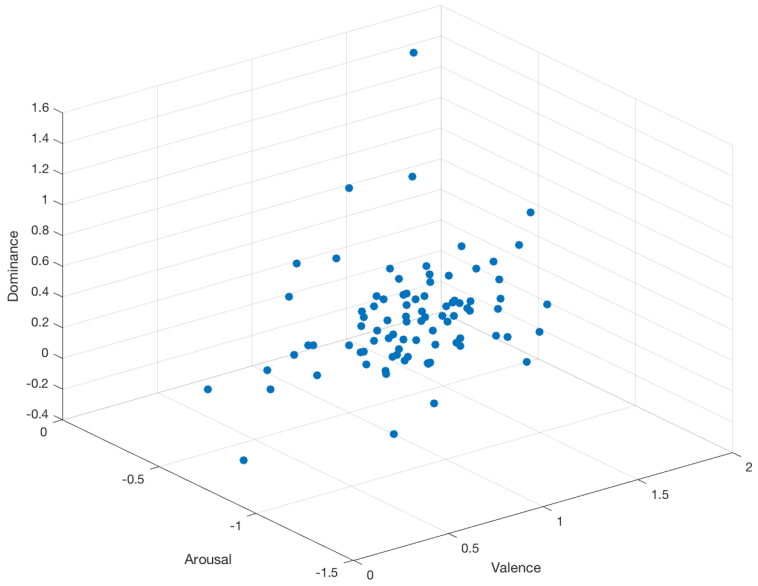

Figure 9: Scatter plot of the song lyrics in the VAD space

5.4 RESULTS & CONCLUSION

Again, in order to evaluate and validate this method, we correlated the distance score of the GMBI items with the ratings given by the study. We show in figure 10 a part of the correlation results.

	stimulating	modern	solid	fresh	adventurous	familiar
simple	0,163	0,164	-0,071	-0,009	0,155	0,007
pure	,215*	,233*	-,216*	-0,17	,211*	-0,143
unique	,245*	,299**	-,302**	-,239*	,257*	-,239*
reflective	0,172	0,195	-0,201	-0,17	0,174	-0,154
intellectual	0,167	0,178	-0,194	-0,147	0,175	-0,144
modern	-0,165	-,212*	0,197	0,137	-0,175	0,145
classic	,241*	,290**	-,282**	-,213*	,249*	-0,207
young	-0,133	-0,175	,220*	0,188	-0,138	0,18
innovative	0,057	0,087	-0,089	-0,086	0,063	-0,088
solid	0,187	,244*	-,217*	-0,165	0,195	-0,158
fresh	-0,026	-0,009	0,08	0,053	-0,037	0,066
inviting	0,088	0,123	-0,097	-0,09	0,086	-0,081
integrating	0,142	0,163	-0,111	-0,085	0,137	-0,063

Figure 10: Correlation scores from the experiment and from our method

These results shows some significant correlations (one star at a 0,05 treshold, two stars at a 0.01 treshold). However, the results are not as expected : we should see the best correlation for the same items, or this is not the case. The results we have here are difficult to interpret so we cannot validate our method

for predicting the results from the study. This can be due to the lexicon which have non accurate values for the GMBI items. It also can be due to the distance we choose as a measure for the score. Once again, these poor results can be explained by the high subjectivity we face in this study. Indeed, we always took the experiments score for each GMBI items but these are relative from a person to each other.

6

CONCLUSION & OUTLOOK

During this internship, we studied several common techniques found in the state of art for text analysis. Our goal was to extract the semantic content of music lyrics, in the context of a development of a music recommendation system based on emotions. First of all, we can say that emotion recognition is a very challenging task because it is highly subjective. In our study, the mood adjectives that we had to automatically detect in the song lyrics were even more difficult to deal with because they were elaborated in an audio branding context. Despite this, we tried to get as much results as we can in three different ways.

First, with the study of two APIs Synesketch and IBM Tone Analyzer. These APIs helped us to compute emotion score for different categories. We correlated these results with the empirical study we had in order to evaluate them. We found out that there is some correlation that we can easily explain. That means there is indeed emotional content conveyed by the lyrics when people are listening to a song, and not only by the audio content.

Secondly, we studied machine learning algorithms, both supervised and unsupervised. We use two different emotion-annotated datasets in order to perform classification task on text data. By trying several classifiers and several sets of features we found out what was the best configuration to tackle the problem. In this respect our goal was to build a sufficient robust model with random text data and then applying it to lyric data. But our maximum score never exceeded 50%, which is still a relatively good score given the lack of data we had. We also study the NMF algorithm applied to a term-document matrix in order to detect automatically song topics. This can be very useful for further research in lyrics emotion detection.

Finally, we exploited another framework, that uses the dimensional approach for emotion. In this last study, we tried to manually build scores from Valence-Arousal-Dominance results and correlates them with our study.

