Utrecht University

MSc. Thesis in
Cognitive Artificial Intelligence

Predicting Relevance of Emotion Tags

Author:
Jeroen Witteveen

Supervisors:
Dr. Frans Wiering
Dr. Anja Volk

December 2015
45 ECTS

Abstract

Emotion tags form an important tool for searching and categorizing large music collections such as online music services, e.g. Last.fm, or production music databases. However, the tags are often not provided or evaluated by experts, resulting in noisy and less useful tag sets. The main goal of this research is to provide a method for automatically evaluating the relevance of those tags. The method consists of using the distance between emotion tag and audio clip, both plotted in arousal/valence (AV) space, as predictor for tag relevance. To this end, first, AV prediction regression models for audio clips are trained and tested with cross validation on three different datasets. Results on the train/test sets are matching state of the art R^2 scores, however, the results deteriorate when the models are validated on other datasets, especially for valence prediction models. Therefore, not the predicted but the human rated AV values of the clips are used in the next step. Second, relevance of four emotion tags (angry, happy, sad and tension) and one set of ten tags together are predicted for audio clips from two datasets with regression models that are trained and tested using two different predictors separately: (1) distance between AV values of the clip and emotion words, and (2) directly with audio features. Except for 'angry', AV distance predictors outperform audio feature predictors (e.g. for 'happy' respectively $R^2 > .65$ vs. $R^2 > .18$). A second evaluation on a self composed dataset with a larger set of different emotion tags did not lead to useful results. Whether the method generalizes to other tags is therefore still inconclusive. These findings (1) indicate that AV prediction of music is still in development phase, and (2) point into the direction of a new promising method for evaluating at least a few emotion tags.

Preface

This thesis is the final work for completing the master Cognitive Artificial Intelligence (CAI). With a bachelor in philosophy and no bèta background, enrolling in the CAI master provided some challenges, but also gave me a great opportunity to further explore my interests in psychology and computer science. These research fields came perfectly together with my interest in music in the last course I followed: 'Sound and Music Technology' where I learned about the field of Music Emotion Recognition (MER).

This subject piqued my interest so I contacted Frans Wiering, one of the lecturers, to discuss the possibilities of doing my thesis in the field of MER. He introduced me to a 'real-world' problem faced by a music company selling production music, and from that moment their problem also became mine for the following year. Fortunately, I thoroughly enjoyed working on it, not in the least because it was a real and practical problem which gave it a great sense of purpose.

I would like to thank Frans Wiering and Anna Aljanaki for their inspirational ideas and dedicated support throughout the process of writing this thesis. Especially Anna for helping me with all the technical parts of this project, providing thorough feedback on various drafts and for trying to let me write less (sorry, I really tried). Finally, I would like to thank Anja Volk for her comments on the second last version of this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Digital music revenues grew from $400 million in 2004 up to $5.9 billion in 2013, constituting 39% of the whole music industry income.[1] Most part of the digital music revenues is still generated by paid downloads, but subscription-based and ad-supported streaming service revenues are rapidly catching up.[2] For example, in 2010 there were eight million paid accounts to subscription services, which rose to 28 million in 2013.[1]

Examples of subscription-based and ad-supported streaming services are Pandora, Spotify and Last.fm.[3] Pandora is an online radio company that streams music with automated recommendations. It had 200 million registered users in the US, Australia and New Zealand in December 2013.[4] Spotify is an online music streaming service with 10 million paying users in May 2014.[5] Last.fm is an online music recommendation system with over 57 million user accounts in 2014, and with 45 million unique tracks of which 20 million are playable.[6]

Together with the growing music industry, digital music collections also grow bigger and bigger due to increasing storage capacity, the use of internet and compacter audio formats (Wieczorkowska et al., 2006; Yang and Chen, 2012). Such collections are used by many for personal, commercial or professional purposes (Casey et al., 2008).

1.1 General problem description

With growing music collections, a proper organization of the data becomes more and more important to maintain accessibility. Various types of information could be used for this purpose: (1) metadata describing the music (e.g. genre, mood, lyrics, theme), (2) content data such as tempo, tonality or chords, and (3) user data such as ratings,

[1] `www.ifpi.org/facts-and-stats.php` 2–12–2014
[2] `www.statista.com/chart/2028/digital-music-revenue/` 2–12–2014
[3] `www.pandora.com`, `www.spotify.com`, `www.last.fm/`
[4] `www.statista.com/topics/1349/pandora/` 2–12–2014
[5] `www.statista.com/statistics/244995/number-of-paying-spotify-subscribers/` 2–12–2014
[6] `www.skilledtests.com/wiki/Last.fm_statistics`

listening statistics, age or gender. This data could be used for various purposes such as indexing, retrieval and recommendation.

In the current research we focus on metadata, specifically emotion[7] tags. Emotion tags are words such as 'angry', 'happy', 'sad' that are relevant to the music. These tags are usually human generated by e.g. online users of social website such as `last.fm` or by owners of (private) collections.

However, in the case of social websites, it becomes increasingly difficult to maintain high quality tags. The music collections grow and data is added by different people who all have their own associations with the music and have a different way of encoding them to tags (Casey et al., 2008).

A similar problem is faced by `allmusic.nl`, a website selling production music. Producers that supply music add a great variety tags, including unrelated tags, to their music to increase findability of their music and thereby decreasing the quality of the tags, e.g. tags that are applied to all items are less informative, or when there are many tags with conflicting meaning applied to the same track ('happy' and 'sad', 'metal' and 'jazz').

In both examples, the low quality of emotion tags decrease the accessibility of the music collections. Improving the quality of tags would directly improve indexing and retrieval of the data. Indirectly it could also improve recommendation systems.

1.2 Research objective

The main objective of this research is to develop and test a new method for automatically predicting weights of emotion tags that are assigned to music. A tag weight indicates how relevant the tag is to the clip, i.e. how well the tag describes the clip. The method could be used to improve quality of tags applied by humans or for auto-tagging. It would thereby improve the accessibility, using emotion words, of large music collections.

Secondary objectives are to build an arousal/valence (AV) prediction model and collect human data to evaluate the tag weight prediction method.

1.3 Method

Various studies in the field of Music Information Retrieval suggest that combining information from tags and audio signal enhances system performances compared to systems that use only one source of information (Bischoff et al., 2009; Wang et al., 2010; Nanopoulos and Karydis, 2011).

[7] For a discussion on why the term 'emotion', instead of the related term 'mood', is used and what it refers to, please read chapter 2 section 2.2.1.2.

In the current research a new method is proposed that combines three sources of information to predict emotion tag weights: AV ratings of emotion words, AV predictions of audio and psychological models relating emotions to AV space. Using these three sources has recently been done by (Saari et al., 2013b), however that method only aimed at improving emotion prediction systems whereas our primary goal is to improve the quality of noisy tag sets.

The main idea behind the presented method is that if the AV values of an emotion tag and the AV values of an audio clip are not similar, that the tag is probably not relevant to the clip and vice versa. In other words, the relation between the AV values of the tag and the clip determine the tag weight (i.e. a value that indicates the relevance) of the tag.

The following main question is formulated: *Can tag weights be predicted based on arousal and valence information of the tag and clip?* The tag weights can be used to clean up noisy tag sets by removing less relevant tags or for auto-tagging.

Figure 1.1 shows the general framework of the tag weight prediction method and how the chapters are connected together. AV prediction regression models for audio clips are trained and tested with extensive grid searches and cross validation on multiple datasets (chapter 3). Because the trained models are not reliable enough in predicting AV values, human AV ratings of the clips are used in the next steps. AV values of the clips are combined with AV rated emotion words, which are collected in chapter 4, to create predictors for the tag weighting method that is described in chapter 5. Tags rated by humans for their relevance to music clips are used as ground truth to train and test the tag weighting regression models. The models are trained and tested on four different tags separately and one set of ten tags. In addition, the models are evaluated on two different datasets to assess the robustness of the models. The best tag weight prediction model is evaluated again on a dataset with self collected data in chapter 6. This second evaluation is different from the evaluation in chapter 5 in that it contains 79 different emotion tags instead of maximum 10 different emotion tags.

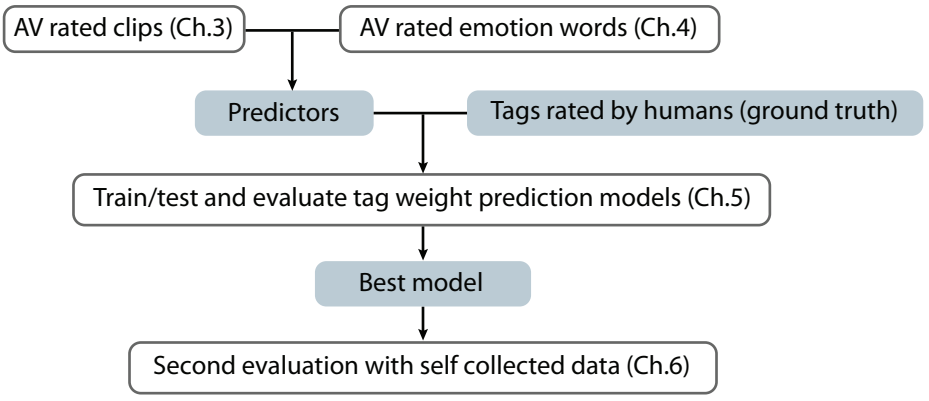


Figure 1.1 Overview of the main chapters and the method.

1.4 Thesis outline

The logic of the core chapters of this thesis are already explained in the previous section and visualized in figure 1.1. In short those chapters contain: AV prediction models (chapter 3), assembly of a dataset with AV rated emotion words (chapter 4), implementation and evaluation of the tag weight prediction method (chapter 5) and a second evaluation of the method (chapter 6). In addition, background information about emotion theories and emotion research related to this thesis is presented in chapter 2. The thesis is concluded by chapter 7 where the main findings are presented and directions for further research are suggested.

Chapter 2

Background

This chapter covers the necessary background knowledge for understanding the concepts and methods used in this thesis and thereby gives an overview of the state of the art in this research field. First, in section 2.1, the psychological background is covered containing definitions and emotion models. Second, in section 2.2, the computational background related to Music Emotion Recognition (MER) is covered. Third, section 2.3 summarizes other related work. And finally, important observations based on discussed literature are highlighted in section 2.4.

2.1 Emotion

Emotions are undoubtedly a very important aspect of our life. Already the old Greek philosophers mentioned emotions in their theories. Aristotle discusses in Poetics that a tragedy (a dramatic play) arouses certain emotions like pity and fear which should accomplish a purification (catharsis) of such emotions in the viewer (Aristotle, 1962). The Stoics (c.300 B.C.E.) argued that impulsive emotions (passions) are false judgements and strived towards a life in which they were not affected by them and in which there were only positive emotions such as joy, caution and reasonable wishing (Rubarth, 2005). Emotions are also important motives for actions, e.g. the following examples in the literature: how would the war in Homer's Iliad have unfolded if Achilles didn't resent Agamemnon for denying his war trophy? And how would Shakespeare's tragic play Romeo and Juliet have ended if Romeo was raised by Stoics?

2.1.1 Terminology: emotion, mood, feeling

Emotion is a concept with many faces, and to make it even worse, other concepts like 'mood', 'feeling', 'affect' and 'temperament' are closely related and used interchangeably (Ketai, 1975; Juslin and Sloboda, 2010). A very useful distinction between those concepts is the timescale in which they occur. Emotions are described as short lived phenomena

(Levenson, 1994) with a duration ranging from a few seconds to a few minutes (Watson, 2000) while mood is described as a background feeling that persists over a longer period of time (Thayer, 1996) lasting for several hours to a few days (Watson, 2000). Whereas mood and emotions are both states, temperament is a trait (a quality of someone's character) which indicates a difference in tendency between individuals for experiencing emotion or mood states (Watson, 2000). The duration of temperament could be a lifetime or change over a period of years (Oatley and Jenkins, 1996).

Another difference between emotion and mood is that moods are often, but not always (e.g. heavy depression), less intensive than emotions. Furthermore, emotions usually have a more obvious cause than moods, which are often caused by a variety of inputs (Thayer, 1996; Grey and Watson, 2001).

Feeling and affect are both different from the others. Feelings are an aspect of emotion and mood, and could be described as the subjective or qualitative experience of a mood or an emotion. Affect, on the other hand, is a broad term that could be used to describe phenomena like mood and emotion (Blagov and Singer, 2004; Spindler, 2009).

These descriptions are merely meant to give an example of how the various concepts around 'emotion' are related and how they could be and are used in the literature. Other uses and definitions are possible and could be correct even if they differ substantially from the above.

2.1.2 Theories of emotion

The first psychological theory of emotion was created by William James. He argues that emotions are the feeling of our bodily changes that follow the perception of something exciting (James, 1884), which is also known as the James-Lange theory of emotion. For example, you see a snake, your heart rate rises which causes a scared feeling. Shiota and Kalat (2012) point out that James' theory is about the feeling aspect of emotion, which is "the perception of the body's action and physiological arousal" (Shiota and Kalat, 2012, p. 14). They argue that an emotional state commonly consists of four aspects: cognition, feeling, behavior and physiological changes (see table 2.1).

Stimulus	**Cognition**	**Feeling**	**Behavior**	**Physiological changes**
Enemy	'Danger'	Fear	Escape	Heart rate, blood pressure
Unexpected event	'What is it?'	Surprise	Stop	"
Obstacle	'Enemy'	Anger	Attack	"

Table 2.1 Four aspects of emotion states with examples. The 'stimulus' is not a part of the emotion state itself, but often included in theories of emotion (Plutchik, 2001).

However, Shiota and Kalat note that not all four of them are necessary. For example, the behavioral aspect is lacking if you feel happy, think that you are happy and show

physiological changes associated with happiness, but behave neutral. Or when an unconsciously perceived stimulus causes you to feel distressed with associated physiological changes and behavior, but without the cognitive appraisal.

The emotion theories that emerged after William James usually focus on one (or more) of the four aspects of emotion (Davies, 2010). For the purposes of this thesis, it is not necessary to explain those theories in depth, however it must be said that because these theories focus on different aspects, their definitions of emotion can be quite different and sometimes even seem contradictory.

A reason for this is, as Russel (2003) puts it, "if, instead, the word emotion refers to a heterogeneous cluster of loosely related events, patterns, and dispositions, then these diverse theories might each concern a somewhat different subset of events or different aspects of those events" (Russell, 2003). So there might not be one natural category or kind on which we try to fit the label 'emotion' (Griffiths, 2008; Shiota and Kalat, 2012) which explains why there can't be a consensus on a clear definition of emotion (Frijda, 1988).

2.1.2.1 Aboutness

Another issue in theories of emotion is whether emotions necessarily have an object onto which they are directed, also called aboutness or intentionality (see 'stimulus' column in table 2.1). For example, to be angry *at* someone or to be happy *about* something. While various authors argue that emotions are necessarily about something (Solomon, 2008), others think that it's not necessarily true (Crane, 2003, p. 37). For example, if you're suddenly feeling happy or sad without being able to say why.

2.1.2.2 Function of emotions

Darwin (1872) already noted that facial expressions of emotions are similar between animals and humans, leading to a theory that emotions have an evolutionary origin (Spindler, 2009). The idea is that emotions evolved because it helped in survival by guiding social interactions and physical fitness (Keltner and James, 1999), e.g.
trust could be gained or enemies be scared of by showing emotions. Emotions also help to decide what is good and bad for our physical fitness. Getting scared of a stronger predator, or getting happy from protein rich food are very useful mechanisms.

2.1.3 Classification and models

Another difference between emotion theories is how emotions are classified and related to each other. Two commonly used models in emotion research are discrete models (Ortony and Turner, 1990; Ekman, 1992; Izard, 2007) and dimensional models (Russell, 1980;

Thayer, 1989; Russell, 2003; Eerola et al., 2009). Both models postulate a fundamentally different theory about how emotions are related to one another.

2.1.3.1 Discrete models

The assumption underlying discrete emotion models is that there exist basic emotions that are fundamentally different from each other. Commonly three characteristics are attributed to basic emotions (Shiota and Kalat, 2012):

1. Basic emotions evolved differently from each other and have different functions.
2. Every healthy person can experience them.
3. All aspects of an emotion (see table 2.1) belong together in a consistent way.

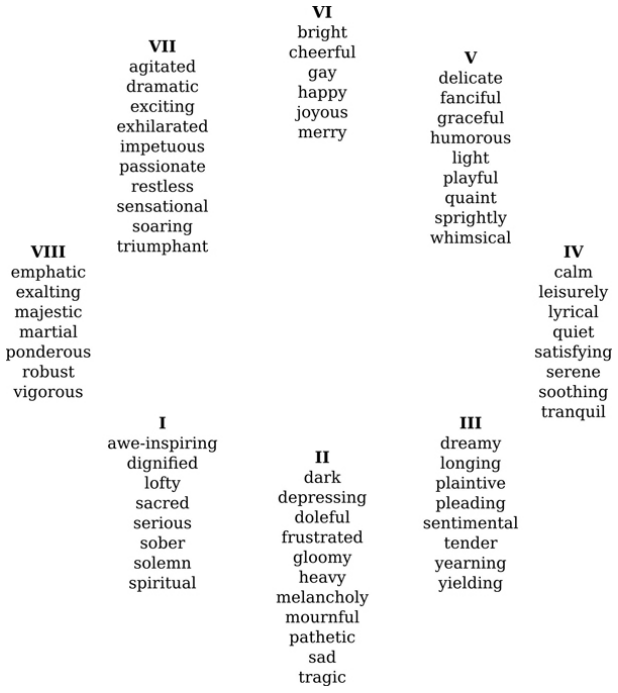


Figure 2.1 Hevner's categorical emotion model.

While various researchers agree that basic emotions exist, there is little agreement on how many and which emotions count as basic emotions, and what criteria need to be satisfied for an emotion to be counted as basic (Ortony and Turner, 1990). The most common emotions listed as 'basic' are: fear, anger, sadness, joy, love and surprise (Plutchik, 2001). However, Ekman (1992), for example, lists ten basic emotions that are distinguished from one another and from other affective phenomena based on nine characteristics. Another example is Hevner's emotion model (Hevner, 1936) with eight categories, see figure 2.1.

Some evidence in favor of discrete models comes from neurological research. For example, Gosselin et al. (2007) reports a patient with bilateral amygdala damage with

an impairment for recognizing sad and scary music, while her happy-music recognition ability functioned normal. However, research like this could only prove that there are neural areas that are necessary for processing information for specific emotions, the other aspects of an emotion (behavior, other physiological changes, feelings) are not taken into account here.

A negative feature of discrete emotion models is low amount of basic emotion classes compared to the richness of the experience (Yang and Chen, 2012), in other words, the resolution of the model is low (Barthet et al., 2013).

2.1.3.2 Dimensional models

Whereas in discrete models all aspects of emotion are taken into account, dimensional models are based on only one aspect of emotion, namely the feeling aspect (Shiota and Kalat, 2012), or as how Russell (2003) calls it, the core affect. All emotions (i.e. feelings) can be described by a few dimensions and are not viewed as individually evolved entities, but as psychological and social constructs (Shiota and Kalat, 2012). Commonly there are two or three dimensions distinguished (Schimmack and Grob, 2000) (see fig. 2.2) which have sometimes different names in the literature denoting approximately the same thing (Spindler, 2009; Yang and Chen, 2012):

- **Valence** or pleasure, pleasantness, evaluation
- **Arousal** or energy, activation, activity
- **Dominance** or attention, potency, tension

However, there are also other dimensions proposed. For example Thayer (1989) proposes a model with two arousal dimensions: tension arousal and energetic arousal. Yet this model is argued to be a 45 degree rotation of a model with arousal/valence dimensions (Eerola and Vuoskoski, 2010). So far there is no consensus on which and how many dimensions there are and what their physiological basis is (Schimmack and Grob, 2000).

A well known dimensional model is the circumplex model of affect by Russell (1980) (see fig. 2.2). In this model, emotion labels are placed on a 2D space based on arousal and valence [1] (AV) ratings acquired by grouping experiments in which similar emotions where grouped together, and from introspective experiments in which subjects had to rate their current feeling on AV scales and for 518 affective adjectives (Russell, 1980). Emotions that are close together indicate a similar feeling and emotions that are opposite of each other indicate opposite feelings. Results from both experiments were mapped on a 2D AV space and showed a consistent model. These findings indicate that the cognitive structure of the feeling aspect of emotions is similar to their semantic structure.

[1] Russell actually uses the term 'pleasantness', which is similar to 'valence' in this context.

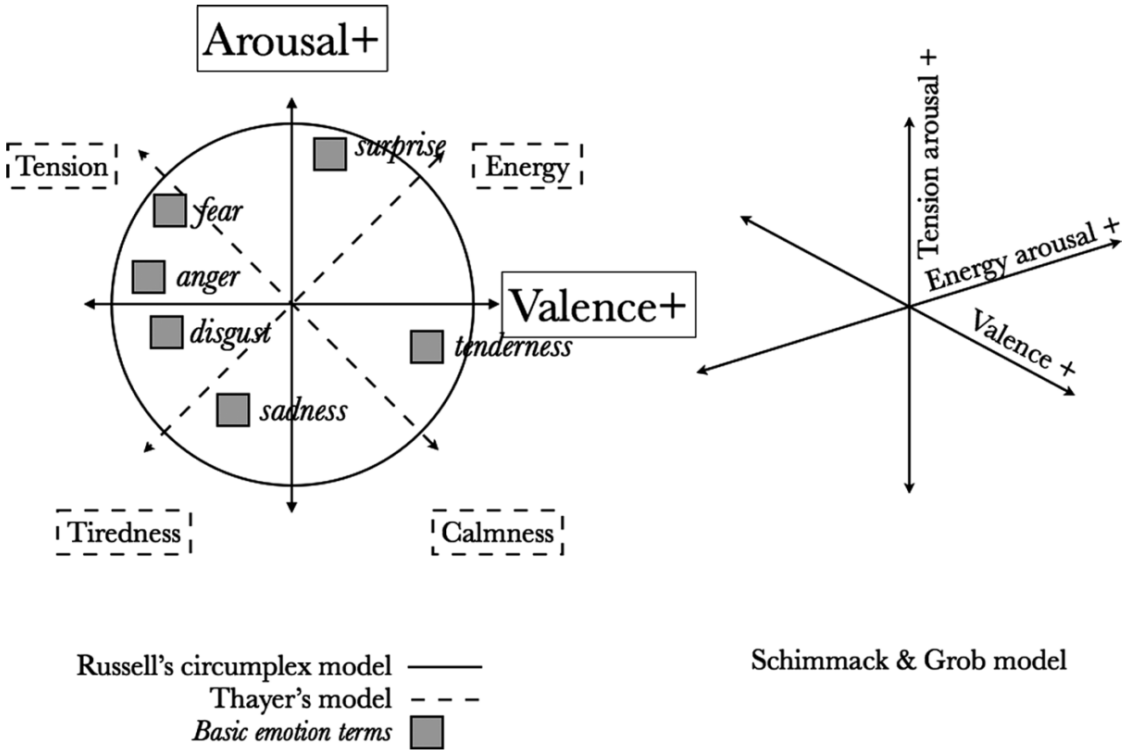


Figure 2.2 Dimensional emotion models, (adapted from Eerola and Vuoskoski (2010))

One of the drawbacks of a AV model is that some emotions (e.g. fear and anger) are mapped very close to each other because of similar AV ratings (Yang and Chen, 2012), indicating that they are similar emotions, while they actually are quite different. An explanation for the lack of variation in the AV model is that the dimensional models primarily focus on the feeling aspect and that e.g. fear and anger have a similar feeling in terms of arousal and valence, while other aspects of emotion like behavior and cognition are able to describe the difference. Thus two dimensions are simply not sufficient to account for all the variance. Some studies show that 3 dimensions are able to account for a larger part of the variance and that the arousal dimension would be more effective if it was split into two dimensions (Schimmack and Grob, 2000; Schimmack and Reisenzein, 2002). Other studies however report that a 2D model could explain the same amount of variance as a 3D model (Eerola et al., 2009; Warriner et al., 2013). The reason for these inconsistent findings is that the latter studies only take the dimensions valence, arousal and dominance into account whereby they report that dominance and valence are strongly correlated. The studies by Schimmack & Grob and Schimmack & Reisenzein however compare also other dimensions: awake-sleepiness and tension-relaxation by which more variance could be explained.

Another problem for dimensional AV models are mixed emotions, which cannot be mapped on the dimensional space. For example, Hunter and Schellenberg (2010) describes two experiments in which music was rated on two separate unipolar scales, one

for sad and another for happy. Some music was rated high on both scales indicating mixed emotions.

2.1.3.3 Comparison

Aspects of emotion Discrete models take all aspects of an emotion into account, whereas dimensional models primarily focus on the feeling aspect. Based on this difference it could be argued that both models are actually noncompetitive since they strictly deal with a different phenomena.

Which and how many? There is no consensus about which emotions are basic in discrete models, but there is also no consensus about which and how many dimensions are optimal.

Focus Discrete models focus mainly on distinctive characteristics between emotions while dimensional models focus on identifying the position of emotions on emotional dimensions (Yang and Chen, 2012).

Genesis A fundamental difference is that basic emotions are assumed to have evolved distinctly while emotions in a dimensional model are assumed to be psychological and social constructs.

Neural process Studies with brain damaged patients show evidence for neural processes underlying both discrete and dimensional theories (Eerola and Vuoskoski, 2010). It could therefore be possible that the underlying biology works in both a discrete and dimensional fashion together.

Abstractness Dimensional models are more abstract (Böck, 2013) because they use characteristics (the dimensions) to categorize emotions, while discrete models directly name emotion itself.

Granularity The discrete model has a much poorer resolution for describing borderline emotions (Yang and Chen, 2012; Barthet et al., 2013).

Ambiguous emotions Some emotions can be ambiguous such as 'mixed' emotions. Both models can't deal with those emotions easily since only one point in the dimensional space can be selected or one category in discrete models.

Comprehensibility Discrete emotion models are easier to understand then dimensional emotion models for categorizing or rating emotions by subjects (Beale and Peter, 2008; Eerola and Vuoskoski, 2010), however the dimensional models are considered to be more accurate for self-assessments (Lichtenstein et al., 2008; Beale and Peter, 2008).

MER In MER the advantage of using a categorical model is that it is easy to incorporate the results in text and metadata based systems. However, as noted earlier, categorical models consist of a small number of basic emotions relative to the full emotional spectrum that humans have. Introducing more (basic) emotions into the model is not a good solution since it allows more ambiguity into the model due to the different conceptualizations of the emotions among people (Yang and Chen, 2012).

These comparisons are not meant to give decisive arguments for one or the other model, but should give an idea about the controversies around discrete and dimensional emotion classification models.

2.1.4 Research methods

There are various methods that could be employed in emotion research, see table 2.2. Each method has its own advantages and disadvantages (Shiota and Kalat, 2012), however only the self-report is discussed since that is the method that is relevant in this research. One general problem with emotion research is the subjective character of the matter. Because of this it is already difficult to exactly define the concept 'emotion', which makes it even harder to do research on it.

Response system	**Measure**
Subjective Experience	Self-report
Physiological measurement	Blood pressure, heart rate, sweating
	Brain activity (EEG, PET, fMRI)
	Hormones
Behaviors	Actions
	Facial and vocal expressions

Table 2.2 Methods of measurement in emotion researach (Mauss and Robinson, 2009; Shiota and Kalat, 2012)

2.1.4.1 Self-report

Self-report is a relatively good and common method (Diener et al., 1991; Robinson and Clore, 2002; Hunter and Schellenberg, 2010; Zentner and Eerola, 2010), and probably the only way (Shiota and Kalat, 2012) to really measure the feeling aspect of emotions. However, there are fields of research for which self-reports on emotion are less usable, for example when researching mental health issues (Shedler et al., 1993).

Validity problems could occur if subjects intentionally give wrong answers for any reason (Zentner and Eerola, 2010), and even if they are as honest as possible, there might be limitations in the awareness of one's emotions (Barker et al., 2002; Zentner and Eerola, 2010). And even if people are sufficiently able to know their own feelings,

the answer to a question might differ over time for the same person, different people could rate the same feeling differently, or just find it hard to verbalize musical emotions (Zentner and Eerola, 2010). These limitations don't make self-reports invalid, however the results should not always be trusted. The interpretation and analysis should be done with these limitations in mind (Barker et al., 2002).

There are roughly two kinds of self-reports: quantitative and qualitative. Typical quantitative self-reports are closed-ended questionnaires or interviews in which the subject has a limited range of predefined answers to choose from. A disadvantage is that the questions can be interpreted differently or subjects are forced to choose between options that are equally wrong or good for them. An advantage is that the results of different subjects can be easily compared. Qualitative self-reports are commonly open-ended questionnaires, (individual or group) interviews or diaries. A disadvantage is that the data must be interpreted and is not easily comparable between different subjects. An advantage is that it captures the experiences of the subject in more detail. Moreover, different interpretations of the questions by the subject can be better obviated.

2.2 Music Emotion Recognition

Music Emotion Recognition (MER) is a research field within Music Information Retrieval (MIR) devoted to the identification of emotions in music. MIR can be characterized as an "interdisciplinary research area encompassing computer science and information retrieval, musicology and music theory, audio engineering and digital signal processing, cognitive science, library science, publishing, and law. Its agenda, roughly, is to develop ways of managing collections of musical material for preservation, access, research, and other uses" (Futrelle and Downie, 2003). MIR differs from music cognition in the sense that the latter tries to model and understand the process of how humans listen to music, while MIR is more focused on modeling the result of the human processes (Aucouturier and Bigand, 2012).

In MIR different types of information (textual, digital signal, music notation) are analyzed for different types of tasks, for example: beat tracking, segmentation, emotion detection, pitch detection, audio fingerprinting, chord estimation and key detection. MER research focuses on the emotion detection task.

2.2.1 Emotion and music

How are emotions and music related? From experience it is clear that music can change the way we feel and that the way we feel can influence the way we appraise music. Furthermore, music can be used "to change emotions, to release emotions, to match their current emotion, to enjoy or comport themselves, and to release stress" (Juslin

and Västfjäll, 2008). However, it is not clear how and why exactly music is able to bring about emotions in the listener (Juslin and Västfjäll, 2008; Yang et al., 2008; Davies, 2010). It is also unknown why we like to listen to music that expresses negative emotions such as sadness (Davies, 2010).

A more fundamental question about the relation between music and emotion is whether music can express emotions. About a century ago McAlpin (1925) argued against the view that music does not express emotions and that music only expresses itself. While it is nowadays a commonly agreed fact in MER research that music expresses emotions (Kim et al., 2010), there are still some philosophical issues with this position (Davies, 2010). Without diving deep into such a discussion, we can safely assume that at least in some cases "music sounds the way emotions feel" (Pratt, 1952), whereby the problematic term 'expressing' is avoided.

2.2.1.1 Felt and perceived emotions

The questions of how and why music is able to bring about emotions in the listener and how music is able to 'express' emotions touches upon an important difference in MER: felt versus perceived emotion. *Felt emotions* are emotions that are induced in the listener by the music whereas *perceived emotions* are emotions that are thought to be expressed by music (Schubert, 2010; Sloboda and Juslin, 2010; Barthet et al., 2013). Zentner et al. (2008) measured that subjects report perceived emotions as more frequently occurring than felt emotions when listening to music. And another study reports a positive although not perfect correlation between felt and perceived emotions (Hunter and Schellenberg, 2010).

However, some emotions are more likely to be perceived than felt suggesting that there is not a one to one relation between felt and perceived emotions (Sloboda and Juslin, 2010). Since this relation is not properly understood, it seems good practice to specify whether felt or perceived emotions are measured in an experiment to avoid subjects using both, which could lead to unreliable results.

The fact that emotions raised by music (felt or perceived) are inherently subjective makes these questions difficult to answer (Panda et al., 2013). Different emotions are felt or perceived by different people for the same song, even for the same person over time the emotions can change, thereby, the words to describe that emotion are often ambiguous (Panda et al., 2013).

2.2.1.2 'Emotion' in music research

'Music emotion' refers to emotions that are somehow induced by music (Juslin and Västfjäll, 2008; Zentner and Eerola, 2010), Thereby, the concept 'emotion' in MER research is used to denote the feeling aspect of emotion and mood (Kim et al., 2010),

although, as Juslin and Västfjäll (2008) argue, feelings induced by music carry more properties of emotions: presence of an object (the music), short duration and autonomic response.

Another discrimination can be made between aesthetic emotions and utilitarian emotions (Scherer, 2004). Utilitarian emotions are the emotions that are described in section 2.1, whereas aesthetic emotions are not concerned with the relevance of a perception to bodily needs, social values, or current goals or plans. For aesthetic emotions "the appreciation of the intrinsic qualities of a piece of visual art or a piece of music is of paramount importance" (Scherer, 2004).

While 'the feeling aspect of emotions' and 'aesthetic emotions' have a considerable overlap in meaning, the latter is more specific on art whereas the former is more general. In the remainder of this research 'emotion' is used to denote the aesthetic/feeling aspect of emotions and mood and it is made explicit in ambiguous situations.

2.2.2 Common methods in MER

There are roughly two types of techniques employed in MER to retrieve information: content based (audio signal, MIDI, music notation, lyrics) and context based (metadata such as tags, reviews, social media) (Casey et al., 2008; Kim et al., 2010). Due to the scope of this research only audio signal and tag based methods are evaluated.

2.2.2.1 Audio signal analysis

The common procedure for audio analysis in MER is to compose a dataset with audio files and label them with emotion words or dimensional ratings. The dataset is then used to train and test learning algorithms (Yang and Chen, 2012). This procedure is divided into five steps and explained in more detail below.

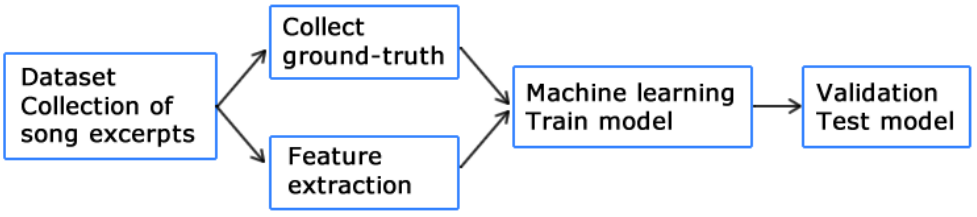


Figure 2.3 Machine learning process in five steps.

Step 1. Dataset. A dataset consisting of music files that is used for training and testing the model should be compiled (see table 2.3 for examples). Ideally a dataset is large encompassing all sorts of music to make it as general as possible. Before a dataset is used it is usually preprocessed to make the music pieces better comparable. Usually excerpts with a duration between 20–30s. are extracted from

each song based on: which part represents the song best, automatic segmentation *or* choosing a standard position (e.g. the middle or after 30s.) (Yang and Chen, 2012). Ideally an excerpt is as short as possible to avoid too much variation, but long enough for a listener to rate it sufficiently (MacDorman et al., 2007). Furthermore the audio is converted to a standard format, for example 22050Hz sampling frequency, 16–bits precision, mono channel (Yang and Chen, 2012) and normalized by sound (Mion and De Poli, 2008).

Name	**Songs**	**Length**	**Format**	**Ground truth**
EmoMusic[1]	744	45s.	mp3	2 dims.
Soundtrack110[2]	110	∼15s.	mp3	5 cats. 3 dims.
Emotify music[3]	400	60s.	mp3	9 cats.

Table 2.3 Datasets with categorical and dimensional mood rated song excerpts. [1] Soleymani et al. (2013), [2] Eerola et al. (2009) and [3] Aljanaki et al. (2014a).

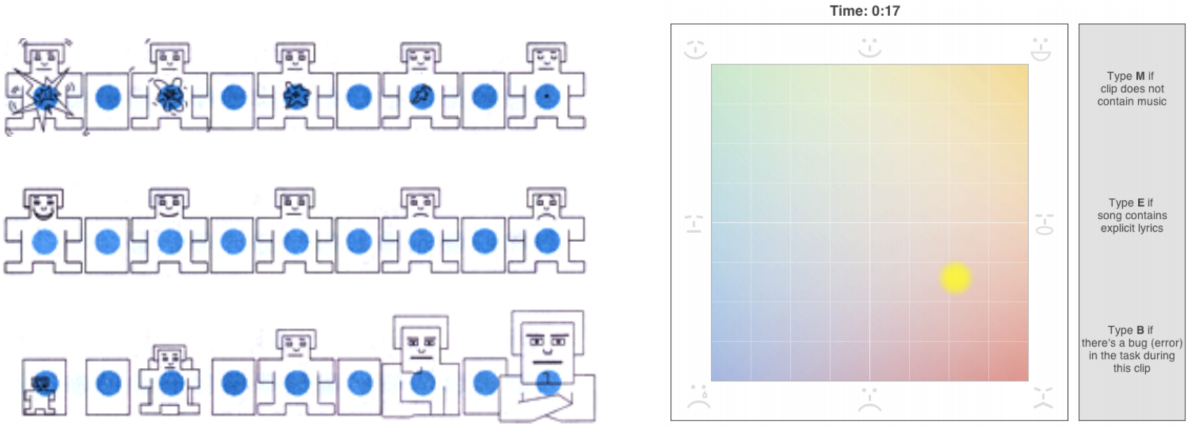
There are however some limitations that restrain the creation of a dataset. Datasets used for MER are usually smaller then 1000 songs due to the fact that music needs to be manually labeled which is a labor intensive work (Yang and Chen, 2012). Furthermore, the audio files cannot be shared publicly due to copyright issues, which causes many researchers to create their own dataset (Yang and Chen, 2012). Thus, there exists no standard and good quality datasets to make the results comparable.

Repetition of audio files is another issue that should be taken into account. Sturm (2012) recognizes different types of repetition: identical excerpts, excerpt of same song, excerpts of different versions of same song, excerpts of same artist. This repetition results in a problem that is called the 'album effect' or 'producer effect' (Kim et al., 2006; Schmidt et al., 2010). Songs that are on the same album produce similar audio features due to the fact that the recording process (instrumentation, microphones, studio acoustics, producer preferences, post production) for an album is quite consistent and differs between albums, even for the same artists. Therefore, the performance of MER systems unjustifiably increases when songs of the same album are used for training and testing. Imagine a dataset of 500 identical songs resulting in a performance of 100%. Ideally songs from different albums (or even from different producers) are chosen, or features that are affected by this effect should not be used.

Step 2. Ground truth. The dataset of music must be manually annotated with either emotion labels (e.g. happy, sad) or a dimensional value (e.g. ratings for arousal and valence) in order to get a ground truth for training and testing the learning algorithms. Ground truths are either provided by experts (generally less than five) whereby songs without consensus are excluded, or by untrained subjects (generally

more than ten) whereby all ratings are averaged (Yang and Chen, 2012). The ratings could either be for excerpts of songs, whole songs or continuous during a song to measure the variation of emotional content over time.

The information is usually collected with self-reports. In the case of categorical labeling, subjects receive a list with emotion words or images depicting the emotions using real or drawn faces. Since dimensional labeling is a bit more abstract, clear instructions about the meaning of what the dimensions mean are important. Collecting dimensional data could be done with rating scales (e.g. numerical, likert scale or with images), online with a slider bar or by selecting a position in the dimensional space, see fig. 2.4. See section 2.1.4.1 for challenges when using self-reports.



(a) Self Assesment Manikin (SAM) (Bradley and Lang, 1994). From high to low: arousal, valence and dominance.

(b) MoodSwings, online game for measuring arousal (horizontal axis) and valence (vertical axis) (Speck et al., 2011).

Figure 2.4 Self-reports

Step 3. Feature extraction. Audio features can be divided into low and high level features. Low-level audio features are audio signal based while high-level features are based on cognitive concepts (genre, mood, tonality, harmony).

Low-level musical features can be extracted from the music in the dataset using Digital Signal Processing (DSP). "Feature extraction is the process of computing a compact numerical representation that can be used to characterize a segment of audio" (Tzanetakis and Cook, 2002). Segmentation techniques are: frame based segmentation (between 10–1000ms.), beat-synchronous segmentation and statistical measures based on multiple features (Casey et al., 2008).

Commonly first the Discrete Fourier Transform is applied which decomposes an audio signal into its basic sinusoid frequencies with amplitudes (see fig. 2.5). From there other features can be extracted such as: Short-Time Magnitude Spectrum,

Mel spectrum, Chromagram, Mel-frequency cepstral coefficients (MFCC) (Casey et al., 2008).

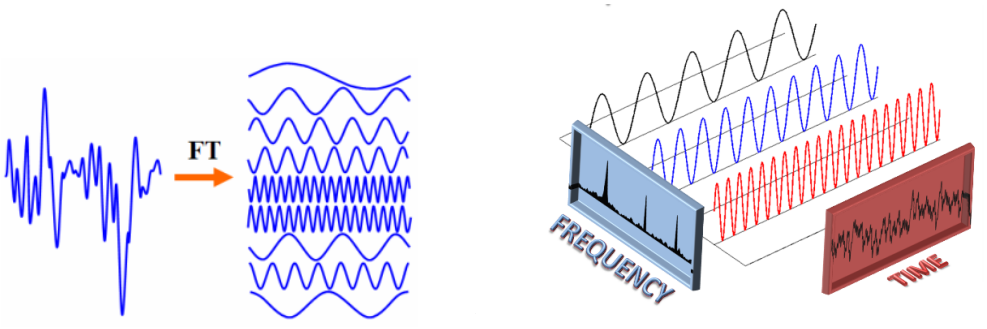


Figure 2.5 Left shows a decomposition of a signal into sinusoids (source: `www.revisemri.com/images/ft.gif`). Right shows a time to frequency transformation (source: `groups.csail.mit.edu/netmit/wordpress/wp-content/themes/netmit/images/sFFT.png`).

There exist several toolboxes for extracting audio features, e.g. MIRToolbox, Marsyas and Psysound.

Getting from low-level to high-level features based on audio analysis is still an unsolved step. This step, or gap, is also called the 'semantic gap': "the semantic gap is the gap between the extracted and indexed low-level features by computers and the high-level concepts (or semantics) of user's queries" (Tsai, 2012).

Various studies use different amounts of audio features (see table 2.4) which is partly due to the specific goal of the study, but also because there is no definitive set of features that work in MER.

Study	Features
Yang et al. (2008)	114
Eerola et al. (2009)	29
Bischoff et al. (2009)	240
Wang et al. (2010)	12
Saari et al. (2011)	66
Song et al. (2012)	55
Panda et al. (2013)	556

Table 2.4 Amount of features extracted by various studies sorted by year.

Step 4. Machine learning. Various supervised and unsupervised learning algorithms are used in MER. In both cases, emotion values for a single clip (static) or sequenced small segments of a clip (dynamic) could be predicted.

Unsupervised algorithms are usually used to find structure in unlabeled data. For example Principal Component Analysis (PCA), Latent Semantic Analysis

(LSA) or Self-Organising Maps (SOM) can be used to reduce the dimensionality of the data, select features or group data.

Supervised learning algorithms are trained to generalize beyond a training data set. A part of the feature set and the ground truths are used to train the algorithm (train set), the other part (test set) is used to check how well the model performs, see validation step.

In MER, the two main approaches are either classification or regression, respectively used for discrete and dimensional emotion models. Differences between both emotion models are discussed in section 2.1.3.2. Classification techniques are used to classify music in emotion categories. The regression approach is a known technique in dimensional emotion recognition where a regressor is trained to predict a real value from observed values (Yang and Chen, 2011). Examples of regression algorithms are Multiple Linear Regression (ML), ϵ-Support Vector Regression (SVR) and ADA-Boost.

The most important factor for the success of a learning algorithm are the features used to train the algorithm (Domingos, 2012). First of all, using too many features for learning could result in a degenerated performance (Mckay and Fujinaga, 2004). Furthermore, not only the amount of features, but also the type is important. Depending on various aspects, e.g. task (mood, genre, beat tracking) or underlying model (discrete or dimensional emotion model), different features can be selected based on theory, experiments or multivariate analysis (Gabrielsson and Lindström, 2010).

Various issues need to be taken into account when employing a learning algorithm. One type of them are overfitting and underfitting errors (see fig. 2.6). *Overfitting* occurs when the algorithm takes unimportant differences in the data into account such as noise and could be caused by high variance in the data. *Underfitting* occurs when the algorithm fails to learn important regularities from the data. Typically there is a trade-off between over- and underfitting errors. Overfitting will lead to models that only represent the training data and don't generalize well over other datasets. Underfitting models will typically generalize well over other data, but will miss important regularities in the data.

There are various techniques to reduce the chance on over- or underfitting errors (Domingos, 2012). In general a good feature selection system that only allows relevant features is important (Mckay and Fujinaga, 2004).

Another problem in machine learning is that generalizing over data becomes exponentially harder as the number of dimensions (features) grows, which is also known as the curse of multidimensionality (Domingos, 2012).

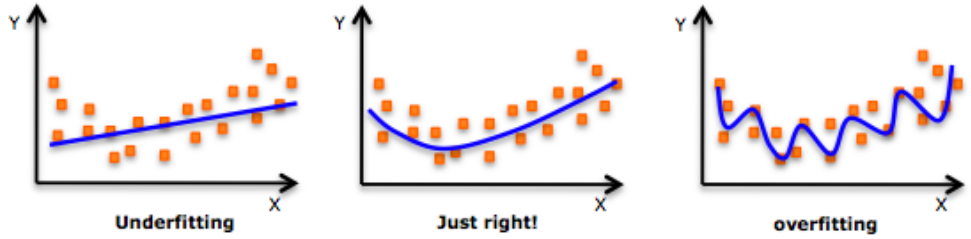


Figure 2.6 Underfitting and overfitting.

Using an optimal amount of relevant features seems to be the key to reduce the mentioned errors. As a rule of thumb a predictors-to-case ratio of 1:10 could be used which means that for every 10 features at least 100 observations are needed (Eerola et al., 2009). However, there is no golden standard so the process of 'crafting' the right features (and validating them, see next step) requires much trial and error (Domingos, 2012; Humphrey et al., 2013).

Also, the type and values of the features need to be considered. Some algorithms require homogenous data (e.g. all data scaled to [-1,1] interval), for best performance and perform poorly with heterogeneous data. Some algorithms perform bad with highly correlated features, and some algorithms work better if there is a complex interaction between the features. Usually several algorithms are tested to see which one performs best.

Step 5. Validation. Each time a 'new' model is produced, its performance is tested with the test set. Usually a cross-validation method is used in which the whole data set (features and ground truths) is divided in equal parts and each time the model is trained, another part is left out and used for the testing phase until every part is left out once (Domingos, 2012). Then all results of the cross-validation are averaged to indicate the success of the model.

Next, various statistical metrics, depending on the hypothesis, can be used to describe performance of the model (Vatolkin, 2012). However, a serious issue is that just reporting the accuracy of a model is not enough (Sturm, 2013a,b). It is important to consider the hypothesis, model, features and performance metric all together to fully understand why and how the system is performing the way it does, otherwise it is hard to tell whether the results are based on the extracted features or just confounding factors (Barthet et al., 2013; Sturm, 2013b).

It is also not a guarantee that if the model performs good on the test set (internal validity), that the model performs good on other (real world) data sets (external validity) (Sturm, 2013b). It might therefore be a good idea to test the system with 'external' data.

2.2.2.2 Tag analysis

Social tags are text labels that are freely applied to songs at social websites such as `last.fm` or `allmusic.com`. Those tags form a huge amount of human generated data that can be used for research or other applications (Levy and Sandler, 2007; Lamere, 2008). Applications of MER systems that use tags are discussed in section 2.3.2.

2.3 Related work

Related or relevant work either consists of research that: uses audio features to predict emotions in a dimensional model, processes and mines social (emotion) tags, or combines social tags and audio analysis (in relation to mood). Methods and results of studies from each of those three categories are discussed.

2.3.1 Audio analysis

Related audio analysis research predicts mood values based on dimensional emotion models such as Russell's AV model. Table 2.5 summarizes some of those studies. They all used a regression approach to predict emotions on two or three dimensional axis, whereby Eerola et al. (2009) found that a 3D model shows no significant improvement over a 2D model.

Author	**Aim**	**Method** (measure)	**Result***
Yang 2008	Emotion prediction	PCA + RFF with 18 ar. and 15 val. predictors + SVR	Arousal 58% Valence 28% (R^2)
Eerola 2009	Compare 2D and 3D emotion models, predict perceived emotion	Normalized features + BIC + PLSR with 2 components	Valence .72 Activity .85 Tension .79 (R^2)
Schmidt 2010	Estimate (time varying) emotional content of music	Individual regressors (LSR, SVR) with distinct features + second-stage regressor (LSR).	.137 ±.005 (Avg. distance)
Panda 2013	Increase emotion prediction by including melodic features	Melodic and standard features + RFF + SVR	Arousal 67% Valence 41% (R^2)

Table 2.5 Related research: emotion prediction based on dimensional models and regression. Only the best results and algorithm configuration (method) that leads to the best results are summarized. Only first authors are mentioned. * Note that R^2 is either presented as percentage or as floating point, which is how they are presented in their original studies.

A difference between the studies is that Yang et al. (2008) and Panda et al. (2013) both use the same dataset with ground-truths which are based on felt emotions, while

Eerola et al. (2009) gathered ground-truths for perceived emotions. Since it is still unclear how important the difference between felt and perceived emotions is for this kind of research (see section 2.2.1), results might be affected by it.