7

APPENDIX A - GMBI LIST AND FACTOR LOADINGS FROM FCTOR ANALYSIS

Item / Factor	Easy-Going	Joyful	Authentic	Progressive
confident	0.141	0.481	0.486	0.202
loving	**0.647**	0.312	0.346	0.057
friendly	0.483	**0.608**	0.248	0.063
honest	0.412	0.370	**0.549**	0.060
trustworthy	0.475	0.361	**0.517**	0.109
happy	0.197	**0.750**	0.161	0.137
beautiful	**0.570**	0.363	0.454	0.123
soft	**0.798**	0.100	0.173	0.053
warm	**0.632**	0.407	0.323	0.008
bright	0.323	**0.530**	0.330	0.203
stimulating	0.212	**0.551**	0.449	0.270
relaxing	**0.783**	0.126	0.258	0.074
chilled	**0.657**	0.122	0.186	0.174
detailed	0.293	0.281	**0.582**	0.238
simple	0.386	0.197	0.098	0.072
pure	0.497	0.282	**0.511**	0.108
unique	0.202	0.273	**0.561**	0.280
reflective	**0.506**	0.116	**0.516**	0.213
intellectual	0.373	0.099	**0.596**	0.239
modern	0.149	0.242	0.049	**0.770**
classic	0.359	0.080	**0.547**	-0.177
young	0.126	0.377	0.017	**0.664**
innovative	0.200	0.280	0.431	**0.544**
solid	0.298	0.327	**0.548**	0.150
fresh	0.273	**0.543**	0.279	0.397
inviting	0.435	**0.555**	0.397	0.176
integrating	0.352	0.406	0.473	0.225
adventurous	0.038	0.424	0.485	0.370
familiar	0.397	0.351	0.416	0.042
serious	0.261	-0.071	**0.564**	0.174
playful	0.152	**0.601**	0.202	0.213
funny	0.099	**0.511**	0.218	0.258
male	-0.085	0.140	0.333	0.109
female	0.382	0.186	0.064	0.150
passionate	0.297	0.420	**0.534**	0.101
sexy	0.322	0.404	0.307	0.306
epic	0.245	0.163	**0.597**	0.258
personal	0.412	0.266	**0.520**	0.161
inspiring	0.398	0.398	**0.545**	0.241
creative	0.214	0.410	**0.506**	0.340
magical	0.426	0.264	0.496	0.267
exciting	0.113	**0.511**	**0.502**	0.312
futuristic	0.076	0.048	0.176	**0.705**
retro	0.164	0.173	0.375	-0.108
timeless	0.400	0.287	**0.541**	-0.008
contemporary	0.268	0.243	0.210	**0.542**
urban	0.067	0.214	0.183	**0.517**
natural	**0.521**	0.353	0.435	0.007
authentic	0.288	0.406	**0.571**	0.074
glamorous	0.381	0.265	0.421	0.253
cool	0.222	0.462	0.355	0.420

BIBLIOGRAPHY

[1] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. "Emotions from text: machine learning for text-based emotion prediction." In: *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics. 2005, pp. 579–586.

[2] Ebba Cecilia Ovesdotter Alm. *Affect in text and speech*. University of Illinois at Urbana-Champaign, 2008.

[3] Rakesh C Balabantaray, Mudasir Mohammad, and Nibha Sharma. "Multi-class twitter emotion classification: A new approach." In: *International Journal of Applied Information Systems* 4.1 (2012), pp. 48–53.

[4] Emmanuel Bigand, Sandrine Vieillard, François Madurell, Jeremy Marozeau, and A Dacquet. "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts." In: *Cognition & Emotion* 19.8 (2005), pp. 1113–1139.

[5] Haji Binali, Chen Wu, and Vidyasagar Potdar. "Computational approaches for emotion detection in text." In: *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on*. IEEE. 2010, pp. 172–177.

[6] Margaret M Bradley and Peter J Lang. "Affective norms for English words (ANEW): Instruction manual and affective ratings." In: (1999).

[7] Sven Buechel and Udo Hahn. "Emotion Analysis as a Regression Problem-Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation." In: *ECAI*. 2016, pp. 1114–1122.

[8] Michel Cabanac. "What is emotion?" In: 60 (Dec. 2002), pp. 69–83.

[9] Corinna Cortes and Vladimir Vapnik. "Support-vector networks." In: *Machine learning* 20.3 (1995), pp. 273–297.

[10] David R Cox. "The regression analysis of binary sequences." In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1958), pp. 215–242.

[11] W Jay Dowling. "The development of music perception and cognition." In: *Foundations of cognitive psychology: Core readings* (2002), pp. 481–502.

[12] Paul Ekman. "An argument for basic emotions." In: *Cognition & emotion* 6.3-4 (1992), pp. 169–200.

[13] Paul Ekman and Wallace V Friesen. "Unmasking the face: A guide to recognizing emotions from facial clues." In: (2003).