Yang et al. (2008) repeated the ground-truth collection process two months later with 22 of the original 253 subjects and found that the annotations given by the same person are quite similar. Based on the test-retest differences, they "compute an upper bound of R^2 for the regression approach: 85.5% for arousal and 58.6% for valence" (Yang et al., 2008).

In the method phase (third column of table 2.5) Yang et al. (2008) and Eerola et al. (2009) both reduce the dimensionality of the feature set with respectively a Principal Component Analysis and a Partial Least Squares Regression. Furthermore all studies except Panda et al. (2013) use a feature selection algorithm RRelieF (Yang et al., 2008; Panda et al., 2013) or Bayesian Information Criterion (BIC) (Eerola et al., 2009).

Yang et al. (2008) and Panda et al. (2013) both report Support Vector Regression (SVR) as the best performing algorithm and Eerola et al. (2009) reports Partial Least Squares as the best approach. Schmidt et al. (2010) uses a different approach by training individual regressors for distinct features and then combining them in a second regression-stage using Least Squares Regression. That approach does not seem to perform extremely better than the others, but it might be worth comparing both approaches in the same study to test which one performs better.

Schmidt et al. (2010) mentions a problem with the data/dimension ratio (exponential more data is required for added dimensions) to explain why combinations of features didn't perform as expected. And Eerola et al. (2009) mentions a problem with the observation:features ratio (for each feature, 10 more predictors are needed) to explain why a linear mapping method might not be sufficient.

2.3.2 Social tags

This section discusses some studies in which (mood) tags from social websites are analyzed. These studies are related because they contain a method for analyzing social emotion tags and/or discuss issues for using tags in MIR.

Tags can be collected based on tracks, tags (Laurier et al., 2009; Song et al., 2013; Saari and Eerola, 2013) or both (Levy and Sandler, 2007), and are usually balanced over genres and artists. The tags can be tokenized to remove common words such as 'the', 'it', 'and', etc. (Levy and Sandler, 2007). Tags can also be lemmatized, i.e. by grouping together inflected forms of the word (Saari and Eerola, 2013). The words in the tags are not stemmed due to the idiosyncratic vocabulary (Levy and Sandler, 2007). Commonly tags that appear less then 100 times and tracks with less then 2 tags are removed from the dataset (Laurier et al., 2009; Saari and Eerola, 2013).

Author	Summary of study
Levy 2007	Method for visualizing social tags in a low-dimensional semantic space. The method is employed for tags in general and mood tags specifically.
Lamere 2008	Review of state of the art in social tagging studies in MIR.
Laurier 2009	Creating a semantic mood space with dimensional, categorical and hierarchical representations to see if there is a difference between experts and community.
Song 2013	About whether social tags predict perceived or induced emotional responses to music. Felt and perceived emotions are highly correlated with each other and with social tags.
Saari 2014	Collect tags from last.fm. Exclude terms associated with less than 100 tracks and tracks with only one mood term. 259,593 tracks, 357 mood terms. Apply TF-IDF, LSA and SVD followed by MDS. The MDS space is linearly transformed to conform an AV space with mood words from emotion research.

Table 2.6 Related research: how tags are used in MIR and MER. Only first authors are mentioned.

When songs and tags are crawled Latent Semantic Analysis (LSA) can be used to analyze the relationship between the tags. First a term-document matrix is created whereby *'Term frequency-inverse document frequency'* (tf-idf) weighting can be used to give less weight to tags that are applied to many items (making the tags less specific to an item) and more weight to tags that are only frequently applied to a few items. Then Singular Value Decomposition (SVD) is used for reducing the dimensionality of the matrix (Levy and Sandler, 2007; Laurier et al., 2009; Saari and Eerola, 2013).

Based on the matrix, track similarity or tag similarity can be calculated (Lamere, 2008). *Tag similarity* (i.e. tags that are often applied to the same item are more similar) is usually measured with co-occurence, Jaccard distance, Dice distance, Overlap distance or Cosine distance. *Item similarity* (items that share common tags are more similar) can be computed with Pearson's correlation, Jaccard distance and co-occurence (Lamere, 2008).

Depending on the aim of the study the dimensionally reduced and term weighted document-term matrix is further processed. Laurier et al. (2009) uses cosine distance to calculate the distance between tags, the Expectation Maximization algorithm to create a categorical representation and a Self-Organising Map (SOM) for a dimensional representation of the data. The categorical representation is compared with Hevner's eight emotion categories and five MIREX clusters, and the dimensional representation is compared with Russel's 2D model. The comparison between experts and community indicates a correlation. Basic emotions are found with clustering mood tags, and the AV model shows similarities with a mood space created with a SOM. However, they present no statistical method or measure for comparing the dimensional expert and community

representations.

Levy and Sandler (2007) also use a SOM to map mood tags. The map shows some similarity with AV space, but is not evaluated further. Also a Correspondence Analysis with two dimensions is used to create a visualization of the data along the two axis.

A recent study (Saari and Eerola, 2013) proposes a new technique called Affective Circumplex Transformation (ACT) for representing moods of music based on social tags and the AV model. First, a non-metric Multidimensional Scaling (MDS) is applied to the SVD matrix. Then the MDS space is linearly transformed to conform an AV model with mood terms which are obtained from psychological research. The result is a model for mapping tags to AV ratings. The ACT model is compared with, and outperforms, other techniques such as SVD, Nonnegative Matrix Factorization and Probabilistic Latent Semantic Analysis for valence, arousal and tension.

In another study, Song et al. (2013) found that felt and perceived emotion ratings on AV scales of test subjects are highly positive agreement. Furthermore, by dividing the AV space into four quadrants and assigning each quadrant a tag: 'happy', 'sad', 'angry' and 'relax', they found positive agreement between the ratings and last.fm tags of the same songs (except for 'relax'). In contrast with other studies, this study found a higher agreement among participants for induced emotions, suggesting that the mechanisms for perceived and felt emotions require more research (see also section 2.2.1.1).

A possible issue with this research is the fact that only four tags are considered, discarding the other tags in that track which could have described another mood. Also the participates rated songs on the AV scales, but then the ratings were reduced into four quadrants which is a huge reduction of the variance in the ratings, increasing the chance of agreement with the tags from last.fm.

2.3.2.1 Problems with tags

Table 2.7 lists various types of problems with tag pools for researchers and users. Using autocorrect for misspellings could produce unwanted results due to idiosyncratic vocabulary and tags with proper nouns (Levy and Sandler, 2007). For the same reasons, stemming algorithms or lemmatizers, which could be used to account for synonyms, might introduce errors.

An advantage of clustering algorithms is that they can map a large variety of tags based on co-occurence which relate less common tags (e.g. different spellings or synonyms) with more canonical tags or disambiguate between word senses based on the other tags (Lamere, 2008).

Different tags with opposite meanings in the same tag pool are a serious issue for supervised learning algorithms, e.g. if a song is tagged with only two labels 'happy' and 'sad', then either the song is categorized as sad or as happy, thereby disqualifying a

Type of problem	Description
Alternative **spellings**, misspellings or synonyms.	Richer pool of tag information from user community (folksonomy). Problem for tag based search or supervised learning, but not necessarily for unsupervised clustering.
Ambiguity of tags.	It is not always clear which meaning is intended, e.g. does the tag 'love' refer to someone's feeling when listening to the song or that someone loves the song.
Different tags with **opposite meanings** in same tag pool.	Issue for categorizing songs based on tags, which one is more right (if that's even possible with subjective content)?
Discursive tags, i.e. tags consisting of multiple words.	Discursive tags are less useful.
Malicious tagging, i.e. intentionally wrongly placed tags.	Serious issue, could result in any of the above mentioned problems and pollutes the data set. E.g. Paris Hilton was number one 'brutal death metal' artist based on tags on last.fm (Lamere, 2008).
Variance and bias in the tag sets.	High variance in all tag sets decreases the distinguishability of the tag set, making it less usable. Wrongly biased tags provide wrong information.
Tagger bias, i.e. tags represent the taste of user community.	Makes results in research less generalizable to whole population.
Uneven tagging or 'cold start' problem.	Popular content is tagged more than unpopular content. While this is a problem for music recommendation services, it also biases the information available for research in the direction of popular music, e.g. in MoodSwings (Kim et al., 2008).

Table 2.7 Overview of some possible issues with tag sets. These problems are further discussed in (Levy and Sandler, 2007; Lamere, 2008; Levy and Sandler, 2009; Nanopoulos and Karydis, 2011)

tag that might be as informative as the other for a part of the user community. The underlying problem here is that it's hard to separate wrong tags (e.g. malicious tagging) from tags that are inconsistent with other tags in the same pool, but which still reflect the opinion of some users.

While high variance in tags can be bad, a greater variety of tags for a song also creates a more nuanced view on how the community judges the song, also known as a 'folksonomy'[2]. Weighting techniques such as tf-idf could be applied to give some tags lower or higher importance to take care of some of those issues. Wrongly biased tags, if subjectively added tags (ignoring malicious tagging here) can be wrong in the first place, can for example be caused by tagger bias.

[2]A combination of the words *folk* and *taxonomy*, referring to a collaboratively created classification scheme (Bischoff et al., 2009).

2.3.3 Social tags and audio analysis

Author	Aim	Method	Result
Bischoff 2009	Use mood tags and audio features for mood classification of songs.	Tags and audio features with ground-truth from allmusic.com. Best classification for audio: SVM with RBF kernel, for tags: Naïve Bayes Multinomial. Both classifiers are linearly combined.	Audio signal performs worse than tags, together they perform best.
Wang 2010	Use tags and audio features for artistic style clustering.	Collect tags and extract audio features. Compare various clustering methods with own method (TC) for tag-artist, content-artist and combined matrices. Use two similarity methods: exp-weighted graph and ϵ NN graphs.	Audio feature classification performs worse, TC method performs best when tags and audio are combined.
Nanopoulos 2011	Solution for 'cold-start' problem.	Train, based on tags and audio features, which audio features (f) are important for similarity. Find nearest neighbor (x) of query track based on f. Get nearest neighbors (y) based on tags of x. Return y.	Proposed system works better than using just audio for training.
Saari 2013b	Predict perceived mood based on audio with new technique: Semantic Layer Projection (SLP).	Tags are used to create a semantic layer: ACT (see Saari2013a prev. sec.). Create mapping between audio features and ACT with PLS. Apply linear regression between layer representations and listener ratings. Compare with baseline: direct mapping of features to ratings with PLS and SVR.	SLP outperforms baseline. R^2: Valence .33 Arousal .78
Saari 2013c	Extends Saari 2013b with tags.	Predict mood: with audio using SLP and ACT, with tags using ACT, and both methods combined.	Audio and tags combined performs best. R^2: Valence .49 Arousal .74 Tension .62

Table 2.8 Related research: audio features and tags combined. Only first authors are mentioned.

One way to use both tags and audio signal is to use tags as ground truth for a system trained with audio features (Song et al., 2013). However, only the studies that employ both tags and audio analysis as a source of information for MIR tasks are reviewed here.

Table 2.8 shows a few studies that combine audio signal analysis with tags for: artistic style clustering (Wang et al., 2010), discrete mood prediction (Bischoff et al., 2009),

dimensional mood prediction (Saari et al., 2013b,a) or proposing a solution for the 'cold-start' problem (Nanopoulos and Karydis, 2011).

Both Bischoff et al. (2009) and Wang et al. (2010) report that the system performs worst with only audio features and best with combined audio features and tags, whereby in the latter study it is found that audio features improve the system only a little when combining with tags. These differences are partly also found in (Saari et al., 2013a) where tags alone perform worst for predicting arousal and audio features alone perform worst for predicting valence. However, for both arousal and valence the combination of tags and audio gives the best results.

An explanation for the difference in results for audio features and tags alone between (Bischoff et al., 2009; Wang et al., 2010) and (Saari et al., 2013a) could be that in the latter tags are also implicitly used for predictions based on audio features alone. Therefore the difference between prediction success for audio features and tag performance could be smaller.

The studies by (Saari et al., 2013b,a) are most relevant since they use tags and audio signal together to predict mood in a dimensional space. Bischoff et al. (2009) also predicts mood, but with a *classification* method and Wang et al. (2010); Nanopoulos and Karydis (2011) are about style clustering and cold start problem.

There are two important differences between the current study and the Semantic Layer Projections (SLP) model proposed by Saari et al. (2013b). First the aim of the current research is to improve the quality of tag sets, while the aim of (Saari et al., 2013b) is to improve dimensional AV mood prediction. Second, in the current research psychological emotion models are used to translate predicted AV ratings into emotion labels, whereas (Saari et al., 2013b) uses those models together with tags to create a semantic layer that is used to improve dimensional mood prediction.

2.4 Conclusion

Based on the discussed literature a few important observations can be made. First, the words 'emotion' and 'mood' are commonly used interchangeably in MER to denote the same phenomenon: the feeling aspect of emotions or moods, also called aesthetic emotions. Second, while tags are a commonly used source of information in MER, no research was found about the validity of the tags. Third, social tag mood spaces seem to conform to dimensional emotion models (Levy and Sandler, 2007; Laurier et al., 2009). Fourth, only recent studies by the same research group (Saari et al., 2013b,a) combined audio features, social tags and emotion models, suggesting that this might be a potential new direction for MER research.

Chapter 3

AV prediction

In this chapter arousal/valence (AV) prediction models will be trained and tested with extensive grid searches and cross validation on three datasets. A fourth dataset (MSE2700) is compiled by combining the three datasets. The datasets all contain audio clips that are rated on AV scales by humans. Each of the three original datasets is used as train (80%) and test set (20%) whereby the two other datasets are used as test set. This set up makes it possible to thoroughly test the robustness of the trained models.

The R^2 scores of the models that are trained and tested on single datasets are state of the art, resembling the results of the studies that the datasets are obtained from. However, when testing the models on the two other datasets, the R^2 scores worsen drastically, especially for the valence models. These results indicate that the models are not robust.

The purpose of the AV prediction models was to be used in the tag weight prediction method described in chapter 5. However, the models turned out to be too unreliable and are therefore not further used. In chapter 5 human rated instead of predicted AV values are used.

3.1 Audio datasets

Various datasets (summarized in table 3.1) containing audio clips with human annotated AV ratings are assembled into one large dataset.

Name		**Clips**	**Length**	**Music type**	**Annotators**
MediaEval 1744[1]	(Soleymani et al., 2013)	1744	45s.	Various	>10 per song
Saari596[2]	(Saari and Eerola, 2013)	596	15–30s.	Various popular	±29 per song
Eerola360[3]	(Eerola and Vuoskoski, 2010)	360	10–30s.	Film music	12 per song

Table 3.1 Annotated audio datasets.

The *MediaEval1744* dataset consists of the train and test sets used in the MediaEval 2014 contest[1] (Soleymani et al., 2013; Aljanaki et al., 2014b). The audio clips of the *Saari596* dataset were collected based on information in the files made available online[2] by Saari and Eerola (2013). The original dataset describes 600 clips, however 4 clips couldn't be retrieved. Also included in the dataset are emotion tags for the clips retrieved from last.fm with an average of eight tags per song. The *Eerola360* dataset consists of film music and is created by Eerola and Vuoskoski (2010). A subset of 110 clips were also annotated by 116 annotators and composed into a separate dataset. Both datasets are available online[3]. The larger dataset is chosen here because it consists of more songs and is still sufficiently annotated by experts.

3.1.1 Preprocessing and assembling the MSE2700 dataset

The three datasets are combined into one large dataset henceforth called MSE2700 referring to the first letters of the datasets with the amount of audio clips (see fig. 3.1). The dataset contains 2700 AV annotated clips which are all down-sampled to the lowest bit rate and frequency found in the source sets which is: 22050Hz, 64kbps, mono channel and mp3 format. Loudness is normalized before the feature extraction process.

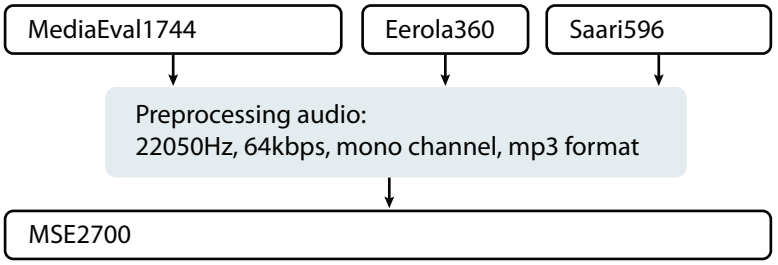


Figure 3.1 Sources and features of the MSE2700 dataset.

3.1.2 Feature extraction

For the current study 376 features are extracted with the standard feature extraction function $mirfeatures()$ from MIRToolbox 1.5[4] (Lartillot and Toiviainen, 2007). See a list of all features in appendix A. The features are normalized between 0 and 1.

The reason for using this feature extraction toolbox (and feature set) is that it is easy to implement and encompasses all commonly used audio features in MER research.

Choosing the right features for machine learning is crucial for the performance of the model (Domingos, 2012; Humphrey et al., 2013), however, there are no proven good

[1] `http://www.multimediaeval.org/mediaeval2014/emotion2014/`
[2] `http://hdl.handle.net/1902.1/21618`
[3] `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/emotion/soundtracks`
[4] `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/mirtoolbox`

features known when it comes to predicting arousal and valence. For example, in a MER review study Barthet et al. (2013) show that timbre related features are most commonly used in MER systems because they result in the best performances. However, as Sturm (2013b) argues, uncontrolled independent variables are often overlooked in the process of validating MER systems. As an example he has created a classification system that correctly classifies two songs, however after applying a filter that changes timbral aspects of the signal that are irrelevant to human ears, the system misclassifies both songs. This shows (i) that feature extracting algorithms are not (yet) able to extract timbre related features properly, and (ii), more general, that MER systems might not work the way we think they do. Supposedly irrelevant changes lead to unexpected results.

While this is a serious issue for MER, AV prediction is a secondary goal of this research. Thereby, using a large dataset that is composed of three different sources might account for some uncontrolled independent variables that could arise from single source datasets.

3.2 Method

All four datasets (MSE2700, MediaEval1744, Saari596 and Eerola360) will each be used in training and testing nine different models (see table 3.2). The models are trained and tested with both normalized and non-normalized features. Furthermore, twelve different amounts of k-best features are used. This set-up will result in a total of: 2 (arousal/valence) $*$ 4 (datasets) $*$ 2 (normalization) $*$ 12 (features) $*$ 9 (different models) = 1728 trained models in total, see fig. 3.2 for an overview. The employed nine models are selected based on commonly used models in MER research and the available regression models in the Scikit-Learn library.

Full name	**Abbreviation**
AdaBoost Regressor	AdaBoost
Bayesian Ridge Regression	BayesRidge
DecisionTree Regressor	DecTree
Gradient Boosting Regression	GradBoost
K-Neighbors Regressor	k-NN
Least Angle Regression	Lars
Linear Regression	LinReg
Passive Aggressive Regression	PAReg
Support Vector Regression	SVR

Table 3.2 Used regression models and how they are abbreviated in this thesis.

3.2.1 Evaluation with R^2

The AV predictions of each model are evaluated with the explained variance (R^2) metric, see equation 3.1, whereby T denotes the true values and P the predicted values. R^2 describes the proportion of variance that is explained by the model. It has an upper limit of 1, denoting a perfect fit of the model and it has no lower limit. Typically, the effect of R^2 can be interpreted as follows: $.04 = Small$, $.25 = Medium$ and $.64 = Large$ (Sullivan and Feinn, 2012).

$$R^2 = 1 - \frac{\sum_i (Ti - Pi)^2}{\sum_i (Ti - \bar{T})^2} \tag{3.1}$$

3.2.2 Validity

The reason for this set-up is that it allows for thorough comparison between datasets, models and feature amount. Thereby, each model will have undergone a three step validation: (1) cross-validated grid search on the training part of the train/test set, (2) tested on the test part of the train/test set and (3) tested on the three other datasets.

If, for example, a model scores high on R^2 for the train/test set but very low on R^2 for the other three datasets, then the model is not very robust.

3.2.3 Implementation

The whole pipeline is implemented in Python[5] version 2.7.5. Various functions of the Scikit-Learn[6] package (version 0.15.2) are used to create each step (including the regression models) of the pipeline. Scikit-Learn is a well documented, open source Python machine learning toolkit.

3.2.4 Pipeline

Figure 3.2 shows the pipeline that is used for training and testing AV prediction models. In the following subsections each part of the pipeline is explained.

Train/test set and test sets Each dataset is used as train/test dataset while the other three are used as test set. The train/test set is randomly split in a train set (80% of the set) and a test set (20% of the set). The test set is, just as the three other datasets, kept apart for the final evaluation.

Missing values Missing values in the feature sets are imputed with the mean of that feature. If all values of a feature were missing, they will all get a 0. A situation

[5] `https://www.python.org/`
[6] `http://scikit-learn.org/stable/index.html`

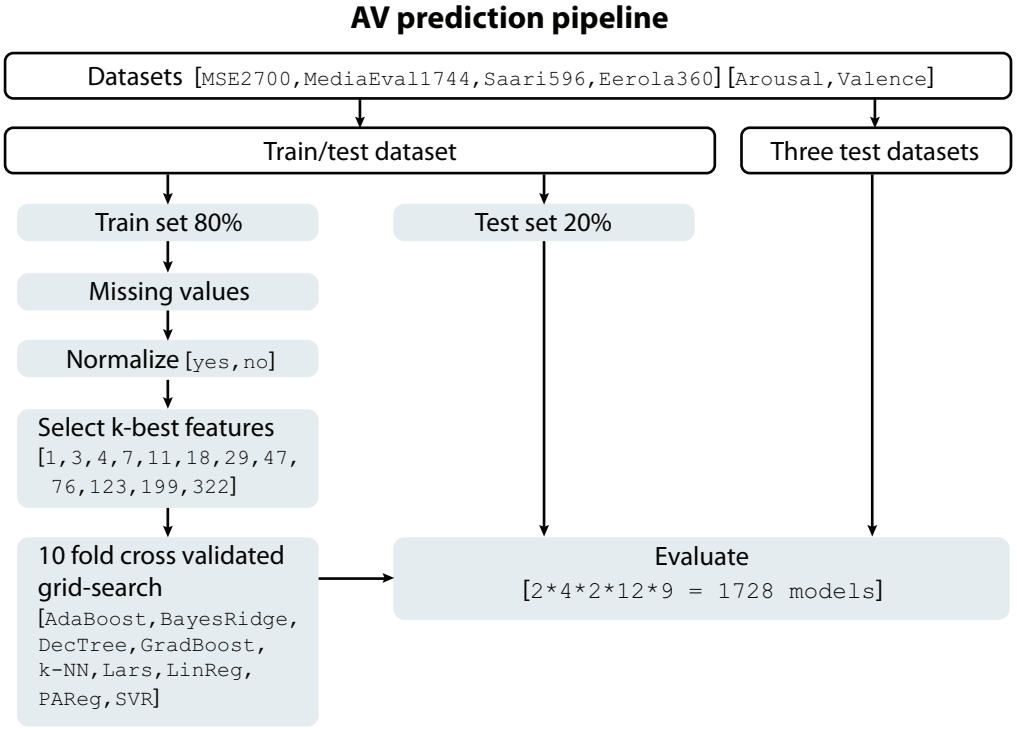


Figure 3.2 Training and testing phases of AV prediction pipeline. All unique permutations of the variables between brackets '[ ]' is performed which results in a total of 980 trained and tested models.

where all values of a feature were missing did not occur often, and might have been caused by some problems in the MIRToolbox. Missing values are imputed on each train and test set separately so that no information of the train set would be used in the testing phase.

Normalize The audio features are either normalized or not. For some models it is known that normalizing the features improves the results dramatically (e.g. SVR), and for some models it does not affect it much (e.g. Lars). Normalization is performed on each train and test set separately so that no information of the train set would be used in the testing phase.

Feature selection Twelve different amounts of k-best features are tested, whereby $k = [1, 3, 4, 7, 11, 18, 29, 47, 76, 123, 199, 322]$. The best features are selected based on F-values between feature and ground truth. For example if k = 3, then the 3 features with the highest F-values are used.

Cross validated gridsearch Nine different models are employed in a 10 fold cross validated grid search in which various parameter settings (varying per model) are tested and optimized. The best parameter setting per model is based on the R^2 score.

Evaluation The 'best' cross-validated parameter settings for each model is used to predict arousal or valence values for the left out 20% of the dataset and for the three other datasets.

Note that arousal and valence are separately trained and tested. So from the 1728 models, half are arousal models and the other half are valence models.

3.3 Results

All models are trained (80%) and tested (20%) on each of the four datasets and tested on the remaining three datasets. Thus each dataset is used once as a train/test set and three times as test set.

First, the best R^2 scores of the models that are only tested on 20% of the train/test set will be compared to the results from the studies the datasets are obtained from. Training and testing on the same dataset is common practice in MER.

Second, the results of those 'best models' on the train/test sets will be compared to their results on the three remaining datasets. This will demonstrate that the 'best models' are actually performing poorly and alternative best models are presented.

Third, the learning curves are presented in appendix B.

3.3.1 Best R^2 scores for train/test sets

The highest R^2 scores on the train/test sets are presented in table 3.3 under the heading 'current'. To put these results in a more meaningful context, the R^2 results from the studies the datasets are obtained from (see table 3.1) are also included.

	Arousal R^2		**Valence R^2**	
Dataset	**Current**	**Original**	**Current**	**Original**
MSE2700	.71	–	.48	–
MediaEval1744	.72	.63*	.49	.35*
Saari596	.80	.77	.33	.25
Eerola360	.66	.85**	.72	.72**

Table 3.3 Best R^2 results on current train/test sets compared with best R^2 results from the studies where the datasets are obtained from. * = These results are obtained from a subset of the MediaEval1744 dataset (700 clips for training, 300 clips for testing). ** = These results are obtained from a smaller subset (110 clips) of the Eerola360 dataset.

The comparison between R^2 scores from current and 'original' studies gives an indication of how well the currently trained/tested models perform. For the results from current and original studies the same train/test logic is used with extracted audio features, AV rated clips and similar learning algorithms.

In the next section it will be shown that the currently obtained R^2 scores are less promising as they appear.

3.3.2 Naive and robust results

One way to test the robustness of a model is to test it on other datasets. In the previous sections each of the four datasets: MSE2700, MediaEval1744, Saari596 and Eerola360, are used to train and test AV prediction models. The results of the best models (based on R^2 score) are presented in table 3.3.

The current question is how well those models (henceforth referred to as 'naive' models) would perform on other datasets. To answer this question, the naive AV models that are trained and tested on each of the MediaEval1744, Saari596 and Eerola360 datasets are also tested on the other two datasets. For example, the best model that is trained (80%) and tested (20%) on the MediaEval1744 dataset (in this scenario the train/test dataset) is also tested on Saari596 and Eerola360 (test datasets). The MSE2700 dataset is not used since it contains all other three datasets, so testing its best model on the other three datasets would be non informative.

In addition to the naive models, it is also examined which models are most robust. For example, all the models that are trained (80%) and tested (20%) on the MediaEval1744 dataset, are also tested on Saari596 and Eerola360. So every model is tested on three datasets and has therefore three R^2 scores. From all these models, the model that scored the least lowest R^2 on all three datasets is picked as the most robust model.

Using the 'least lowest' as criteria for robustness clearly indicates the minimum capabilities of the models. Another criteria could have been average performance. The downside of average performance is that it could select a best robust model that scores good (i.e. high R^2) on two datasets but relatively bad on the third. So, while the model scores on average best, its minimum capabilities could be worse compared to other models.

Table 3.4 contains the AV prediction R^2 results of the best naive and robust models on the datasets. The presented R^2 results are obtained from 20% of the train/test set (marked with '$*$') and the other two datasets that are used as test sets. In addition, figure 3.3 provides a visual presentation of the R^2 scores. The train/test sets are marked with '$*$' and labeled on the x-axis.

In the next sections the results are discussed in more detail.

3.3.2.1 Valence models

Valence naive models score relatively high on train/test sets, however the results on the test sets are in most cases below zero indicating that the models are not able to generalize

	Arousal			Valence		
	R^2	**Features**	**Model**	R^2	**Features**	**Model**
Best naive models						
MediaEval1744 *	0.72	199	LinReg	0.49	123	GradBoost
Saari596	0.65			−0.31		
Eerola360	0.01			−0.14		
MediaEval1744	0.44	199	BayesRidge	−0.22	199	AdaBoost
Saari596 *	0.8			0.33		
Eerola360	−0.03			0.13		
MediaEval1744	−0.15	76	AdaBoost	−1.48	199	GradBoost
Saari596	0.09			−1.06		
Eerola360 *	0.66			0.72		
Best robust models						
MediaEval1744 *	0.52	322	SVR	−0.01	7	SVR
Saari596	0.65			−0.12		
Eerola360	0.38			−0.07		
MediaEval1744	0.52	76	k-NN	−0.06	123	SVR
Saari596 *	0.57			0.07		
Eerola360	0.32			0.11		
MediaEval1744	0.26	3	k-NN	−0.09	3	k-NN
Saari596	0.37			−0.35		
Eerola360 *	0.33			0.21		

Table 3.4 Best naive and robust models on three datasets. The '$*$' marks the train/test set on which the model is trained (80%) and tested (20%). The other two models are used as test set.

over other datasets. Thereby, also the best robust models preform very poorly since there is no model that scores a positive R^2 on all three datasets.

3.3.2.2 Arousal models

Arousal models have high R^2 scores for the naive train/test ($*$) sets, but not for all its test sets. The best robust models have overall lower, but more equal R^2 scores with $R^2 \approx .4$ for MediaEval1744$*$ and Saari596$*$, and $R^2 \approx .3$ for Eerola360$*$. These scores are not impressive but show at least a medium effect for the best robust models.

3.3.2.3 Eerola360 dataset

The naive arousal models trained on MediaEval1744$*$ and Saari596$*$ are relatively good at predicting arousal values of each other, but unable to sufficiently predict arousal values of the Eerola360 dataset. Conversely, models trained on Eerola360$*$ are unable to sufficiently predict arousal values for the MediaEval1744 and Saari596 datasets.

This difference could be explained by the fact that Eerola is rated on 'energy' instead of 'arousal'. While these two dimensions are in the literature sometimes described as

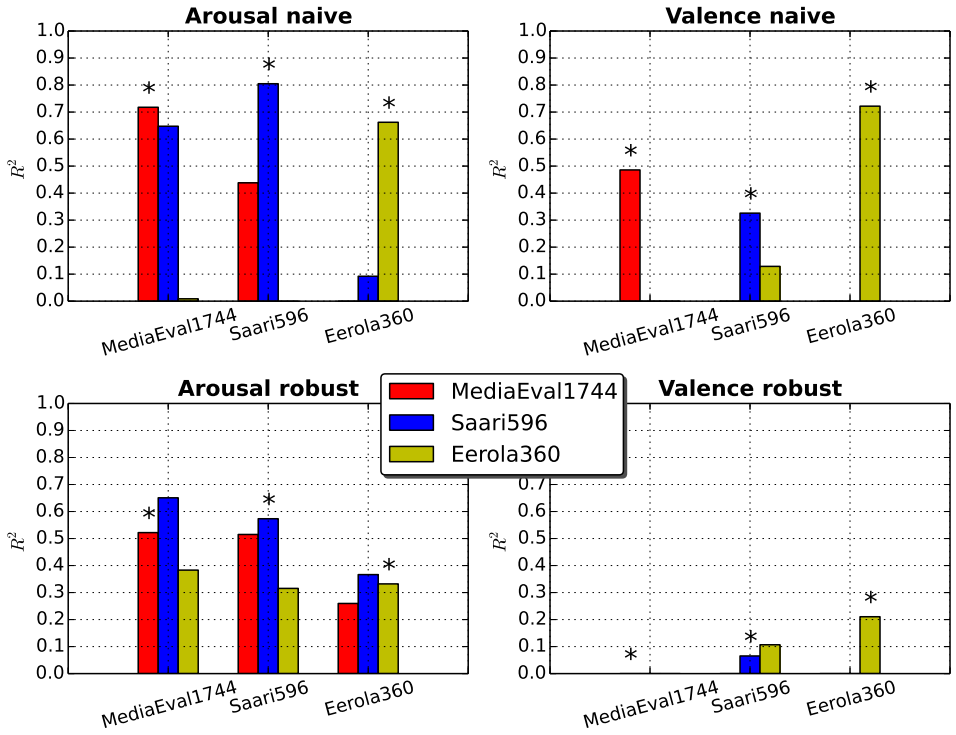


Figure 3.3 Models that are trained and tested on the train/test sets (marked with '∗' and labeled on the x-axis) are evaluated on the other two datasets.

exchangeable, it could be that the different terminology caused different rating behavior of the raters.

3.3.2.4 Overfitting

Eerola et al. (2009) advises a features/observation ratio of 1:10, which means that for each extra feature used in training phase, there should be ten more rated audio clips. Violation of this rule of thumb might result in overfitting models.

As can be seen in table 3.4, only the robust arousal and valence models for Eerola360∗ and the robust valence model for MediaEval1744∗ satisfy this rule, and all the other trained models violate it. So are these other models victim to overfitting based on this rule? While it is certainly possible that these models are overfitting, it can not be concluded with certainty[7]. However, elaborating this issue in depth is beyond the scope of this thesis.

Another reason for suspecting overfitting is when a model scores well on similar data to the train set, but not so well on other datasets. This is the case for the naive arousal and valence model. While overfitting might be suspected there, another explanation for these results could be that the data is bad resulting in bad models. There might be a difference in how the AV truth values of the data sets are acquired making it difficult

[7] `www.jakevdp.github.io/blog/2015/07/06/model-complexity-myth/` (22–11–2015)

to predict values of one dataset when trained on the other. Also it could be that the models are unable to account for the different types of music: film music (Eerola360) vs. popular music (Saari596).

An argument against the idea that the results of the naive AV models are caused by bad data are the results of the robust arousal models. The robust arousal models have all modest R^2 scores. If the data was bad, these results would have been more similar to the results of the naive models.

It is therefore concluded that it is likely that the naive arousal and valence models are overfitting. The robust arousal models show no indication of overfitting and the robust valence models are so bad that the question whether there is overfitting is pointless.

3.3.3 Summary

In general, the results suggest that arousal prediction models that are trained and tested with audio features are less susceptible to confounding factors compared to valence models. Furthermore, it can be concluded that the trained valence models are not generalizable to other datasets and their initial relatively high R^2 valence scores (see table 3.3) should not be trusted.

3.4 Conclusion

The goal of this chapter was to create AV prediction models that could be used in the proposed tag weight prediction method in this thesis. After extensive training and testing nine different models on four datasets, it is concluded that the currently trained models are not sufficient for being further used. The arousal prediction results are modest with robust R^2 scores of $\approx .4$. The valence prediction results are just bad with no model that scores a positive R^2 on all tested datasets. These results endorse the current state of affairs in AV prediction systems in MER.

The main result of this chapter is that high R^2 scores can be obtained when the models are trained and tested on the same dataset, even when the tested data is not seen during training. However, testing the models on other datasets shows that these results are not robust and are probably overfitted on the train/test dataset. These results relate to the critique of (Sturm, 2013b) on the validity of MER systems.

Chapter 4

Emotion words in AV space

This chapter describes the construction of a large dataset with arousal/valence (AV) rated emotion words. The purpose of this dataset is to be used in the tag weight prediction method described in chapter 5 where emotion tags need to be mapped in AV space. The resulting emotion wordlist contains 1397 AV rated emotion words.

First, a set of emotion words is compiled based on various emotion studies (section 4.1). Second, a large dataset of words rated on AV scales is used to collect AV ratings of the emotion words (section 4.2). Finally, the resulting dataset of AV rated emotion words is analyzed and visualized in AV space (section 4.3).

4.1 Emotion words

A list of 3099 emotion words (hereafter referred to as *emotion wordlist*) is compiled from various emotion studies, see table 4.1. The emotion words consist of nouns, verbs, adjectives and adverbs.

First in section 4.1.1 the origin and use of these words in their original studies are discussed. Second, in section 4.1.2, some problems and the usefulness of this emotion wordlist for the purpose of this thesis is discussed.

4.1.1 How are the emotion words selected?

The studies and sources listed in table 4.1 employed various methods for selecting the emotion words.

Shaver et al. (1987) and Zentner et al. (2008) first conducted experiments to select emotion words that were easily understood by test subjects. Russell and Mehrabian (1977) and Russell (1980) selected respectively 151 and 28 emotion words to cover a wide range of emotions in the affective space.

Two studies specifically targeted music emotions. Zentner et al. (2008) selected emotion terms that were most suitable for describing music. The selection was made

Emotion words source	Word count
Russell (1980)	28
Hevner (1936)	66
Zentner et al. (2008)	67
Shaver et al. (1987)	136
Hu et al. (2009)	138
Russell and Mehrabian (1977)	151
`allmusic.com`	280
Saari and Eerola (2013)	687
WordNet-affect-1.1	746
WordNet-affect-1.0	2904
Unique emotion words	*3099*

Table 4.1 Sources of emotion words.

based on ratings of test subjects who had to rate how often they felt or perceived those emotions when listening to music.

Hu et al. (2009) employed four steps to filter a large dataset of words for musical emotion words. First, WordNet-affect (Strapparava and Valitutti, 2004) is used to filter the emotion words. Second, words that also contained musical meanings are removed, e.g. 'trance' or 'beat'. Third, judgmental words are removed, e.g. 'bad','good' or 'poor'. At last, words with ambiguous meanings are removed, e.g. does someone 'love' the track, or does the track elicit the feeling of 'love'?

The emotion words form `allmusic.com` are relevant since they are used to search music, however it is not clear how and why those words are selected.

WordNet-affect was developed as a linguistic resource of lexical affective knowledge (Strapparava and Valitutti, 2004). Both WordNet-Affect[1] sources (Strapparava and Valitutti, 2004) are included in WordNet Domains 3.2[2] (Magnini and Cavagli, 2000), which is a large database of English words with one or more domain labels.

The affective words in WordNet-affect-1.0 are labeled with word type (noun, verb, adjective or adverb), and with 'emotion type' information (emotion, mood, trait, cognitive state, physical state, hedonic signal, emotion-eliciting situation, emotional response, behavior, attitude, sensation). Identical words with different meanings are treated as separate words. The affective words are initially selected by hand and later extended by derivations, e.g. nouns from adjectives.

The WordNet-affect-1.1 set is a hierarchically (based on valence) ordered subset of version 1.0. It provides a more structured semantic organization and additional labels.

[1] In 'wn-domains-3.2/WordNet-affect-1.1/a-hierarchy.xml' the word 'simpathy' is corrected to 'sympathy', and the line '$< categname = "surprise" isa = "astonishment" / >$' is removed since astonishment already refers to surprise

[2] `http://wndomains.fbk.eu/publications.html`

4.1.2 Critical review of the *emotion wordlist* and its use

The 3099 unique words from the composed *emotion wordlist* are at least related to emotion. There are however some complications. A list of unique emotion words does not disambiguate between different meanings of a word (such as WordNet-affect does), for example an emotion word could be used to describe the content of a song, but also to express someone's opinion about the song. Furthermore, a word in the *emotion wordlist* does not necessarily refer to an affective meaning. For example the word 'dark' could refer to a dark feeling. However, describing the absence of light with the word 'dark' is not necessarily affect related.

The *emotion wordlist* is mainly compiled to filter tag sets for emotion words. Those emotion words are then weighted based on how well they describe the emotional content of the song. Therefore, if there are more ambiguous emotion words in the *emotion wordlist*, it means that the chance on erroneous filtering also increases. Erroneous filtering means that tags that don't describe, or are not intended to describe, the emotional content are selected. E.g. 'love' might be selected while it was meant as someone's opinion on the music.

Filtering differently intended tags from the tag sets is impossible since the intentions of the tagger are impossible to determine based on the available data, and might even for the tagger himself be unknown or change over time.

Another solution would be to compile an *emotion wordlist* with unambiguous words such as Hu et al. (2009) did. This would decrease the amount of erroneous selection, but also decrease the amount of tags that can be used.

A third, and optimal, solution would be to treat every tag as if it was intended to describe the musical content of a song. Differently intended tags could still describe the musical content and if they are not, then they should be filtered by the proposed method.

The main reason to choose the third solution is that removing possible ambiguous tags from the list would limit the method to a small list of emotion words (138 by Hu et al. (2009) versus 1379 in the current research). Moreover the intentions could not be known, so filtering for possible ambiguous tags partly removes tags describing the musical content.

4.2 AV rated words

This section describes how the AV ratings for the emotion words are obtained. To this end two corpora with AV annotated words created by Warriner et al. (2013) and Bradley and Lang (1999) (further referred to as respectively Warriner and ANEW) are used. In both studies people rated english words on arousal, valence and dominance scales. The

ANEW dataset consists of 1035 words which are (except for the word 'grime') also used in the Warriner dataset that consists of 13915 words, see table 4.2.

AV rated words source	**Word count**
Bradley and Lang (1999) 'ANEW'	1034
Warriner et al. (2013) 'Warriner'	13915
Total unique AV rated words	*13916*

Table 4.2 Sources of AV rated words.

4.2.1 How are the ratings obtained?

Most words in the Warriner dataset are rated by 18 to 30 persons. The words were rated on three scales happy-unhappy (valence), excited-calm (arousal) and controlled-in control (dominance). For each word the participant was asked to rate the word based on what they felt when reading the word. The questionnaires were distributed via mechanical turk, so a large diversity of participants rated the words in their own environment.

In the ANEW study, the word ratings are obtained during one hour sessions with Introductory Psychology class students. How many students participated or how many ratings a word received is not reported. The words were also rated based on how the student felt when reading the word.

4.2.2 Consistency of the ratings

Both datasets are rated by different people under different circumstances, which provides the opportunity to compare the consistency of the ratings for words that are in both datasets. Figure 4.1 shows the correlation between overlapping words for arousal and valence ratings. Valence ratings have the highest correlation, $r = 0.95$, and arousal also has a high correlation of $r = 0.76$. Both correlations are significant, $p < .001$.

Furthermore, in the Warriner dataset, the mean standard deviation of valence ratings is 1.82 while it is 2.14 for the arousal ratings. In the ANEW dataset, the mean standard deviation of valence ratings is 1.66 and 2.37 for arousal. The higher mean standard deviation on arousal ratings in both datasets suggest that rating words on arousal scale is more subjective or difficult.

4.3 AV rated emotion words

The 3099 unique emotion words are matched with the Warriner dataset resulting in 1379 AV rated emotion words, see table 4.3. For the current research the ratings of

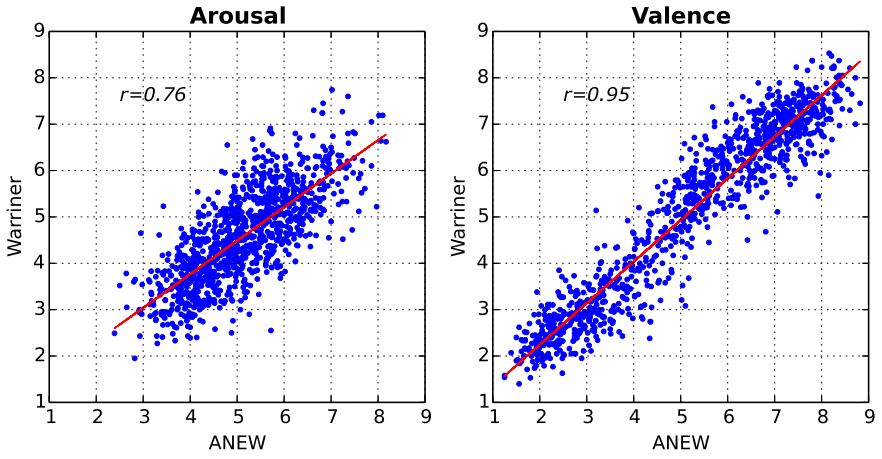


Figure 4.1 Linear correlation between AV ratings for 1033 overlapping words of the Warriner and ANEW datasets. For both regressions $p < .001$.

the Warriner dataset are used so that all ratings are from the same source. Thereby all words in the ANEW dataset are also in the Warriner dataset.

A reason that less then half of the unique emotion words (1379 out of 3099) were found in the AV rated word files is that many emotion words are deviations or conjugations, which are not all rated for AV values. See figure 4.2 for a visual representation of the AV rated emotion words in AV space.

Sources	**Word count**
Emotion wordlist	3099
AV rated words (Warriner)	13915
Total AV rated emotion words	*1379*

Table 4.3 AV rated emotion words.

4.3.1 Discussion

The constructed AV rated emotion word list contains 1379 words. The AV ratings of words were obtained by asking what the participant felt when reading the word. This will result in rated emotion words based on daily use. However, ratings might have been different when the question was to rate the words while thinking of a musical context. For example, 'happy' in musical context might have a higher arousal compared to 'happy' in general.

Another issue is the various intensities of emotions. For example, the word 'sad' could be used to describe music that is really sad but in a pleasant way (relatively higher valence), or music that is really sad in a more unpleasant way (relatively lower valence). However, the AV ratings of the word only denote one point in AV space. To account for granularity in the use of emotion words, the standard deviations of the

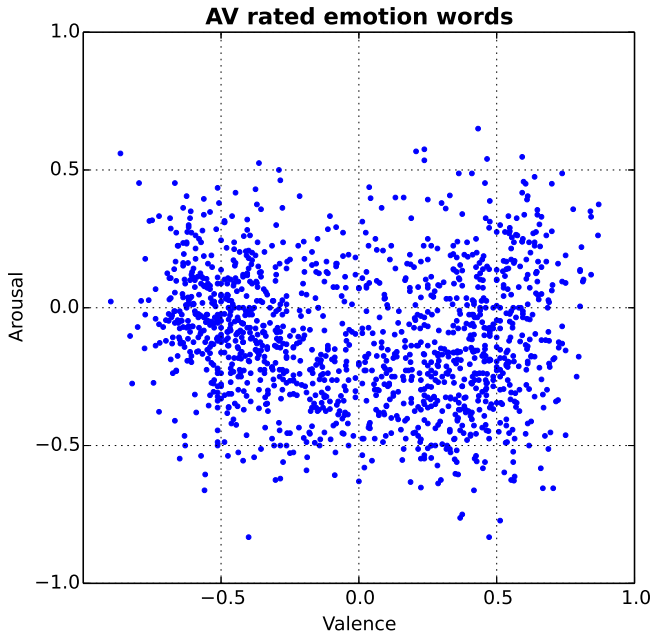


Figure 4.2 AV rated emotion words (1379) of *EmotionWordlist* plotted in AV space. Every dot represents a word.

AV ratings are also included in the dataset. The standard deviations could be used to create an area around the AV point in AV space denoting the area in which the tag is of relevance.

Ambiguity issues for creating the emotion wordlist are already discussed in section 4.1.2. Those are argued not to be extremely problematic since whatever the intended meaning of a word was, if the tag weight is predicted, it is done based on the same AV rating for the same word every time. So while the weighted tags don't discriminate between different possible meanings of the tag (and thereby losing some variety of the tags), the tags are consistently weighted for one specific meaning, which seems to be a fair trade off.

4.4 Conclusion

A dataset with 1379 AV rated emotion words is created containing: words, arousal rating, valence rating, arousal standard deviation and valence standard deviation. Ambiguity and granularity issues are discussed and argued not to be too problematic. The fact that the emotion words are not rated in a musical context is more problematic since there might be a difference between daily use of emotion words compared to use of those words in musical context. Testing whether there is such a difference or obtaining AV ratings for emotion words in musical context is beyond the scope of this thesis and might

a topic for further research. Therefore the current AV rated emotion word dataset is deemed sufficient for being used in the tag weight prediction method in chapter 5.

Chapter 5

Predicting weights of emotion tags

In this chapter tag weights of emotion tags for audio clips will be predicted. Tag weights indicate how relevant a tag is for a certain audio clip. This is useful for e.g. the music industry and social websites.

Based on different features of the emotion tags and the music, regression models are trained and tested to predict the importance of the tags. Two different predictors are employed and compared. (1) The distance between AV values of both the audio clips and emotion tags (collected in chapter 4) are used to predict tag weights. The used AV values of the audio clips are not predicted (see chapter 3) but human generated AV ratings. (2) Audio features of the clips are used to directly predict tag weights.

All the tags used in this chapter are from two datasets (Saari596 and Eerola110, see table 5.1) and are rated by humans on how well they describe the music. These ratings are used as ground truth to evaluate the trained models.

The results show that the AV distance predictors outperform the audio feature predictors. Furthermore, some emotion words seem to be more suitable for this method than others.

5.1 Method

The method consists of training regression models to predict emotion tag relevance for audio clips based on predictors. Here we propose a set of predictors based on the distance between the audio clip position and the emotion tag position in AV space. The current question is whether the relation between the AV values of the clip and AV values of the tags could be used as a reliable predictor of the relevance of the tag.

	Eerola110	**Saari596**
Source	Eerola and Vuoskoski (2010)	Saari and Eerola (2013)
Clips	110	596
Clip length	10–30s.	15–30s.
Music type	Film music	Various popular
Annotators	116 per song	±29 per song
Rated on	valence, energy, tension, anger, fear, happy, sad, tender	valence, arousal, tension, atmospheric, happy, dark, sad, angry, sensual, sentimental

Table 5.1 Annotated datasets used to train and test the tag weight prediction models.

Various variables are employed to test the method and answer the question. In the next section (5.2) the method and its variables are explained in detail and motivated. The results are presented and discussed in section 5.3.

5.2 Pipeline

Figure 5.1 shows the complete pipeline that is used to test the tag weight prediction method. The variables are put between brackets '[ ]' and are explained together with the other parts of the pipeline in the following sections.