[14] Beverley Fehr and James A Russell. "Concept of emotion viewed from a prototype perspective." In: *Journal of experimental psychology: General* 113.3 (1984), p. 464.

[15] Alf Gabrielsson and Erik Lindström. "The influence of musical structure on emotional expression." In: (2001).

[16] Maria Giatsoglou, Manolis G Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch Chatzisavvas. "Sentiment analysis leveraging emotions and word embeddings." In: *Expert Systems with Applications* 69 (2017), pp. 214–224.

[17] Ruchi Hirat and Namita Mittal. "A Survey On Emotion Detection Techniques using Text in Blogposts." In: *International Bulletin of Mathematical Research* 2.1 (2015), pp. 180–187.

[18] Tin Kam Ho. "Random decision forests." In: *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. Vol. 1. IEEE. 1995, pp. 278–282.

[19] Xiao Hu. *Improving music mood classification using lyrics, audio and social tags*. University of Illinois at Urbana-Champaign, 2010.

[20] Xiao Hu and J Stephen Downie. "Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata." In: (2007).

[21] Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. "Emotion analysis of songs based on lyrical and audio features." In: *arXiv preprint arXiv:1506.05012* (2015).

[22] Patrik N Juslin and Petri Laukka. "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening." In: *Journal of New Music Research* 33.3 (2004), pp. 217–238.

[23] Patrik N Juslin and John A Sloboda. "Music and emotion: Theory and research." In: (2001).

[24] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. "Music emotion recognition: A state of the art review." In: ().

[25] Florian Kleedorfer, Peter Knees, and Tim Pohle. "Oh Oh Oh Whoah! Towards Automatic Topic Detection In Song Lyrics." In: *Ismir*. 2008, pp. 287–292.

[26] Uros Krcadinac, Philippe Pasquier, Jelena Jovanovic, and Vladan Devedzic. "Synesketch: An open source library for sentence-based emotion recognition." In: *IEEE Transactions on Affective Computing* 4.3 (2013), pp. 312–325.

[27] Gunter Kreutz, Ulrich Ott, Daniel Teichmann, Patrick Osawa, and Dieter Vaitl. "Using music to induce emotions: Influences of musical preference and absorption." In: *Psychology of Music* 36.1 (2008), pp. 101–126.

[28] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. "Multimodal music mood classification using audio and lyrics." In: *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*. IEEE. 2008, pp. 688–693.

[29] Cyril Laurier, Perfecto Herrera, M Mandel, and D Ellis. "Audio music mood classification using support vector machine." In: *MIREX task on Audio Mood Classification* (2007), pp. 2–4.

[30] Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. "Music Mood Representations from Social Tags." In: (2009), pp. 381–386.

[31] Ricardo Malheiro, Renato Panda, Paulo Gomes, and R Paiva. "Music emotion recognition from lyrics: A comparative study." In: 6th International Workshop on Machine Learning et al. 2013.

[32] Rudolf Mayer, Robert Neumayer, and Andreas Rauber. "Rhyme and Style Features for Musical Genre Classification by Song Lyrics." In: *Ismir*. 2008, pp. 337–342.

[33] Josh H. McDermott, Alan F. Schultz, Eduardo A. Undurraga, and Ricardo A. Godoy. "Indifference to dissonance in native Amazonians reveals cultural variation in music perception." In: *Nature* 535.7613 (July 2016), pp. 547–550.

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[35] George A Miller. "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation." In: *EMNLP*. Vol. 14. 2014, pp. 1532–1543.

[37] Robert Plutchik. "A general psychoevolutionary theory of emotion." In: *Theories of emotion* 1.3-31 (1980), p. 4.

[38] Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. "EmpaTweet: Annotating and Detecting Emotions on Twitter." In:

[39] JA Russell. "A circumspect model of affect, 1980." In: *J Psychol Soc Psychol* 39.6 (1980), p. 1161.

[40] Klaus R Scherer and Marcel R Zentner. "Emotional effects of music: Production rules." In: ().

[41] Carlo Strapparava and Rada Mihalcea. "Learning to identify emotions in text." In: (2008), pp. 1556–1560.

[42] Carlo Strapparava, Alessandro Valitutti, et al. "WordNet Affect: an Affective Extension of WordNet." In: 4 (2004), pp. 1083–1086.

[43] Robert E Thayer. "The biopsychology of mood and arousal." In: (1990).

[44] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. "Norms of valence, arousal, and dominance for 13,915 English lemmas." In: *Behavior research methods* 45.4 (2013), pp. 1191–1207.

[45] Yi-Hsuan Yang and Homer H. Chen. "Machine Recognition of Music Emotion: A Review." In: *ACM Trans. Intell. Syst. Technol.* 3.3 (May 2012), 40:1–40:30.

[46] Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen. "Toward multi-modal music emotion classification." In: (2008), pp. 70–79.