The whole pipeline is implemented in Python[1] version 2.7.5. Various functions of the Scikit-Learn[2] package (version 0.15.2) are used to create each step (including the regression models) of the pipeline.

5.2.1 Tags

Tags are either individually [$happy, sad, angry, tension$] or together [$tagset$] trained and tested. There are two reasons for this setup. First, it allows for comparing results of individual tags to find out whether there is a difference between individual tag weight prediction models. For example, it could be that some tags are more or less predictable than others. Second, it indicates whether tag weight prediction improves by training the tags on separate models compared to predicting all tags together with one model.

The individual tags are selected based on the overlapping rated tags in the Saari596 and Eerola110 datasets (see section 5.2.3). It should be noted that in the Eerola110 dataset not the word 'angry' is rated but the word 'anger'.

In the [$tagset$] option a set of tags together is used to train and test the models. In the train/test phase on the Saari596 dataset [$angry, atmospheric, dark, happy, sad,$

[1] `https://www.python.org/`

[2] Scikit-Learn is a well documented, open source Python machine learning toolkit. `http://scikit-learn.org/stable/index.html`

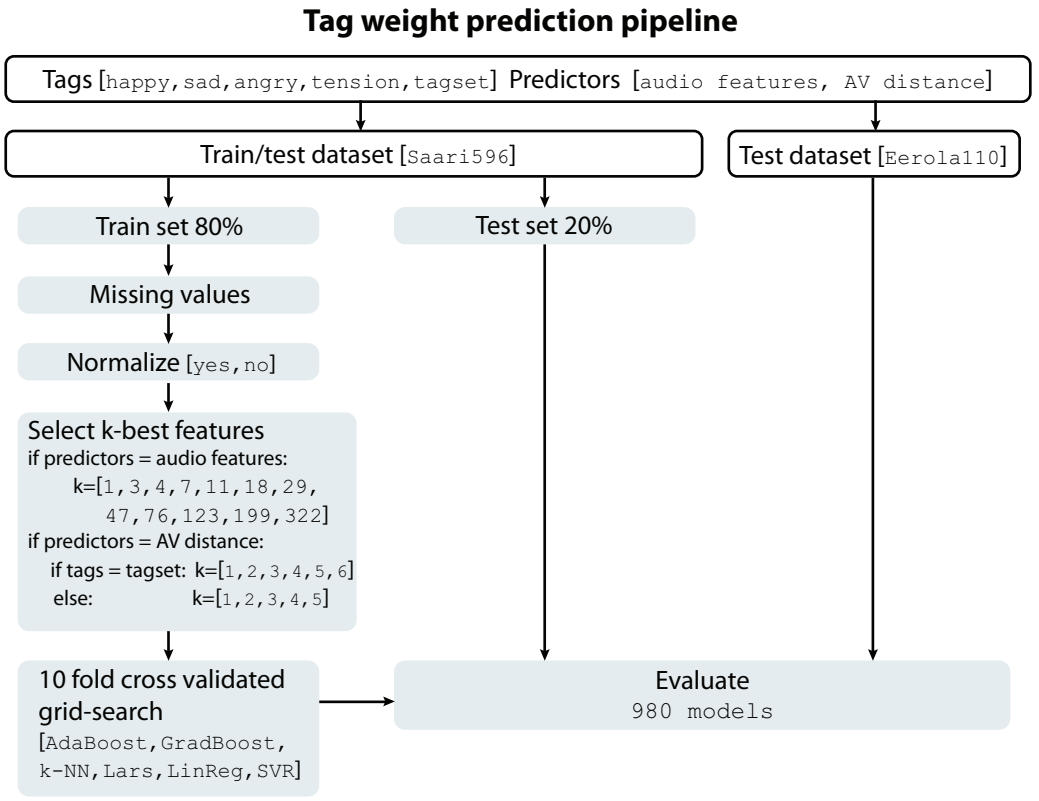


Figure 5.1 Training and testing phases of tag weight prediction pipeline. All unique permutations of the variables between brackets '[ ]' are performed, which results in a total of 980 trained and tested models.

$sensual, sentimental, tension$] are used and in the test phase on the Eerola110 dataset [$angry, happy, sad, tension, fear, tender$] are used. The tags in both sets are selected based on all available emotion tags in both datasets. The difference between the sets should not be problematic since the purpose is to test whether AV distance is a general predictor of emotion tags.

An overview of all the tags is presented in figure 5.2 where the tags are plotted in AV space based on the AV ratings in the Warriner dataset. The green ellipses around the tags indicate the standard deviations of the ratings.

5.2.2 Predictors

Two sets of predictors are created. The first one (audio features) contains 376 audio feature extracted from the audio clips with MIRToolbox 1.5[3] (Lartillot and Toiviainen, 2007). The second predictor set (AV distance) contains six predictors, see table 5.2 and figure 5.3. The results from using both predictor sets will be compared to see if the

[3] `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/mirtoolbox`

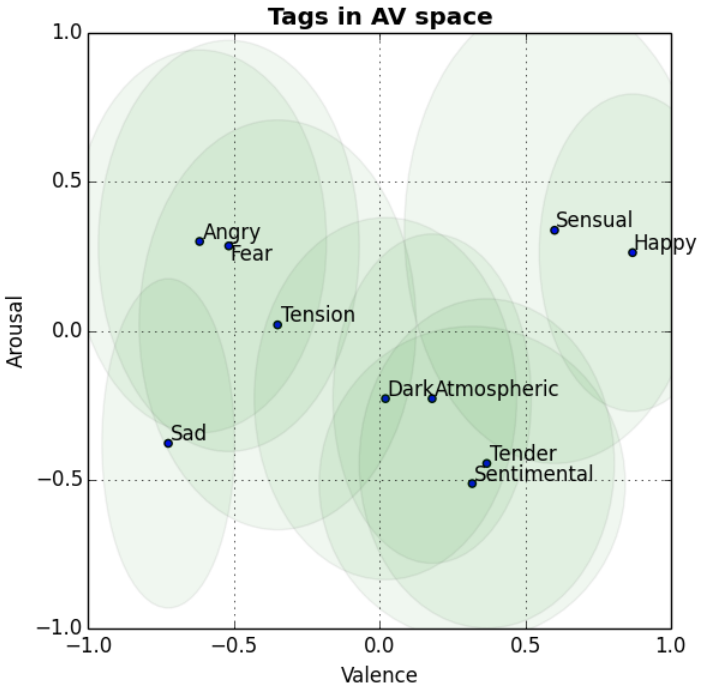


Figure 5.2 All tags used in this chapter plotted in AV space. AV values of the tags are from the Warriner dataset (see chapter4).

newly proposed AV distance predictors outperform the more traditionally used audio features.

Predictor	**Abbreviation**	**Formula**
A rating of the clip	clipA	–
V rating of the clip	clipV	–
AV distance [tag,clip] in AV space	clipTagDist	–
A distance [tag,clip]	arousalDist	–
V distance [tag,clip]	valenceDist	–
Std. of the tag AV rating	stdArea	$std.A * std.V * \pi$

Table 5.2 Six AV distance tag weight predictors. A = arousal, V = valence, Std = standard deviation.

Five of the six AV distance predictors are used to predict the individual tags: [$happy$, $sad, angry, tension$]. A sixth predictor (standard deviation area: stdArea) is included to predict the set of tags [$tagset$]. *StdArea* is the area around a tag and formed by the arousal standard deviation (std.A) of the tag multiplied by the valence standard deviation (std.V) of the tag multiplied by π ($std.A * std.V * \pi$). A large *stdArea* indicates that there is less consensus among the raters for a tag since the ratings were more scattered. Vice versa, a smaller *stdArea* indicates more consensus.

The *stdArea* is useless as predictor for individual tags since it will be just one value. When predicting multiple tags together it makes more sense to add this sixth predictor

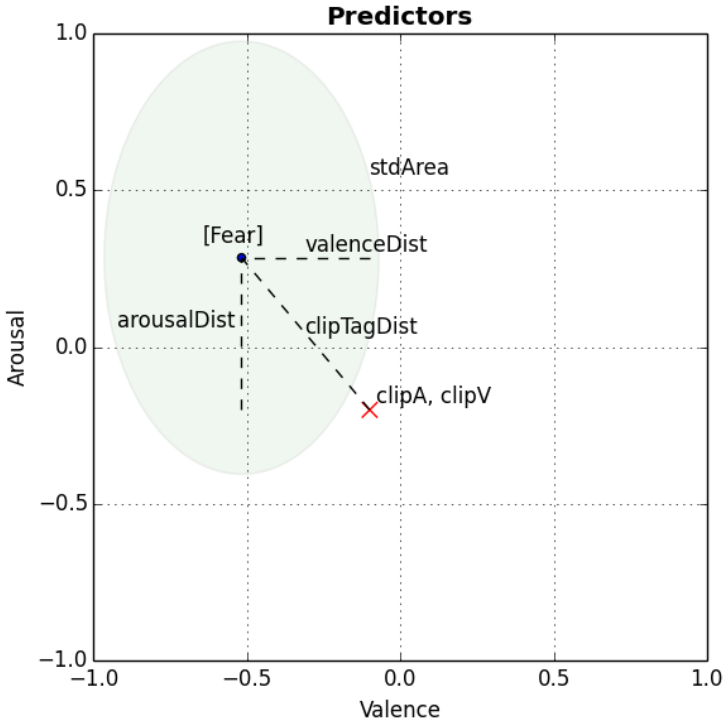


Figure 5.3 Six AV distance tag weight predictors for the tag 'fear' and an example clip (red cross). The green area around the tag (stdArea) is based on the AV standard deviations of the tag rating from the Warriner dataset (see chapter 4).

since the value differs for different tags.

For creating the AV distance predictors, true clip AV values instead of predicted are used. The trained and tested AV prediction models from chapter 3, especially the valence models, were not robust enough to get reliable AV predictions. Testing the tag weight method with unreliable AV predictions makes it difficult to properly evaluate the proposed method. Using the human rated AV values of the audio clips makes it possible to purely test the method in an optimal situation.

5.2.3 Datasets

For testing this method, two audio datasets and one dataset of emotion words are used. The audio datasets are Saari596 and the Eerola110, which contain respectively 596 and 110 audio clips. Those audio clips are rated on various scales, see table 5.3.

The emotion word dataset (see chapter 4) contains emotion words with AV ratings. From this dataset the AV ratings of the emotion tags described in section 5.2.1 are obtained.

All AV ratings are normalized between [-1,1] and all ratings on the emotion words are normalized between [0,1].

Dataset	Clips	Rated on
Saari596	596	valence, arousal, tension, atmospheric, happy, dark, sad, angry, sensual, sentimental
Eerola110	110	valence, energy*, tension, anger, fear, happy, sad, tender

Table 5.3 Two rated audio clip datasets. * = energy is similar to arousal (Spindler, 2009; Yang and Chen, 2012).

5.2.4 Missing values

This option is only used for the audio features. Missing values in the feature sets are imputed with the mean of that feature. If all values of a feature were missing, they will all get a 0. A situation where all values of a feature were missing did not occur often, and might have been caused by some problems in the MIRToolbox. Missing values are imputed on each train and test set separately so that no information of the train set would be used in the testing phase.

5.2.5 Normalize

The predictors are either normalized or not on all three predictor sets. The only reason for this variable is to find the optimal models. Normalization is performed separately on each of the three feature sets, otherwise information from the training phase will be used in the test phases.

5.2.6 Feature selection

The k-best feature selection process is a bit complicated due to the different predictors and tags. If the audio features are used as predictor, then $k = [1, 3, 4, 7, 11, 18, 29, 47, 76, 123, 199, 322]$. If the AV distances are used as predictor and the tags are individually trained and tested, then $k = [1, 2, 3, 4, 5]$. If the AV distances are used for the set of tags, then $k = [1, 2, 3, 4, 5, 6]$.

The k-best features are selected in the training phase and the same features are used during both testing phases. The best features are selected based on F-values between predictor and truth value. For example if $k = 3$, then the 3 predictors with the highest F-values are used.

Correlation coefficients of the AV distance features with the tag ratings are presented in table 5.4. The negative correlations are due to the fact that a large distance (ideally) means a low tag weight.

Tag	**1 (best)**	r	**2**	r	**3**	r
Angry	clipTagDist	−0.78	clipV	−0.76	arousalDist	−0.74
Happy	arousalDist	−0.91	clipV	0.91	clipTagDist	−0.86
Sad	clipTagDist	−0.75	clipA	−0.66	valenceDist	−0.62
Tension	clipA	0.69	clipV	−0.69	arousalDist	−0.6
Tag Set	clipTagDist	−0.48	arousalDist	−0.45	valenceDist	−0.22

4	r	**5**	r	**6 (worst)**	r
clipA	0.61	valenceDist	−0.35	–	–
valenceDist	−0.16	clipA	0.14	–	–
arousalDist	−0.35	clipV	−0.35	–	–
clipTagDist	−0.45	valenceDist	−0.03	–	–
clipV	−0.13	tagStdArea	−0.13	clipA	−0.12

Table 5.4 For each tag the five or six AV distance predictors ordered from most relevant (1) to least relevant (5 or 6) are presented. Correlation coefficient r indicates how well the predictor correlates with the tag rating. All r values have $p < .001$, except Tension/valenceDist with $p < .01$. The meaning of the predictor abbreviations can be found in table 5.2.

5.2.7 Regression models

Five different models $[AdaBoost, GradBoost, k-NN, LinReg, SVR]$ are employed in a 10-fold cross validated grid search in which various parameter settings (varying per model) are tested and optimized. The best parameter setting per model is based on the R^2 score.

5.2.8 Evalution

The 'best' cross-validated parameter settings for each model is used to predict tag weights for the left out 20% of the Saari596 dataset and for the full Eerola110 dataset. The predictions are evaluated with the human ratings of the tags. In total 980 models are trained and tested.

5.3 Results

The main results are presented in figure 5.4, containing the best robust models per tag on both sets of predictors (AV distance and Audio features) and on both datasets: Saari596 (training and testing) and Eerola110 (only testing). Table 5.5 also includes the names of the models and amount of selected features. The criteria for robustness is the 'least lowest' R^2 score (also used in chapter 3). For example, there are various models trained and tested to predict 'angry' with AV distance predictors. For each model there are two R^2 scores (i.e. for Saari595 and Eerola110). The model that has the 'least lowest'

R^2 score is the most robust, because that one has the best minimum capabilities of the models.

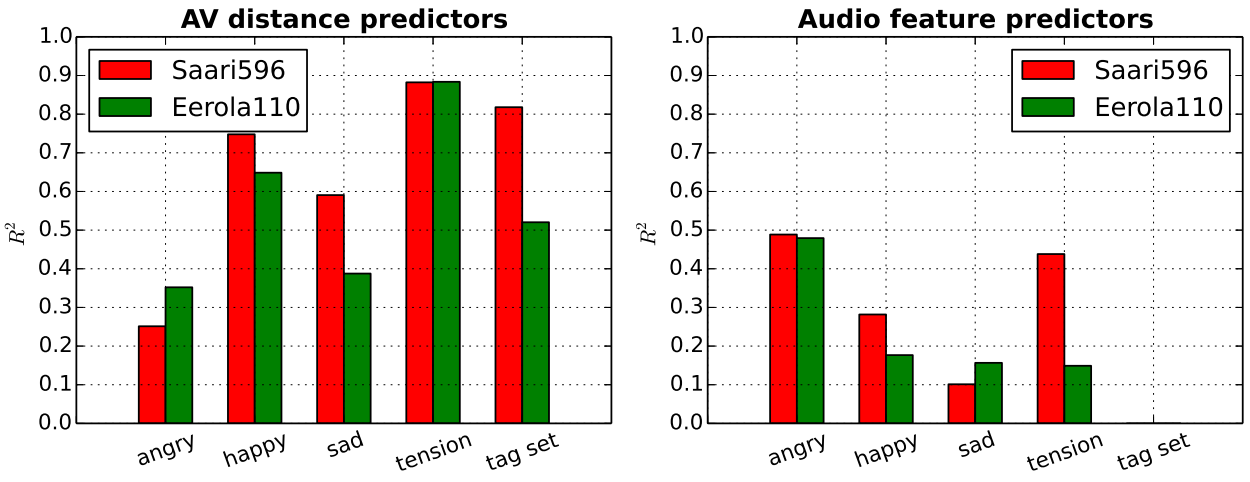


Figure 5.4 Best robust models for the Saari596 and Eerola110 test set on four individual tags and a set of tags.

Tag	R^2 Saari596	R^2 Eerola110	Model	Features
AV distance predictors				
Angry	0.25	0.35	GradBoost	1
Happy	0.75	0.65	GradBoost	5
Sad	0.59	0.39	k-NN	5
Tension	0.88	0.88	LinReg	5
Tag Set	0.82	0.52	k-NN	6
Audio feature predictors				
Angry	0.49	0.48	GradBoost	18
Happy	0.28	0.18	GradBoost	123
Sad	0.1	0.16	k-NN	29
Tension	0.44	0.15	AdaBoost	47
Tag Set	−0.05	−0.06	k-NN	18

Table 5.5 Best robust models for the Saari596 and Eerola110 test set on four individual tags and a set of tags. Name of the regression models and the amount of features are also included.

The results are discussed in the following subsections. First the results of the individual tags and the set of tags will be evaluated, second, the differences between the two predictor sets will be discussed. Third, the robustness of the models will be evaluated. In addition, the learning curves for each model and the best models per tag (set) are presented in appendix C.

5.3.1 Individual tags and tag set

For the four individual trained tags, 'tension' clearly stands out with $R^2 = .88$ as best robust result. Also performing well is 'happy' with $R^2 = .65$. 'Angry' and 'sad' have respectively $R^2 = .48$ and $R^2 = .39$ as best robust result.

The high R^2 for 'tension' could be explained based on the literature. In the current study a 2D emotion space (arousal/valence) is used, however some studies suggest a third dimension: dominance or tension. Yet it is argued that this third dimension does not add significant explanatory power to the emotion model because this third dimension is highly correlated with valence (Eerola et al., 2009; Warriner et al., 2013) (see also chapter 2, section 2.1.3.2). Here we use arousal and valence information to predict the tag weights, so a high correlation of arousal and/or valence with a tag will likely result in a good prediction.

Another standing out result is that 'angry' is better predicted with the audio features while the five AV distance predictors show correlations higher than 'tension', see table 5.4. This can be explained by the fact that only the best robust models are presented. When looking at the results of all 'angry' models, it shows various models trained with the AV distance predictors that score $R^2 > .7$ on the Saari596 test set, but $R^2 < .01$ on Eerola110. So the 'good' predictor correlation might only hold for the Saari596 dataset and not for the Eerola110 set. A reason for this might be that the Eerola110 dataset is actually rated for 'anger' instead of 'angry', making the datasets less compatible for this tag. Another reason could be that Eerola110 contains film music soundtracks which have a very specific way of expressing anger compared to regular pop music as in Saari596.

The results for the tag sets are not bad for the AV distance predictors with R^2=.52 since it performs better than 'angry' and 'sad'. The tag set model performs not strikingly different from the individual tag models. However, when the audio feature predictors are used, the tag set model preforms worse than the individual tag models with $R^2 < .0$.

The individual tag results suggest that some tags are better predicted then others and also with different models. Further elaboration on the results for each model on each tag can be found in appendix C.

5.3.2 Predictors: AV distance vs. Audio features

Two different predictor sets (AV distance and Audio features) are used to evaluate which one is better suitable for predicting tag weights. It is immediately visible from figure 5.4 that the AV distance predictor outperform the audio features for 'happy', 'sad', 'tension' and 'all', only the tag 'angry' is predicted better with the audio features.

From these results it is concluded that AV distance predictors are generally better suitable for predicting tag weights compared to audio features. However, this might not

be the case for all tags, e.g. 'angry'. Combining both predictor sets might lead to better results.

More information about the learning curves for both predictor sets and amounts of k-best features can be found in appendix C.

5.3.3 Robustness of the models

One way to test the robustness of trained regression models is to evaluate them on unseen data (i.e./ not used during training) and preferably on another dataset. If a model also performs good on other datasets it shows that the model is more generalizable. However it does not rule out the possibility that the model is based on confounding factors since it is possible that by chance the same confounding factors are present in all test sets.

The models are trained on 80% of the Saari596 dataset, the other 20% is used for testing and not seen during training. Thereby, another dataset, Eerola110, is also used for testing. Figure 5.4 and table 5.5 show the R^2 scores of the best robust models.

5.4 Conclusion

The main goal of this chapter was to test a new method for predicting emotion tag weights. To this end a pipeline with various variables was created to train and test regression models.

The results show that the AV distance predictors are better suited for predicting emotion tag weights compared to audio features. Furthermore, individual trained models for each tag do not largely outperform models that are trained on a set of tags. At last, the models are tested for robustness on another dataset. While this does not necessarily exclude the possibility of confounding factors, it makes the models more reliable.

The main conclusion is that the proposed method shows a promising direction into predicting tag weights. The main obstacle to utilize this method is that we used human AV ratings of the audio clips to test the potential of the method. Creating sufficiently reliable AV prediction models for audio clips is still a huge challenge, see also chapter 3.

The next step will be to test if the currently found results can be repeated for other emotion tags individually or in sets. Furthermore, the AV distance predictors might gain in predictive power when combined with other predictors. Another idea that could be tested is to find out whether there are more basic emotion tags that could be used to predict other more complex emotion tags, just like arousal and valence are used here to predict emotion tags. That way a dimensional approach would be combined with a categorical approach.

Chapter 6

Data collection and method evaluation

This chapter describes the process of collecting human data to evaluate the tag weight prediction method. The reason for this experiment is to test the method on real world data, i.e. tags assigned by Last.fm users. The main difference with the evaluation in chapter 5 is that in the current chapter 79 unique tags are used in contrast to the small tag set in chapter 5.

The first goal is to create a dataset containing short audio clips with emotion words that are rated for how well they describe the music. The second goal is to use this data to evaluate the tag weight prediction method.

Due to time limitations it was decided to keep the experiment small resulting in 49 audio clips with each five tags which are rated by 27 participants. Unfortunately, the ratings of the words are scattered in such a way that the results are not sufficient for properly testing the tag weight prediction method. An evaluation of the method is nevertheless conducted with, as expected, inconclusive results. In the discussion it is discussed what caused the low quality of the dataset and what could have improved it.

6.1 Data collection method

The objective is to collect human data on how well (i.e. average ratings of the participants) emotion words described music clips. Specifically perceived emotions[1] are targeted since they are less subjective. In the experiment short audio clips are presented with 5 emotion words per clip that have to be rated on a five point scale for how well they describe the music.

[1]See chapter 2, section 2.2.1.1, for the difference between felt and perceived emotions.

6.1.1 Experimental set-up

The clips used in the experiment were selected from the Saari596 dataset because every clip in that dataset contains tags which are originally assigned by last.fm users. These tags provide a perfect example of noisy tag sets for which the tag weight prediction method is designed.

The 49 clips with the most tags were selected from the Saari596 dataset, see figure 6.1. For each clip 5 tags from its tag set were randomly selected after being filtered for ambiguous tags, e.g. chill and slow[2]. This resulted in $49 * 5 = 245$ tags to be rated, which was estimated to take about 20 minutes per person at a normal pace without breaks.

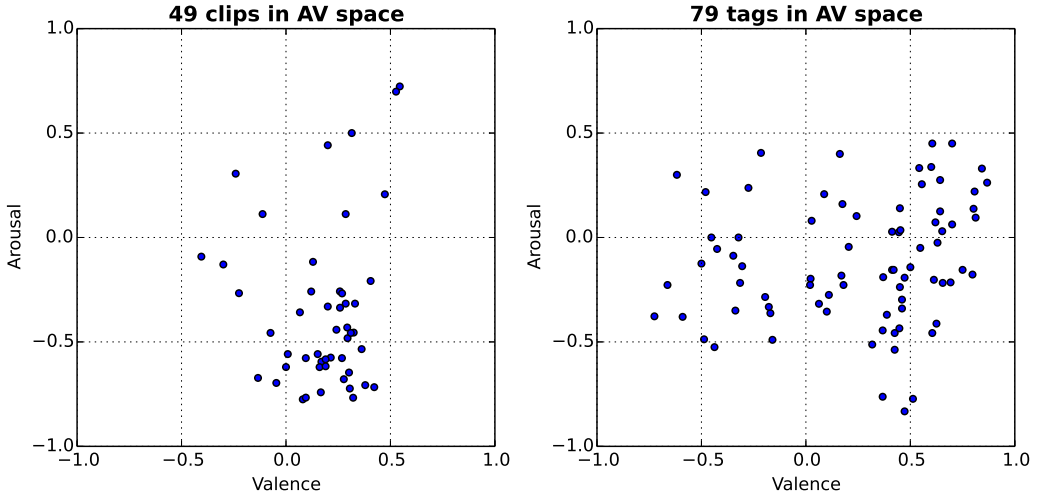


Figure 6.1 49 clips and 79 unique tags in AV space.

Since the tags were selected randomly some tags occurred more often than others. In total there are 79 unique tags of which more than half only occurs once or twice. 'Sad' is the most occurring word (18 times). The selected clips and tags are presented in appendix D. During the experiment the clips were presented at random order.

Due to the fact that most participants would be native Dutch speakers it was expected that not all words would be properly understood by everyone. To decrease the chance of rating words that are not understood, two measures were taken. First, clicking the words redirected the participant to a google translate page with the translation(s) of the word in Dutch. Second, there is an option "I don't understand the word" that enabled the participant to skip the word.

[2]Chill could be understood as shiver or as relaxed. Slow could describe an emotional state but also describe the tempo of the music.

6.1.2 Implementation

A forced choice online questionnaire that collected user responses and saved them to csv files was created using HTML, PHP and CSS. Figure 6.2 (right) shows the interface. The advantage of using an online questionnaire is that participants could do the experiment at any location with internet.

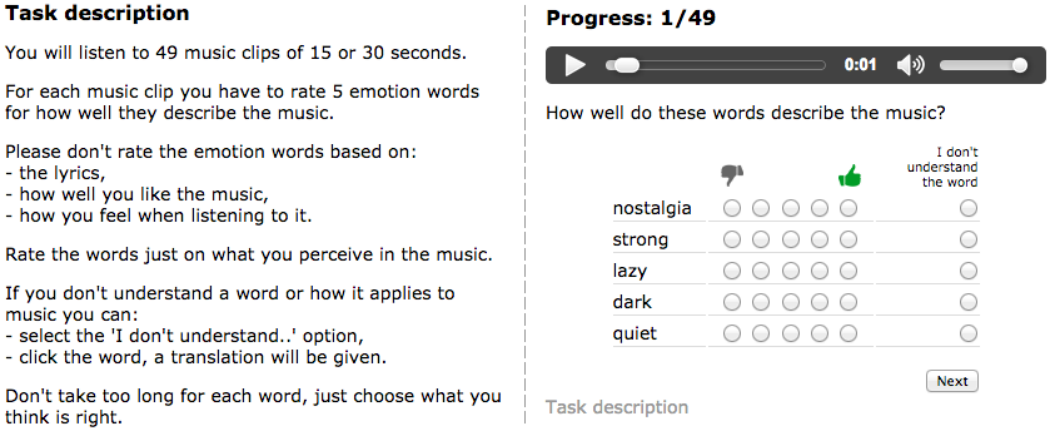


Figure 6.2 Task description (left) and user interface (right) for rating emotion words.

6.1.3 Task description

In the task description it is explained that the emotion words need to be rated on perceived emotions: 'how well they describe the music' and 'what you perceive in the music' (see figure 6.2, left). During test trials it turned out that people were influenced by the lyrics when rating the words. Since the tag weighting method does not use lyrical information, it was decided to explicitly inform the participants to ignore the lyrics. Furthermore the translation option (by clicking on the word), and what to do when a word is not understood is explained.

Additionally, the full task description also appeared when hovering the 'task description' text at the bottom left of the tag rating interface screen.

6.1.4 Participants

The participants did not have to meet special requirements. A total of 28 people participated of which 14 females and 14 males with ages ranging from 19 to 64 years and a mean of 30 years.

6.2 Results

Three participants didn't finish the experiment for unknown reasons, however the parts that they did fill out are still used. The results of one participant were disregarded based

on the feedback he gave afterwards. It was clear that he did not understood how to rate the words and used the rating scale in opposite manner for negative words.

In total 71 times a word was skipped with the 'I don't understand' button. Only two subjects used this option more then 10 times. Spread over all clips, 'soothing' was most often skipped with 7 times. 'Lush' and 'atmospheric' were both 6 times skipped, suggesting that those words were difficult to understand.

Krippendorff's alpha is calculated[3] to assess the agreement among the participants. Krippendorff's alpha is used because the data contains missing values (not all participants rated all tags). When the participants agree perfectly, $\alpha = 1$, indicating a perfect reliability and $\alpha = 0$ when the results are produced as if by chance, indicating an absence of reliability (Krippendorff, 2011).

The result of the Krippendorff's alpha (calculated on an interval scale) is $\alpha = 0.30$, which indicates a very low agreement. The low alpha value is not unexpected due to the subjective nature of the experiment and small test group. It also begs the question whether the dataset is reliable enough to test the method, which will be discussed in the discussion (section 6.4).

6.3 Tag weight prediction method evaluation

The collected data is used to evaluate the tag weight prediction method. For the evaluation the best robust method from chapter 5 is used: k-NN, non normalized data and 6 AV distance predictors. With this model tag weights are predicted for 245 tags from the experiment.

Figure 6.3 shows the tag weight predictions (x-axis) and human rated tag weights (y-axis). The explained variance is extremely low, $R^2 = -1.09$. Also correlation is low $r = .21$ with $p < .01$. The human ratings and tag weight predictions for each tag are presented in appendix D (table D.2).

It is however impossible to conclude anything about the viability of the tag weight prediction method. The Krippendorff's alpha suggests that the collected data is not reliable, so any bad results could be produced by the data or by the model, or even other factors. In any case, further testing with the currently collected data is pointless. Results and improvements will be discussed in the next section.

6.4 Discussion

First the feedback of the participants of the experiment will be discussed, second the low Krippendorff's alpha on the results will be discussed. Third improvements to the

[3]The following python script was used to calculate Krippendorff's alpha: `http://grrrr.org/data/dev/krippendorff_alpha/krippendorff_alpha.py` (7 august 2015)

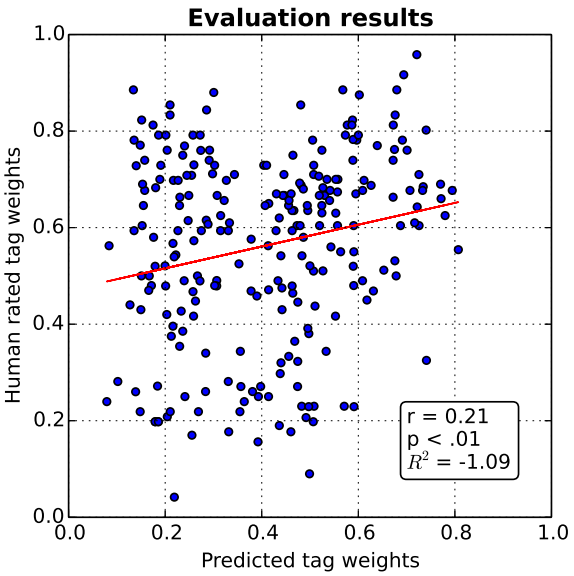


Figure 6.3 Final result

current experiment will be suggested.

6.4.1 User comments

Feedback of the participants seems to corroborate the low Krippendorff's alpha. It is often reported that it was very hard to discriminate between felt and perceived emotions when rating the words especially for music they already knew. Also some words were ill understood, and even less understood for how they could be related to the music, e.g. 'soothing'. This might be due to the fact that most participants were non-native english speakers. One (the only) native english participant did not report any trouble in understanding the words or applying them to the music. Two other points that were commonly heard in the feedback were that it was hard to ignore the lyrics and that some words could be interpreted in various ways, e.g. 'content'.

6.4.2 Krippendorff's alpha

To better interpret the low Krippendorff's alpha, agreement results of two other studies that conducted a similar experiment are discussed. In Saari and Eerola (2013) Cronbach's alpha values of at least $\alpha = .84$ were obtained in a listening experiment that collected ratings of the emotion words of the original Saari596 dataset. A difference with the current study is that they used 59 participants of which more than half was a musician or trained professional. Another difference is that the clips were only rated on 10 scales.

In another study (Soleymani et al., 2013), where the AV ratings of a part of the MediaEval1744 dataset is obtained, ordinal scale Krippendorff's alpha of the ratings

was $\alpha = .32$ for valence and $\alpha = .35$ for arousal rated clip. These results are accepted as in fair range of agreement. Furthermore, each song is rated by at least 10 subjects.

While the interpretation of the Cronbach's alpha in the first study seems fair, it is unclear why this is the case in the second study. Their Krippendorff's alpha is almost as low as in the current study and should indicate that the data is not reliable.

It is assumed that the currently collected data is insufficient for evaluating the tag weight prediction method. However, since in another study similar alpha results are accepted, tag weights are predicted for the 245 tags and evaluated with the collected data (see section 6.3). The results are inconclusive and it is thereby impossible to determine whether this is caused by the dataset or by the method.

6.4.3 Improvements

This experiment could benefit from various adjustments which are listed below:

- A training phase for the experiment to be more certain that participants understand the task.
- A short description of the words could be provided to make sure that participants properly understand the words and to disambiguate between different word meanings. This could also help non-native English speakers.
- Instead of a five point scale to rate the tags, a slider could have been used. A slider gives a much higher rating granularity which more accurately describes the relevance of a tag. Thereby, it might also be more intuitive for the participants instead of forcing them to choose between five options.
- The selection of the clips was based on the amount of Last.fm tags. The 49 clips with the most tags were selected for this experiment. As can be seen in figure 6.1, most clips have a low arousal and moderate valence. It would have been better to select clips based on their AV rating so that they would be evenly distributed in AV space.
- For each clip 5 tags were randomly selected from its Last.fm tag set for the experiment. While the intention was to resemble a noisy tag set, this created the problem that there was no control group. For example, adding clearly unfitting and clearly fitting tags to the tag sets would make it possible to asses the quality of the ratings.
- In addition to the previous point, a lower variety of tags could lead to more sophisticated results. In the current experiment 79 of the 245 tags were unique whereby each unique word occurred 3.1 times on average. Using a lower variety of tags would also make it possible to calculate the Krippendorff's alpha per tag.

- Another experimental set-up could also be used. Instead of letting people rate how well a tag describes music clips, it could be asked to evaluate the tag weights predicted by the proposed method. This is a more passive task and doesn't require actively assessing to what extend a tag fits the music. That way people directly asses the quality of the predictions.

6.5 Conclusion

Main conclusion is that the evaluation of the tag weight prediction method with the data collected in this chapter is inconclusive. The main reason is that the collected data is not sufficient with an agreement of $\alpha = .3$. Various factors why the collected data is not sufficient are discussed. Improvements such as a training phase, word description and a rating slider are suggested. More importantly, the current selection procedure of the tags and clips allowed for many uncontrolled variables that could influence the results.

Chapter 7

Conclusion

The main conclusions of this thesis are presented in section 7.1. Future work and improvements are provided in section 7.2.

7.1 Conclusion

The main objective of this research was to test a new tag weight prediction method, which could be used to filter emotion tags of music. To this end, arousal/valence (AV) prediction models for music are trained and tested in chapter 3. However, the models were deemed insufficient for predicting AV values based on evaluation results (see section 7.1.1), so human AV ratings of the clips are further used. The AV values of audio clips are combined with AV ratings of emotion words, which are described in chapter 4, to create predictors for predicting tag weights in chapter 5. The tag weight prediction method is evaluated with a small set of tags in chapter 5 and a large set of tags with self collected ground truths in chapter 6 (see section 7.1.2).

To answer the main question: *Can tag weights be predicted based on arousal and valence information of the tag and clip?*, AV information of tags and clips can be used to predict weights of emotion tags. However, the method is only tested with human generated AV values of the audio clips. Fully implementing the method would require better AV prediction models which not yet exist. Furthermore, the method is only tested on a small set of tags for which the results are promising, but it has yet to be determined whether those results generalizes to other tags.

7.1.1 AV prediction

Originally, the AV prediction models from chapter 3 were intended to be used to predict AV values for the tag weight prediction method. However, the AV predictions models from chapter 3 turned out to be not reliable enough. The results from training and testing AV prediction models were nonetheless quite interesting.

The currently employed method for predicting AV values is based on common methods in MER and extended with extensive grid searches and evaluations on various datasets, which is rarely done. The evaluations on other datasets showed that only the arousal models are moderately robust.

The naive valence models, trained and tested on each of the three datasets separately, reached promising R^2 scores (.49, .33 and .72 on respectively MediaEval1744, Saari596 and Eerola360), which are similar to the results of the studies the datasets are obtained from. However, these naive models were unable to predict AV values of the two other, not seen during training phase, datasets, with $R^2 < .13$ in all cases. When comparing the results of all the trained valence models on the three datasets, it showed that there are no models that are able to sufficiently predict valence values on all three datasets. Even the best robust models scored $R^2 < .11$ on all datasets.

These findings indicate that especially valence prediction models are unable to generalize over other datasets. The finding that arousal is better predicted compared to valence conforms findings of other studies, e.g in Saari and Eerola (2013); Eerola and Vuoskoski (2010), and is therefore not unexpected.

Main conclusion from this is that AV models show decreased performance on datasets that are not used for training, especially for valence models. One explanation could be that the currently used predictors are insufficient to capture the characteristics of valence and, to a lesser extend, arousal. Another explanation could be that the datasets are differently created (music selection and gathering of AV values) in a way that the models are unable to generalize over them.

Which explanation is right (or maybe both), is impossible to say based on the current research. It is however clear that there are unknown confounding factors that influence the results (Sturm, 2013b,a) and that methodology for predicting AV values is still in development phase and not yet ready for reliably predicting AV values.

7.1.2 Tag weight prediction

The tag weight prediction models are trained and tested on one dataset and tested on another dataset (see chapter 5). The best results on both datasets indicate that the AV distance predictors outperform the audio feature predictors. Furthermore, some tags are better predicted than others, e.g. tension: $R^2 = .88$ vs. angry: $R^2 = .35$. In addition to predicting weights of the tags individually, the tag weights are also predicted for a small set of tags together. Results show that the tag set prediction is not better or worse compared to the individual tags.

These promising results indicate that the proposed method is viable at least for some tags. A second evaluation of the method in chapter 6 was inconclusive due to low agreement on the collected data, making it impossible to properly evaluate the method.

7.2 Future work

There are various next steps that can be taken from the current work. A dataset with audio clips, preferably from different sources to test robustness, could be rated for a larger set of emotion words. The proposed tag weight prediction method can then be more extensively tested. Thereby, the relation between music related emotion words can be studied.

Emotion prediction based on commonly used audio features such as can be found in MIRToolbox 1.5 seem to be insufficient and lead to non-generalizable models. Therefore, researchers should devote their full attention to finding new predictors. One direction to improve emotion prediction could be to first development good higher level feature recognition systems, e.g. for melody (Panda et al., 2013), harmony, chords, tonality and tempo. These features might be better compared to low level features at predicting emotion since they encompass important, but not often investigated, cues for emotion recognition (Gabrielsson and Lindström, 2010).

Using other datasets to test the robustness and generalizability of the emotion prediction models seems to be a good practice, however it should be investigated to what extent different datasets could be used for this purpose. More specifically, what are the criteria for such datasets? And what kind of performance should be expected from emotion prediction models on other datasets?

The currently used dataset with AV rated words could be improved by having the words rated in a musical context. With this data it could also be studied whether there is a difference between words rated in a normal context and in a musical context.

And at last, instead of tag filtering, the proposed tag weighting method could be used to suggest tags for music or auto-tagging. The method is not specifically tested for this purpose, however the difference between tag filtering and tag recommendation is more a semantical than a practical one since tag weights could be used for both. One difference is that noisy tag sets are assumed to have at least some correct tags (depending on the noisiness), while in tag recommendation the music typically doesn't have any tags. Auto-tagging would target the 'cold start' problem in which newly created music is not (yet) tagged on social websites.

Appendix A

Extracted features with MIRToolbox

Table A shows all the features that are extracted with MIRToolbox 1.5[1] (Lartillot and Toiviainen, 2007) and which are used to train and test the AV models.

Table A.1 All music features extracted with MIRToolbox.

dynamics rms Mean	dynamics rms Std
dynamics rms Slope	dynamics rms PeriodFreq
dynamics rms PeriodAmp	dynamics rms PeriodEntropy
fluctuation peak PeakPosMean	fluctuation peak PeakMagMean
fluctuation centroid Mean	rhythm tempo Mean
rhythm tempo Std	rhythm tempo Slope
rhythm tempo PeriodFreq	rhythm tempo PeriodAmp
rhythm tempo PeriodEntropy	rhythm attack time Mean
rhythm attack time Std	rhythm attack time Slope
rhythm attack time PeriodFreq	rhythm attack time PeriodAmp
rhythm attack time PeriodEntropy	rhythm attack slope Mean
rhythm attack slope Std	rhythm attack slope Slope
rhythm attack slope PeriodFreq	rhythm attack slope PeriodAmp
rhythm attack slope PeriodEntropy	spectral centroid Mean
spectral centroid Std	spectral centroid Slope
spectral centroid PeriodFreq	spectral centroid PeriodAmp
spectral centroid PeriodEntropy	spectral brightness Mean
spectral brightness Std	spectral brightness Slope
spectral brightness PeriodFreq	spectral brightness PeriodAmp
spectral brightness PeriodEntropy	spectral spread Mean
spectral spread Std	spectral spread Slope

Continued on next page

[1] `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/mirtoolbox`

Table A.1 – *Continued from previous page*

spectral spread PeriodFreq	spectral spread PeriodAmp
spectral spread PeriodEntropy	spectral skewness Mean
spectral skewness Std	spectral skewness Slope
spectral skewness PeriodFreq	spectral skewness PeriodAmp
spectral skewness PeriodEntropy	spectral kurtosis Mean
spectral kurtosis Std	spectral kurtosis Slope
spectral kurtosis PeriodFreq	spectral kurtosis PeriodAmp
spectral kurtosis PeriodEntropy	spectral rolloff95 Mean
spectral rolloff95 Std	spectral rolloff95 Slope
spectral rolloff95 PeriodFreq	spectral rolloff95 PeriodAmp
spectral rolloff95 PeriodEntropy	spectral rolloff85 Mean
spectral rolloff85 Std	spectral rolloff85 Slope
spectral rolloff85 PeriodFreq	spectral rolloff85 PeriodAmp
spectral rolloff85 PeriodEntropy	spectral spectentropy Mean
spectral spectentropy Std	spectral spectentropy Slope
spectral spectentropy PeriodFreq	spectral spectentropy PeriodAmp
spectral spectentropy PeriodEntropy	spectral flatness Mean
spectral flatness Std	spectral flatness Slope
spectral flatness PeriodFreq	spectral flatness PeriodAmp
spectral flatness PeriodEntropy	spectral roughness Mean
spectral roughness Std	spectral roughness Slope
spectral roughness PeriodFreq	spectral roughness PeriodAmp
spectral roughness PeriodEntropy	spectral irregularity Mean
spectral irregularity Std	spectral irregularity Slope
spectral irregularity PeriodFreq	spectral irregularity PeriodAmp
spectral irregularity PeriodEntropy	spectral mfcc Mean 1
spectral mfcc Mean 2	spectral mfcc Mean 3
spectral mfcc Mean 4	spectral mfcc Mean 5
spectral mfcc Mean 6	spectral mfcc Mean 7
spectral mfcc Mean 8	spectral mfcc Mean 9
spectral mfcc Mean 10	spectral mfcc Mean 11
spectral mfcc Mean 12	spectral mfcc Mean 13
spectral mfcc Std 1	spectral mfcc Std 2
spectral mfcc Std 3	spectral mfcc Std 4
spectral mfcc Std 5	spectral mfcc Std 6
spectral mfcc Std 7	spectral mfcc Std 8
spectral mfcc Std 9	spectral mfcc Std 10
spectral mfcc Std 11	spectral mfcc Std 12
spectral mfcc Std 13	spectral mfcc Slope 1
spectral mfcc Slope 2	spectral mfcc Slope 3
spectral mfcc Slope 4	spectral mfcc Slope 5
spectral mfcc Slope 6	spectral mfcc Slope 7

Continued on next page

Table A.1 – *Continued from previous page*

spectral mfcc Slope 8	spectral mfcc Slope 9
spectral mfcc Slope 10	spectral mfcc Slope 11
spectral mfcc Slope 12	spectral mfcc Slope 13
spectral mfcc PeriodFreq 1	spectral mfcc PeriodFreq 2
spectral mfcc PeriodFreq 3	spectral mfcc PeriodFreq 4
spectral mfcc PeriodFreq 5	spectral mfcc PeriodFreq 6
spectral mfcc PeriodFreq 7	spectral mfcc PeriodFreq 8
spectral mfcc PeriodFreq 9	spectral mfcc PeriodFreq 10
spectral mfcc PeriodFreq 11	spectral mfcc PeriodFreq 12
spectral mfcc PeriodFreq 13	spectral mfcc PeriodAmp 1
spectral mfcc PeriodAmp 2	spectral mfcc PeriodAmp 3
spectral mfcc PeriodAmp 4	spectral mfcc PeriodAmp 5
spectral mfcc PeriodAmp 6	spectral mfcc PeriodAmp 7
spectral mfcc PeriodAmp 8	spectral mfcc PeriodAmp 9
spectral mfcc PeriodAmp 10	spectral mfcc PeriodAmp 11
spectral mfcc PeriodAmp 12	spectral mfcc PeriodAmp 13
spectral mfcc PeriodEntropy 1	spectral mfcc PeriodEntropy 2
spectral mfcc PeriodEntropy 3	spectral mfcc PeriodEntropy 4
spectral mfcc PeriodEntropy 5	spectral mfcc PeriodEntropy 6
spectral mfcc PeriodEntropy 7	spectral mfcc PeriodEntropy 8
spectral mfcc PeriodEntropy 9	spectral mfcc PeriodEntropy 10
spectral mfcc PeriodEntropy 11	spectral mfcc PeriodEntropy 12
spectral mfcc PeriodEntropy 13	spectral dmfcc Mean 1
spectral dmfcc Mean 2	spectral dmfcc Mean 3
spectral dmfcc Mean 4	spectral dmfcc Mean 5
spectral dmfcc Mean 6	spectral dmfcc Mean 7
spectral dmfcc Mean 8	spectral dmfcc Mean 9
spectral dmfcc Mean 10	spectral dmfcc Mean 11
spectral dmfcc Mean 12	spectral dmfcc Mean 13
spectral dmfcc Std 1	spectral dmfcc Std 2
spectral dmfcc Std 3	spectral dmfcc Std 4
spectral dmfcc Std 5	spectral dmfcc Std 6
spectral dmfcc Std 7	spectral dmfcc Std 8
spectral dmfcc Std 9	spectral dmfcc Std 10
spectral dmfcc Std 11	spectral dmfcc Std 12
spectral dmfcc Std 13	spectral dmfcc Slope 1
spectral dmfcc Slope 2	spectral dmfcc Slope 3
spectral dmfcc Slope 4	spectral dmfcc Slope 5
spectral dmfcc Slope 6	spectral dmfcc Slope 7
spectral dmfcc Slope 8	spectral dmfcc Slope 9
spectral dmfcc Slope 10	spectral dmfcc Slope 11
spectral dmfcc Slope 12	spectral dmfcc Slope 13

Continued on next page

Table A.1 – *Continued from previous page*

spectral dmfcc PeriodFreq 1	spectral dmfcc PeriodFreq 2
spectral dmfcc PeriodFreq 3	spectral dmfcc PeriodFreq 4
spectral dmfcc PeriodFreq 5	spectral dmfcc PeriodFreq 6
spectral dmfcc PeriodFreq 7	spectral dmfcc PeriodFreq 8
spectral dmfcc PeriodFreq 9	spectral dmfcc PeriodFreq 10
spectral dmfcc PeriodFreq 11	spectral dmfcc PeriodFreq 12
spectral dmfcc PeriodFreq 13	spectral dmfcc PeriodAmp 1
spectral dmfcc PeriodAmp 2	spectral dmfcc PeriodAmp 3
spectral dmfcc PeriodAmp 4	spectral dmfcc PeriodAmp 5
spectral dmfcc PeriodAmp 6	spectral dmfcc PeriodAmp 7
spectral dmfcc PeriodAmp 8	spectral dmfcc PeriodAmp 9
spectral dmfcc PeriodAmp 10	spectral dmfcc PeriodAmp 11
spectral dmfcc PeriodAmp 12	spectral dmfcc PeriodAmp 13
spectral dmfcc PeriodEntropy 1	spectral dmfcc PeriodEntropy 2
spectral dmfcc PeriodEntropy 3	spectral dmfcc PeriodEntropy 4
spectral dmfcc PeriodEntropy 5	spectral dmfcc PeriodEntropy 6
spectral dmfcc PeriodEntropy 7	spectral dmfcc PeriodEntropy 8
spectral dmfcc PeriodEntropy 9	spectral dmfcc PeriodEntropy 10
spectral dmfcc PeriodEntropy 11	spectral dmfcc PeriodEntropy 12
spectral dmfcc PeriodEntropy 13	spectral ddmfcc Mean 1
spectral ddmfcc Mean 2	spectral ddmfcc Mean 3
spectral ddmfcc Mean 4	spectral ddmfcc Mean 5
spectral ddmfcc Mean 6	spectral ddmfcc Mean 7
spectral ddmfcc Mean 8	spectral ddmfcc Mean 9
spectral ddmfcc Mean 10	spectral ddmfcc Mean 11
spectral ddmfcc Mean 12	spectral ddmfcc Mean 13
spectral ddmfcc Std 1	spectral ddmfcc Std 2
spectral ddmfcc Std 3	spectral ddmfcc Std 4
spectral ddmfcc Std 5	spectral ddmfcc Std 6
spectral ddmfcc Std 7	spectral ddmfcc Std 8
spectral ddmfcc Std 9	spectral ddmfcc Std 10
spectral ddmfcc Std 11	spectral ddmfcc Std 12
spectral ddmfcc Std 13	spectral ddmfcc Slope 1
spectral ddmfcc Slope 2	spectral ddmfcc Slope 3
spectral ddmfcc Slope 4	spectral ddmfcc Slope 5
spectral ddmfcc Slope 6	spectral ddmfcc Slope 7
spectral ddmfcc Slope 8	spectral ddmfcc Slope 9
spectral ddmfcc Slope 10	spectral ddmfcc Slope 11
spectral ddmfcc Slope 12	spectral ddmfcc Slope 13
spectral ddmfcc PeriodFreq 1	spectral ddmfcc PeriodFreq 2
spectral ddmfcc PeriodFreq 3	spectral ddmfcc PeriodFreq 4
spectral ddmfcc PeriodFreq 5	spectral ddmfcc PeriodFreq 6

Continued on next page

Table A.1 – *Continued from previous page*

spectral ddmfcc PeriodFreq 7	spectral ddmfcc PeriodFreq 8
spectral ddmfcc PeriodFreq 9	spectral ddmfcc PeriodFreq 10
spectral ddmfcc PeriodFreq 11	spectral ddmfcc PeriodFreq 12
spectral ddmfcc PeriodFreq 13	spectral ddmfcc PeriodAmp 1
spectral ddmfcc PeriodAmp 2	spectral ddmfcc PeriodAmp 3
spectral ddmfcc PeriodAmp 4	spectral ddmfcc PeriodAmp 5
spectral ddmfcc PeriodAmp 6	spectral ddmfcc PeriodAmp 7
spectral ddmfcc PeriodAmp 8	spectral ddmfcc PeriodAmp 9
spectral ddmfcc PeriodAmp 10	spectral ddmfcc PeriodAmp 11
spectral ddmfcc PeriodAmp 12	spectral ddmfcc PeriodAmp 13
spectral ddmfcc PeriodEntropy 1	spectral ddmfcc PeriodEntropy 2
spectral ddmfcc PeriodEntropy 3	spectral ddmfcc PeriodEntropy 4
spectral ddmfcc PeriodEntropy 5	spectral ddmfcc PeriodEntropy 6
spectral ddmfcc PeriodEntropy 7	spectral ddmfcc PeriodEntropy 8
spectral ddmfcc PeriodEntropy 9	spectral ddmfcc PeriodEntropy 10
spectral ddmfcc PeriodEntropy 11	spectral ddmfcc PeriodEntropy 12
spectral ddmfcc PeriodEntropy 13	timbre zerocross Mean
timbre zerocross Std	timbre zerocross Slope
timbre zerocross PeriodFreq	timbre zerocross PeriodAmp
timbre zerocross PeriodEntropy	timbre lowenergy Mean
timbre spectralflux Mean	timbre spectralflux Std
timbre spectralflux Slope	timbre spectralflux PeriodFreq
timbre spectralflux PeriodAmp	timbre spectralflux PeriodEntropy
tonal chromagram peak PeakPosMean	tonal chromagram peak PeakPosStd
tonal chromagram peak PeakPosSlope	tonal chromagram peak PeakPosPeriodFreq
tonal chromagram peak PeakPosPeriodAmp	tonal chromagram peak PeakPosPeriodEntropy
tonal chromagram peak PeakMagMean	tonal chromagram peak PeakMagStd
tonal chromagram peak PeakMagSlope	tonal chromagram peak PeakMagPeriodFreq
tonal chromagram peak PeakMagPeriodAmp	tonal chromagram peak PeakMagPeriodEntropy
tonal chromagram centroid Mean	tonal chromagram centroid Std
tonal chromagram centroid Slope	tonal chromagram centroid PeriodFreq
tonal chromagram centroid PeriodAmp	tonal chromagram centroid PeriodEntropy
tonal keyclarity Mean	tonal keyclarity Std
tonal keyclarity Slope	tonal keyclarity PeriodFreq
tonal keyclarity PeriodAmp	tonal keyclarity PeriodEntropy
tonal mode Mean	tonal mode Std
tonal mode Slope	tonal mode PeriodFreq
tonal mode PeriodAmp	tonal mode PeriodEntropy
tonal hcdf Mean	tonal hcdf Std
tonal hcdf Slope	tonal hcdf PeriodFreq
tonal hcdf PeriodAmp	tonal hcdf PeriodEntropy

Appendix B

Learning curves of AV prediction models

B.1 Learning curves

Figure B.1 shows the learning curves of all nine models for each train/test set (MSE2700, MediaEval1744, Saari596 and Eerola360).

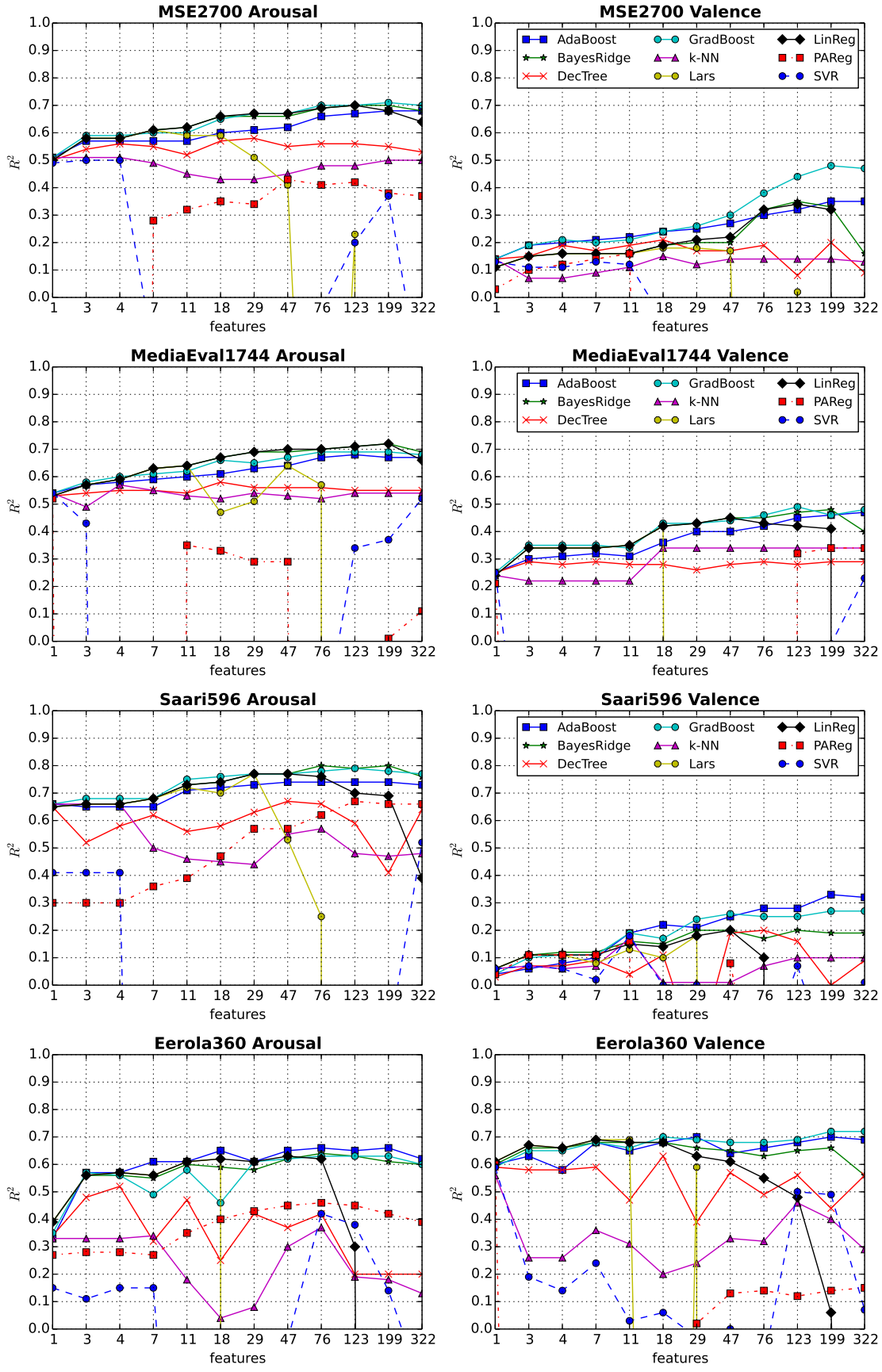


Figure B.1 AV learning curves for nine models on four datasets.

Appendix C

Results of tag weight prediction models

C.1 R^2 results per model

The best robust five models per tag are presented in figure C.1. These models are selected by picking the model that has the best worst R^2 score on both datasets (Saari596 and Eerola110).

For the results with the AV distance trained models, it shows that all 'tension' models preform almost equally. For the other tags and the tag set, differences between models are larger. For example SVR is largely outperformed by k-NN and GradBoost for respectively 'happy' and 'angry'. However, the results don't point into the direction of a irrefutable best model. In general GradBoost or k-NN performs best.

For the audio feature predictors it is more difficult to compare the different models since the R^2 results are low for most models. The robust result for the GradBoost 'angry' model stands out with $R^2 = .48$, which is high compared to the results of the other tags. Also standing out are the results for the tag set, since all 'robust' models score $R^2 < .0$. This suggests that there is no relation between various rated tags together and audio features.

What can be conclude from these results is that while individually trained models could enhance prediction scores for tags, choosing the right predictor set for a tag is more important.

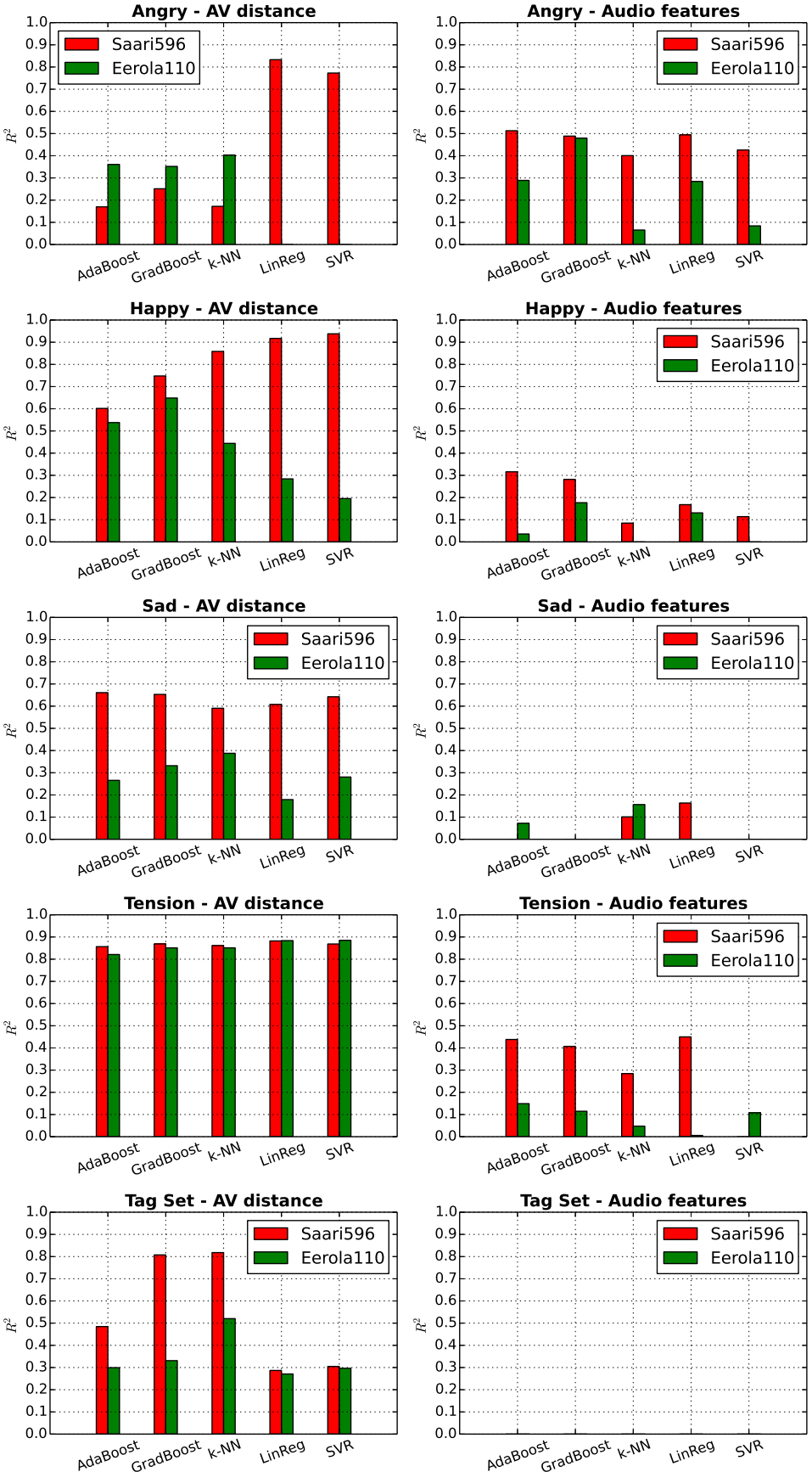


Figure C.1 Best robust models for the Saari596 and Eerola110 test set on four individual tags and a set of tags.

C.2 Learning curves

The learning curves of all models are presented in figure C.2. The results on the AV distance models are far more uniform compared to the results on the audio feature models on the Saari596 dataset, which was also the training dataset. Results on the Eerola110 dataset are in most cases more scattered suggesting overfitting of some models, e.g. SVR in 'happy - AV distance'.

The results on Eerola110 for 'angry - AV distance' and 'tension - AV distance' stand out, the first because the results for all models drop below .0 while for tension the R^2 scores almost perfectly follow those of the Saari596 dataset.

Another notable result is the sudden increase of k-NN and GradBoost in 'tag set - AV distance' after three predictors. However only for k-NN the results on Eerola110 also increase suggesting overfitting of the GradBoost model.

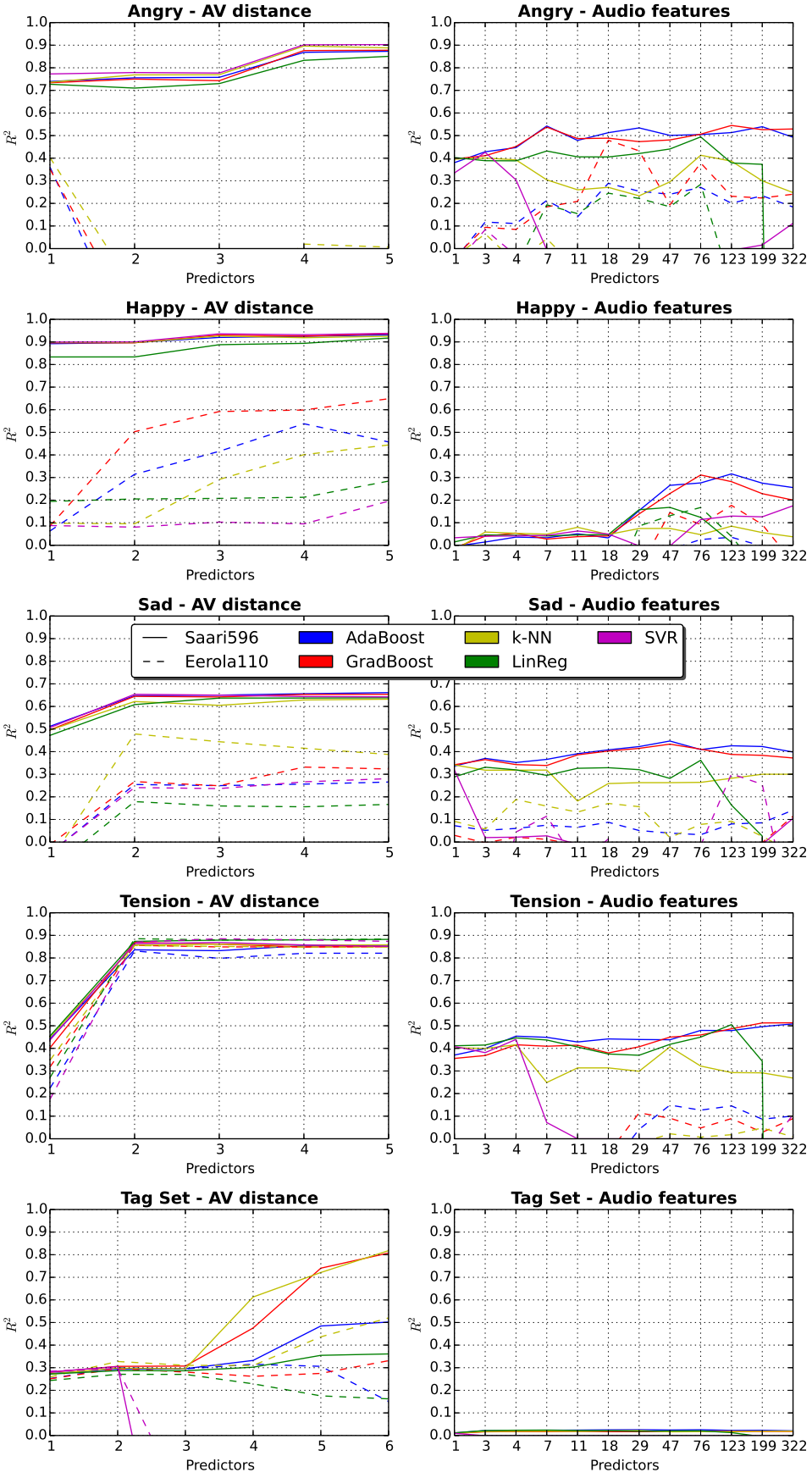


Figure C.2 Learning curves for the tag weight prediction models.

Appendix D

Datasets and results of the experiment

Table D.1 Dataset used in the questionnaire. The clip id's refer to the clips in the Saari596 dataset.

Clip ID	Tags				
3	fun	dreamy	sexy	black	cold
4	hypnotic	melancholy	calm	depressive	atmospheric
13	dread	desperate	epic	shiver	strong
14	power	moody	relaxing	sad	intense
40	romantic	relaxing	power	nostalgia	sweet
43	melancholy	shiver	emotional	sentimental	sad
44	romantic	cold	soothing	hope	relaxing
52	calm	fun	sad	wonder	bittersweet
53	epic	sweet	moody	sensual	lyrical
62	sentimental	calm	melancholy	romantic	atmospheric
66	epic	hope	intense	naive	sad
77	dreamy	relaxing	power	hope	emotional
79	relaxing	sad	interesting	epic	depressive
85	nostalgia	strong	lazy	dark	quiet
90	sentimental	sad	angry	emotional	tired
96	power	shock	relaxing	nostalgia	happy
101	sad	power	dark	depressive	calm
102	melancholy	power	gentle	sad	amiable
112	sweet	magical	content	melancholy	relaxing
116	melancholy	epic	sad	relaxing	emotional
127	pure	warm	dreamy	bittersweet	sad
136	melancholy	bittersweet	depressive	deep	sad
158	atmospheric	depressive	sexy	emotional	bittersweet

Continued on next page

Table D.1 – *Continued from previous page*

Clip ID	Tags				
182	elegant	dreamy	soothing	peaceful	romantic
189	atmospheric	relaxing	romantic	power	sexy
204	sentimental	sad	loneliness	cold	cynical
238	relaxing	hope	romantic	sad	gentle
241	sentimental	sad	calm	dreamy	romantic
243	energetic	optimistic	emotional	fun	hope
250	atmospheric	quiet	calm	dreamy	sweet
256	precious	calm	melancholy	happy	encouraging
278	emotional	romantic	sad	warm	pleasure
291	light	romantic	sweet	tender	passionate
301	dreamy	emotional	joy	hope	happy
344	weary	sad	melancholy	cold	depressive
377	dreamy	pleasure	relaxing	dark	sexy
381	pleasure	sentimental	sweet	relaxing	bittersweet
406	relaxing	melancholy	sleepy	dreamy	free
408	quiet	romantic	calm	sweet	dreamy
468	pure	positive	happy	enchanting	intimate
496	calm	melancholy	loneliness	sweet	dreamy
498	sad	epic	explosive	hope	relaxing
502	emotional	hope	hypnotic	peaceful	sad
504	surreal	difficult	fiery	euphoric	spiritual
526	fond	happy	fun	sexy	energetic
533	warm	lush	nostalgia	melancholy	power
553	hypnotic	relaxing	stylish	sensual	sweet
554	fun	detached	passionate	aggressive	rousing
576	lazy	bliss	quiet	happy	soothing

Table D.2 For each tag the mean AV ratings with standard deviations (std.) from the experiment and AV predictions are presented. The difference is the mean rating minus the predicted weight.

Clip Id	Tag	Mean rating and std.	Predicted weight	Difference
3	black	0.27 ± 0.35	0.18	0.09
3	cold	0.22 ± 0.26	0.21	0.01
3	dreamy	0.85 ± 0.19	0.21	0.64
3	fun	0.18 ± 0.22	0.46	-0.28
3	sexy	0.42 ± 0.34	0.55	-0.14
4	atmospheric	0.6 ± 0.28	0.59	0.01
4	calm	0.47 ± 0.3	0.17	0.3
4	depressive	0.52 ± 0.35	0.59	-0.07

Continued on next page

Table D.2 – *Continued from previous page*

Clip Id	Tag	Mean rating and std.	Predicted weight	Difference
4	hypnotic	0.7 ± 0.3	0.56	0.14
4	melancholy	0.69 ± 0.27	0.48	0.21
13	desperate	0.5 ± 0.32	0.68	−0.18
13	dread	0.46 ± 0.31	0.46	0.0
13	epic	0.28 ± 0.29	0.33	−0.05
13	shiver	0.38 ± 0.34	0.5	−0.12
13	strong	0.2 ± 0.2	0.19	0.01
14	intense	0.78 ± 0.26	0.51	0.28
14	moody	0.59 ± 0.37	0.23	0.37
14	power	0.49 ± 0.36	0.27	0.22
14	relaxing	0.64 ± 0.3	0.47	0.17
14	sad	0.66 ± 0.28	0.46	0.2
40	nostalgia	0.56 ± 0.25	0.54	0.02
40	power	0.68 ± 0.28	0.49	0.19
40	relaxing	0.49 ± 0.28	0.61	−0.12
40	romantic	0.45 ± 0.34	0.62	−0.17
40	sweet	0.43 ± 0.29	0.15	0.28
43	emotional	0.68 ± 0.24	0.18	0.5
43	melancholy	0.62 ± 0.32	0.29	0.33
43	sad	0.47 ± 0.24	0.41	0.06
43	sentimental	0.69 ± 0.29	0.48	0.21
43	shiver	0.34 ± 0.3	0.28	0.06
44	cold	0.17 ± 0.25	0.26	−0.09
44	hope	0.52 ± 0.23	0.18	0.34
44	relaxing	0.71 ± 0.21	0.73	−0.02
44	romantic	0.67 ± 0.26	0.43	0.24
44	soothing	0.68 ± 0.24	0.73	−0.05
52	bittersweet	0.47 ± 0.39	0.63	−0.16
52	calm	0.67 ± 0.24	0.45	0.22
52	fun	0.27 ± 0.32	0.4	−0.13
52	sad	0.59 ± 0.25	0.43	0.16
52	wonder	0.48 ± 0.35	0.31	0.17
53	epic	0.47 ± 0.32	0.26	0.21
53	lyrical	0.53 ± 0.29	0.35	0.17
53	moody	0.68 ± 0.2	0.61	0.07
53	sensual	0.39 ± 0.33	0.24	0.15
53	sweet	0.21 ± 0.21	0.2	0.0
62	atmospheric	0.55 ± 0.36	0.81	−0.25
62	calm	0.78 ± 0.27	0.6	0.18

Continued on next page

Table D.2 – *Continued from previous page*

Clip Id	Tag	Mean rating and std.	Predicted weight	Difference
62	melancholy	0.71 ± 0.34	0.24	0.46
62	romantic	0.66 ± 0.3	0.54	0.12
62	sentimental	0.82 ± 0.23	0.59	0.23
66	epic	0.57 ± 0.29	0.26	0.31
66	hope	0.35 ± 0.33	0.23	0.12
66	intense	0.67 ± 0.27	0.53	0.14
66	naive	0.21 ± 0.18	0.49	−0.29
66	sad	0.36 ± 0.25	0.47	−0.1
77	dreamy	0.5 ± 0.31	0.15	0.35
77	emotional	0.74 ± 0.22	0.16	0.58
77	hope	0.68 ± 0.29	0.16	0.52
77	power	0.54 ± 0.38	0.48	0.06
77	relaxing	0.51 ± 0.26	0.53	−0.02
79	depressive	0.23 ± 0.23	0.57	−0.34
79	epic	0.32 ± 0.26	0.44	−0.12
79	interesting	0.65 ± 0.27	0.41	0.24
79	relaxing	0.71 ± 0.31	0.51	0.2
79	sad	0.23 ± 0.24	0.51	−0.28
85	dark	0.2 ± 0.28	0.19	0.01
85	lazy	0.84 ± 0.25	0.29	0.56
85	nostalgia	0.7 ± 0.27	0.23	0.47
85	quiet	0.53 ± 0.36	0.68	−0.14
85	strong	0.48 ± 0.31	0.2	0.28
90	angry	0.24 ± 0.29	0.08	0.16
90	emotional	0.76 ± 0.13	0.2	0.56
90	sad	0.67 ± 0.25	0.44	0.22
90	sentimental	0.73 ± 0.23	0.51	0.22
90	tired	0.7 ± 0.25	0.61	0.09
96	happy	0.78 ± 0.24	0.69	0.09
96	nostalgia	0.52 ± 0.36	0.2	0.32
96	power	0.78 ± 0.3	0.14	0.65
96	relaxing	0.18 ± 0.21	0.33	−0.15
96	shock	0.32 ± 0.36	0.47	−0.15
101	calm	0.55 ± 0.34	0.59	−0.04
101	dark	0.61 ± 0.37	0.33	0.28
101	depressive	0.67 ± 0.28	0.59	0.08
101	power	0.25 ± 0.23	0.39	−0.14
101	sad	0.7 ± 0.31	0.54	0.16
102	amiable	0.51 ± 0.21	0.65	−0.14

Continued on next page

Table D.2 – *Continued from previous page*

Clip Id	Tag	Mean rating and std.	Predicted weight	Difference
102	gentle	0.67 ± 0.21	0.46	0.21
102	melancholy	0.6 ± 0.29	0.47	0.13
102	power	0.23 ± 0.23	0.48	−0.25
102	sad	0.62 ± 0.22	0.44	0.18
112	content	0.45 ± 0.29	0.47	−0.03
112	magical	0.39 ± 0.33	0.5	−0.1
112	melancholy	0.59 ± 0.28	0.31	0.28
112	relaxing	0.59 ± 0.23	0.46	0.13
112	sweet	0.59 ± 0.3	0.14	0.46
116	emotional	0.71 ± 0.26	0.34	0.37
116	epic	0.22 ± 0.29	0.27	−0.05
116	melancholy	0.76 ± 0.22	0.29	0.47
116	relaxing	0.58 ± 0.29	0.49	0.09
116	sad	0.63 ± 0.33	0.55	0.08
127	bittersweet	0.68 ± 0.22	0.73	−0.06
127	dreamy	0.73 ± 0.28	0.3	0.43
127	pure	0.6 ± 0.28	0.73	−0.12
127	sad	0.65 ± 0.26	0.52	0.12
127	warm	0.59 ± 0.28	0.33	0.26
136	bittersweet	0.66 ± 0.33	0.77	−0.11
136	deep	0.66 ± 0.3	0.32	0.33
136	depressive	0.55 ± 0.27	0.56	−0.01
136	melancholy	0.73 ± 0.32	0.26	0.47
136	sad	0.7 ± 0.32	0.55	0.15
158	atmospheric	0.63 ± 0.31	0.78	−0.15
158	bittersweet	0.68 ± 0.29	0.79	−0.12
158	depressive	0.34 ± 0.25	0.53	−0.19
158	emotional	0.67 ± 0.21	0.31	0.36
158	sexy	0.44 ± 0.32	0.51	−0.07
182	dreamy	0.88 ± 0.17	0.3	0.58
182	elegant	0.69 ± 0.32	0.77	−0.08
182	peaceful	0.76 ± 0.23	0.27	0.49
182	romantic	0.71 ± 0.27	0.45	0.26
182	soothing	0.66 ± 0.28	0.67	−0.01
189	atmospheric	0.79 ± 0.21	0.6	0.19
189	power	0.09 ± 0.16	0.5	−0.41
189	relaxing	0.75 ± 0.26	0.46	0.29
189	romantic	0.48 ± 0.36	0.3	0.18
189	sexy	0.43 ± 0.32	0.44	−0.01

Continued on next page

Table D.2 – *Continued from previous page*

Clip Id	Tag	Mean rating and std.	Predicted weight	Difference
204	cold	0.24 ± 0.25	0.36	-0.12
204	cynical	0.26 ± 0.33	0.28	-0.02
204	loneliness	0.74 ± 0.26	0.29	0.45
204	sad	0.67 ± 0.22	0.52	0.15
204	sentimental	0.74 ± 0.25	0.59	0.15
238	gentle	0.6 ± 0.26	0.15	0.45
238	hope	0.65 ± 0.29	0.41	0.24
238	relaxing	0.48 ± 0.31	0.46	0.02
238	romantic	0.67 ± 0.25	0.25	0.42
238	sad	0.47 ± 0.3	0.38	0.09
241	calm	0.79 ± 0.17	0.2	0.59
241	dreamy	0.79 ± 0.2	0.27	0.52
241	romantic	0.65 ± 0.25	0.49	0.15
241	sad	0.6 ± 0.31	0.52	0.08
241	sentimental	0.79 ± 0.26	0.57	0.22
243	emotional	0.48 ± 0.36	0.17	0.31
243	energetic	0.77 ± 0.22	0.64	0.13
243	fun	0.48 ± 0.29	0.59	-0.11
243	hope	0.58 ± 0.21	0.18	0.4
243	optimistic	0.69 ± 0.23	0.15	0.54
250	atmospheric	0.78 ± 0.25	0.59	0.19
250	calm	0.16 ± 0.25	0.39	-0.24
250	dreamy	0.54 ± 0.33	0.44	0.1
250	quiet	0.2 ± 0.23	0.51	-0.31
250	sweet	0.25 ± 0.22	0.41	-0.16
256	calm	0.22 ± 0.25	0.35	-0.14
256	encouraging	0.65 ± 0.32	0.23	0.42
256	happy	0.64 ± 0.35	0.46	0.17
256	melancholy	0.45 ± 0.35	0.26	0.19
256	precious	0.3 ± 0.31	0.44	-0.14
278	emotional	0.81 ± 0.21	0.17	0.64
278	pleasure	0.23 ± 0.24	0.5	-0.27
278	romantic	0.88 ± 0.14	0.6	0.27
278	sad	0.46 ± 0.32	0.39	0.07
278	warm	0.73 ± 0.22	0.19	0.54
291	light	0.63 ± 0.33	0.52	0.11
291	passionate	0.67 ± 0.26	0.56	0.12
291	romantic	0.66 ± 0.26	0.23	0.43
291	sweet	0.73 ± 0.16	0.14	0.59

Continued on next page

Table D.2 – *Continued from previous page*

Clip Id	Tag	Mean rating and std.	Predicted weight	Difference
291	tender	0.71 ± 0.2	0.52	0.19
301	dreamy	0.43 ± 0.32	0.23	0.19
301	emotional	0.28 ± 0.25	0.1	0.18
301	happy	0.81 ± 0.25	0.67	0.14
301	hope	0.59 ± 0.32	0.27	0.32
301	joy	0.83 ± 0.24	0.68	0.16
344	cold	0.26 ± 0.33	0.38	−0.12
344	depressive	0.6 ± 0.33	0.69	−0.08
344	melancholy	0.73 ± 0.24	0.41	0.32
344	sad	0.69 ± 0.27	0.63	0.06
344	weary	0.76 ± 0.25	0.67	0.09
377	dark	0.2 ± 0.27	0.18	0.02
377	dreamy	0.83 ± 0.22	0.21	0.62
377	pleasure	0.6 ± 0.3	0.48	0.12
377	relaxing	0.85 ± 0.16	0.48	0.37
377	sexy	0.64 ± 0.32	0.5	0.14
381	bittersweet	0.8 ± 0.24	0.74	0.06
381	pleasure	0.27 ± 0.2	0.47	−0.2
381	relaxing	0.6 ± 0.24	0.56	0.05
381	sentimental	0.89 ± 0.19	0.57	0.32
381	sweet	0.63 ± 0.27	0.31	0.31
406	dreamy	0.89 ± 0.23	0.13	0.75
406	free	0.56 ± 0.3	0.08	0.48
406	melancholy	0.58 ± 0.32	0.38	0.2
406	relaxing	0.76 ± 0.25	0.52	0.24
406	sleepy	0.76 ± 0.34	0.7	0.06
408	calm	0.79 ± 0.22	0.26	0.53
408	dreamy	0.77 ± 0.28	0.15	0.62
408	quiet	0.68 ± 0.24	0.71	−0.03
408	romantic	0.73 ± 0.29	0.4	0.33
408	sweet	0.82 ± 0.18	0.15	0.67
468	enchanting	0.65 ± 0.27	0.15	0.49
468	happy	0.25 ± 0.21	0.24	0.01
468	intimate	0.44 ± 0.34	0.13	0.31
468	positive	0.42 ± 0.29	0.2	0.22
468	pure	0.61 ± 0.22	0.72	−0.11
496	calm	0.81 ± 0.17	0.58	0.24
496	dreamy	0.7 ± 0.27	0.22	0.48
496	loneliness	0.75 ± 0.26	0.24	0.51

Continued on next page

Table D.2 – *Continued from previous page*

Clip Id	Tag	Mean rating and std.	Predicted weight	Difference
496	melancholy	0.71 ± 0.29	0.25	0.45
496	sweet	0.79 ± 0.19	0.19	0.61
498	epic	0.61 ± 0.25	0.25	0.37
498	explosive	0.81 ± 0.18	0.59	0.23
498	hope	0.22 ± 0.27	0.15	0.07
498	relaxing	0.04 ± 0.09	0.22	−0.18
498	sad	0.33 ± 0.29	0.46	−0.12
502	emotional	0.77 ± 0.23	0.24	0.53
502	hope	0.49 ± 0.28	0.31	0.18
502	hypnotic	0.57 ± 0.27	0.22	0.35
502	peaceful	0.71 ± 0.27	0.3	0.41
502	sad	0.68 ± 0.26	0.53	0.16
504	difficult	0.68 ± 0.25	0.54	0.14
504	euphoric	0.2 ± 0.27	0.18	0.02
504	fiery	0.47 ± 0.34	0.44	0.03
504	spiritual	0.42 ± 0.35	0.26	0.16
504	surreal	0.56 ± 0.37	0.41	0.15
526	energetic	0.96 ± 0.09	0.72	0.24
526	fond	0.54 ± 0.36	0.22	0.32
526	fun	0.89 ± 0.14	0.68	0.21
526	happy	0.92 ± 0.14	0.69	0.22
526	sexy	0.38 ± 0.31	0.21	0.16
533	lush	0.54 ± 0.25	0.22	0.32
533	melancholy	0.49 ± 0.35	0.24	0.25
533	nostalgia	0.4 ± 0.37	0.22	0.18
533	power	0.19 ± 0.26	0.44	−0.25
533	warm	0.5 ± 0.32	0.17	0.33
553	hypnotic	0.7 ± 0.27	0.19	0.51
553	relaxing	0.51 ± 0.28	0.51	0.0
553	sensual	0.49 ± 0.29	0.43	0.06
553	stylish	0.52 ± 0.34	0.5	0.02
553	sweet	0.26 ± 0.21	0.14	0.12
554	aggressive	0.34 ± 0.3	0.36	−0.01
554	detached	0.5 ± 0.35	0.27	0.23
554	fun	0.65 ± 0.32	0.46	0.19
554	passionate	0.23 ± 0.32	0.59	−0.36
554	rousing	0.61 ± 0.34	0.29	0.32
576	bliss	0.33 ± 0.32	0.74	−0.42
576	happy	0.27 ± 0.24	0.36	−0.09

Continued on next page

Table D.2 – *Continued from previous page*

Clip Id	**Tag**	**Mean rating and std.**	**Predicted weight**	**Difference**
576	lazy	0.7 ± 0.31	0.33	0.37
576	quiet	0.74 ± 0.25	0.67	0.07
576	soothing	0.64 ± 0.27	0.72	-0.08

Bibliography

Aljanaki, A., Wiering, F., and Veltkamp, R. C. (2014a). Collecting annotations for induced musical emotion via online game with a purpose Emotify. Technical Report April, Utrecht.

Aljanaki, A., Yang, Y.-H., and Soleymani, M. (2014b). Emotion in music task at MediaEval 2014. In *MediaEval*, pages 5–6.

Aristotle (1962). *Poetics*. Clarendon Press, Oxford.

Aucouturier, J.-J. and Bigand, E. (2012). Mel Cepstrum & Ann Ova: The difficult dialog between MIR and music cognition. *ISMIR*, (Ismir):397–402.

Barker, C., Pistrang, N., and Elliott, R. (2002). *Research methods in clinical psychology - an introduction for students*. John Wiley & Sons, Chichester, 2 edition.

Barthet, M., Fazekas, G., and Sandler, M. (2013). Music emotion recognition: From content- to context-based models. In Aramaki, M., Barthet, M., Kronland-Martinet, R., and Ystad, S., editors, *CMMR 2012 - From Sounds to Music and Emotions: Revised Selected Papers*, pages 228–252. Springer.

Beale, R. and Peter, C. (2008). The Role of Affect and Emotion in HCI. In Peter, C. and Beale, R., editors, *Affect and emotion in human-computer interaction: From theory to applications*, pages 1–11. Springer.

Bischoff, K., Firan, C. S., Paiu, R., Nejdl, W., Laurier, C., and Sordo, M. (2009). Music mood and theme classification - A hybrid approach. In *ISMIR*, pages 657–662.

Blagov, P. S. and Singer, J. A. (2004). Four dimensions of self-defining memories (specificity, meaning, content, and affect) and their relationships to self-restraint, distress, and repressive defensiveness. *Journal of Personality*, 72(3):481–512.

Böck, R. (2013). Multimodal automatic user disposition recognition in human-machine interaction. Technical report.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. *Technical Report C-1, The Center for Research in Psychophysiology*.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696.

Crane, T. (2003). *The mechanical mind*. Roudledge, New York, 2nd edition.

Darwin, C. (1872). *The expression of the emotions in man and animals*.

Davies, S. (2010). Emotions expressed and aroused by music: Philosophical perspectives. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of Music and Emotion*, chapter 1. Oxford University Press, New York.

Diener, E., Sandvik, E., Pavot, W., and Gallagher, D. (1991). Response artifacts in the mearsurement of subjective well-being. *Social Indicators Research*, 24:35–56.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.

Eerola, T., Lartillot, O., and Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. *ISMIR*.

Eerola, T. and Vuoskoski, J. K. (2010). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49.

Ekman, P. (1992). An argument for basic emotions. *Cognition and emotion*, 6(3):169–200.

Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, 43(5):349–358.

Futrelle, J. and Downie, J. S. (2003). Interdisciplinary research issues in music information retrieval. *Journal of New Music Research*, 32(2):121–131.

Gabrielsson, A. and Lindström, E. (2010). The role of structure in the musical expression of emotions. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of Music and Emotion*, chapter Ch. 14, page 367400. Oxford University Press, New York.

Gosselin, N., Peretz, I., Johnsen, E., and Adolphs, R. (2007). Amygdala damage impairs emotion recognition from music. *Neuropsychologia*, 45(2):236–44.

Grey, E. K. and Watson, D. (2001). Emotion, mood and temperament: similarities, differences, and a synthesis. In Payne, R. L. and Cooper, C. L., editors, *Emotions at work: theory, research and applications for management*, chapter 2, pages 21–43. John Wiley and Sons, Ltd.

Griffiths, P. E. (2008). Is emotion a natural kind? In Lycan, W. G. and Prinz, J. J., editors, *Mind and cognition: an anthology*, chapter 56, pages 850–862. Blackwell Publishing Ltd, 3th edition.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268.

Hu, X., Downie, J. S., and Ehmann, A. F. (2009). Lyric text mining in music mood classification. *ISMIR 2009*, pages 411–416.

Humphrey, E. J., Bello, J. P., and LeCun, Y. (2013). Feature learning and deep architectures: new directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481.

Hunter, P. G. and Schellenberg, G. E. (2010). Music and emotion. In Jones, M. R., Fay, R. R., and Popper, A. N., editors, *Music and Perception*, volume 36 of *Springer Handbook of Auditory Research*, chapter 5, pages 129–164. Springer New York, New York, NY.

Izard, C. E. (2007). Basic Emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, 2(3):260–280.

James, W. (1884). What is an emotion? *Mind*, 9:188–205.

Juslin, P. N. and Sloboda, J. (2010). *Handbook of music and emotion: Theory, research, applications*. Oxford University Press.

Juslin, P. N. and Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *The Behavioral and brain sciences*, 31(5):559–75; discussion 575–621.

Keltner, D. and James, J. G. (1999). Functional Accounts of Emotions. *Cognition and emotion*, 13(5):467–480.

Ketai, R. (1975). Affect, mood, emotion and feeling: Semantic considerations. *American Journal of Psychiatry*, 132(11):1215–1217.

Kim, Y. E., Schmidt, E., and Emelle, L. (2008). MoodSwings: A collaborative game for music mood label collection. In *ISMIR*, pages 231–236.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnbull, D. (2010). Music emotion recognition: A state of the art review. *ISMIR*, (Ismir):255–266.

Kim, Y. E., Williamson, D. S., and Pilli, S. (2006). Towards quantifying the "Album Effect " in artist identification. In *ISMIR 2006*.

Krippendorff, K. (2011). Computing Krippendorff ' s Alpha Reliability.

Lamere, P. (2008). Social tagging and music information retrieval. *Journal of New Music Research*, 37(2).

Lartillot, O. and Toiviainen, P. (2007). MIR in Matlab (II): a toolbox for musical reature extraction form audio. *Proceedings of 5th International Conference on Music Information Retrieval*.

Laurier, C., Sordo, M., Serrà, J., and Herrera, P. (2009). Music mood representations from social tags. In *ISMIR*, number Ismir, pages 381–386.

Levenson, R. W. (1994). Human emotion: a functional view. In Ekman, P. and Davidson, R. J., editors, *The nature of emotion: Fundamental questions*, pages 123–126. Oxford University Press, New York.

Levy, M. and Sandler, M. (2007). A semantic space for music derived from social tags. *Austrian Compuer Society*.

Levy, M. and Sandler, M. (2009). Music information retrieval using social tags and audio. In *IEEE Transactions on Multimedia*, pages 1–14.

Lichtenstein, A., Oehme, A., Kupschick, S., and Jürgensohn, T. (2008). Comparing Two Emotion Models for Deriving Affective States from Physiological Data. In Peter, C. and Beale, R., editors, *Affect and emotion in human-computer interaction: From theory to applications*, pages 35–50. Springer.

MacDorman, K. F., Ough, S., and Chin-Chang, H. (2007). Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, 36(4):281–299.

Magnini, B. and Cavagli, G. (2000). Integrating Subject Field Codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, 23(2):209–237.

McAlpin, C. (1925). Is music the language of emotion? *The Musical Quarterly*, 11(3):427–443.

Mckay, C. and Fujinaga, I. (2004). Automatic genre classification using large high-level musical feature sets. In *ISMIR 2004*.

Mion, L. and De Poli, G. (2008). Score-independent audio features for description of music expression. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):458–466.

Nanopoulos, A. and Karydis, I. (2011). Know thy neighbor: Combining audio features and social tags for effective music similarity. In *ICASSP 2011*, pages 165–168, Prague, Czech Republic.

Oatley, K. and Jenkins, J. M. (1996). *Understanding emotions*. Blackwell, Oxford, UK.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, 97(3):315–331.

Panda, R., Rocha, B., and Paiva, R. P. (2013). Dimensional music emotion recognition: Combining standard and melodic audio features. In *CMMR*, pages 1–11.

Plutchik, R. (2001). The nature of emotions. *American Scientist*, 89(4):344–350.

Pratt, C. C. (1952). Music as the language of emotion.

Robinson, M. D. and Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6):934–960.

Rubarth, S. (2005). Stoic philosophy of mind. *Internet Encyclopedia of Philosophy*.

Russell, J. A. (1980). A Circumplex model of Affect. *Journal of personality and social psychology*.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294.

Saari, P. and Eerola, T. (2013). Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2548–2560.

Saari, P., Eerola, T., Fazekas, G., Barthet, M., Lartillot, O., and Sandler, M. (2013a). The role of audio and tags in music mood prediction: a study using semantic layer projection. In *ISMIR*.

Saari, P., Eerola, T., Fazekas, G., and Sandler, M. (2013b). Using semantic layer projection for enhancing music mood prediction with audio features. In *SMC*, pages 722–728, Stockholm, Sweden.

Saari, P., Eerola, T., and Lartillot, O. (2011). Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1802–1812.

Scherer, K. R. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *Journal of New Music Research*, 33(3):239–251.

Schimmack, U. and Grob, A. (2000). Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14:325–345.

Schimmack, U. and Reisenzein, R. (2002). Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, 2(4):412–417.

Schmidt, E. M., Turnbull, D., and Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression categories and subject descriptors. *ISMIR*, pages 267–273.

Schubert, E. (2010). Continuous self-report methods. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of Music and Emotion*, chapter 9, pages 223–253. Oxford University Press, New York.

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061–86.

Shedler, J., Mayman, M., and Manis, M. (1993). The illusion of mental health. *American Psychologist*, 48(11):1117–1131.

Shiota, M. N. and Kalat, J. W. (2012). *Emotion*. Wadsworth Cengage Learning.

Sloboda, J. A. and Juslin, P. N. (2010). At the interface between the inner and outer world: Psychological perspectives. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of Music and Emotion*, chapter 4, pages 73–97. Oxford University Press, New York.

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., and Yang, Y.-H. (2013). 1000 Songs for emotional analysis of music. In *CrowdMM*, pages 1–6, Barcelona, Spain. ACM Press.

Solomon, R. C. (2008). The philosophy of emotions. In Lewis, M., Haviland-Jones, J. M., and Barrett, L. F., editors, *Handbook of emotions*, chapter 1. The Guilford Press, New York.

Song, Y., Dixon, S., and Pearce, M. (2012). Evaluation of musical features for emotion classification. *ISMIR*, (Ismir):523–528.

Song, Y., Dixon, S., Pearce, M., and Halpern, A. (2013). Do online social tags predict perceived or induced emotional responses to music? In *ISMIR*.

Speck, J. A., Schmidt, E. M., Morton, B. G., and Kim, Y. E. (2011). A comparative study of collaborative vs. traditional musical mood annotation. *ISMIR*, (Ismir):549–554.

Spindler, O. (2009). *Affective space interfaces*. PhD thesis, Technischen Universität Wien.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Sturm, B. L. (2012). An analysis of the GTZAN music genre dataset. *MIRUM 2012*, pages 7–12.

Sturm, B. L. (2013a). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, (11):371–406.

Sturm, B. L. (2013b). Evaluating music emotion recognition: Lessons from music genre recognition? In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference*.

Sullivan, G. M. and Feinn, R. (2012). Using Effect Size - or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(September):279–82.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford University Press, New York.

Thayer, R. E. (1996). *The origin of everyday moods: Managing energy, tension and stress*. Oxford University Press, New York.

Tsai, C.-f. (2012). Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*, 2012:19.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *Proceedings of the IEEE*, 10(5):293–302.

Vatolkin, I. (2012). Multi-objective evaluation of music classification. In Gaul, W. A., Geyer-Schulz, A., Schmidt-Thieme, L., and Kunze, J., editors, *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 401–410. Springer, Heidelberg.

Wang, D., Li, T., and Ogihara, M. (2010). Are tags better than audio features? The effect of joint use of tags and audio content features for artistic style clustering. *11th International Society on Music Information ...*, (Ismir):57–62.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, pages 1–35.

Watson, D. (2000). *Mood and temperament*. The Guilford Press, New York.

Wieczorkowska, A., Synak, P., and Ras, Z. W. (2006). Multi-label classification of emotions in music. *Advances in Soft Computing*, 35.

Yang, Y.-H. and Chen, H. H. (2011). *Music Emotion Recognition*. CRC Press, Inc., Boca Raton, USA.

Yang, Y.-H. and Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):1–30.

Yang, Y.-h., Lin, Y.-c., Su, Y.-f., and Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457.

Zentner, M. and Eerola, T. (2010). Self-report measures and models. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of Music and Emotion*, chapter 8, pages 187–222. Oxford University Press, New York.

Zentner, M., Grandjean, D., and Scherer, K. R. (2008). Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, 8(4):494–521.