

MUSIC EMOTION RECOGNITION: A MULTIMODAL
MACHINE LEARNING APPROACH



by

Cemre Gokalp

Submitted to the Graduate School of Management

in partial fulfillment of the requirements

for the degree of Master of Science

Sabancı University

July 2019

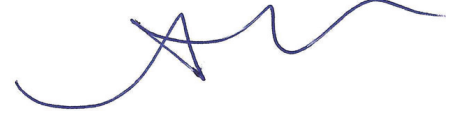
MUSIC EMOTION RECOGNITION: A MULTIMODAL MACHINE

LEARNING APPROACH

Approved by:

Assoc. Prof. Abdullah Daşqı

(Thesis Supervisor)

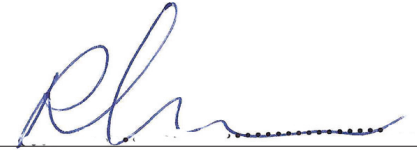


Assist. Prof. Ahmet Onur Durahim

(Thesis Co-Supervisor)



Assoc. Prof. Raha Akhavan-Tabatabaei



Assoc. Prof. Ayse Kocabiyikoglu



Assoc. Prof. Mumtaz Karatas



Date of approval:



© Cemre Gokalp, 2019

All Rights Reserved.

MUSIC EMOTION RECOGNITION: A MULTIMODAL MACHINE LEARNING APPROACH

Cemre Gökçalp

Business Analytics, Master's Thesis, 2019

Thesis Supervisor: Assoc. Prof. Abdullah Daşçı

Thesis Co-Supervisor: Assist. Prof. Ahmet Onur Durahim

Keywords: Music Emotion Recognition, Music Information Retrieval, Machine Learning, Feature Selection, Multi-Modal Analysis

ABSTRACT

Music emotion recognition (MER) is an emerging domain of the Music Information Retrieval (MIR) scientific community, and besides, music searches through emotions are one of the major selection preferred by web users.

As the world goes to digital, the musical contents in online databases, such as Last.fm have expanded exponentially, which require substantial manual efforts for managing them and also keeping them updated. Therefore, the demand for innovative and

adaptable search mechanisms, which can be personalized according to users' emotional state, has gained increasing consideration in recent years.

This thesis concentrates on addressing music emotion recognition problem by presenting several classification models, which were fed by textual features, as well as audio attributes extracted from the music. In this study, we build both supervised and semi-supervised classification designs under four research experiments, that addresses the emotional role of audio features, such as tempo, acousticness, and energy, and also the impact of textual features extracted by two different approaches, which are TF-IDF and Word2Vec. Furthermore, we proposed a multi-modal approach by using a combined feature-set consisting of the features from the audio content, as well as from context-aware data. For this purpose, we generated a ground truth dataset containing over 1500 labeled song lyrics and also unlabeled big data, which stands for more than 2.5 million Turkish documents, for achieving to generate an accurate automatic emotion classification system. The analytical models were conducted by adopting several algorithms on the cross-validated data by using Python. As a conclusion of the experiments, the best-attained performance was 44.2% when employing only audio features, whereas, with the usage of textual features, better performances were observed with 46.3% and 51.3% accuracy scores considering supervised and semi-supervised learning paradigms, respectively. As of last, even though we created a comprehensive feature set with the combination of audio and textual features, this approach did not display any significant improvement for classification performance.

MÜZİK DUYGUSU TANIMA: ÇOK-MODLU MAKİNE ÖĞRENMESİ YAKLAŞIMI

Cemre Gökalp

İş Analitiği, Yüksek Lisans Tezi, 2019

Tez Danışmanı: Doç. Dr. Abdullah Daşçı

Tez Eş-Danışmanı: Dr. Öğr. Üyesi Ahmet Onur Durahim

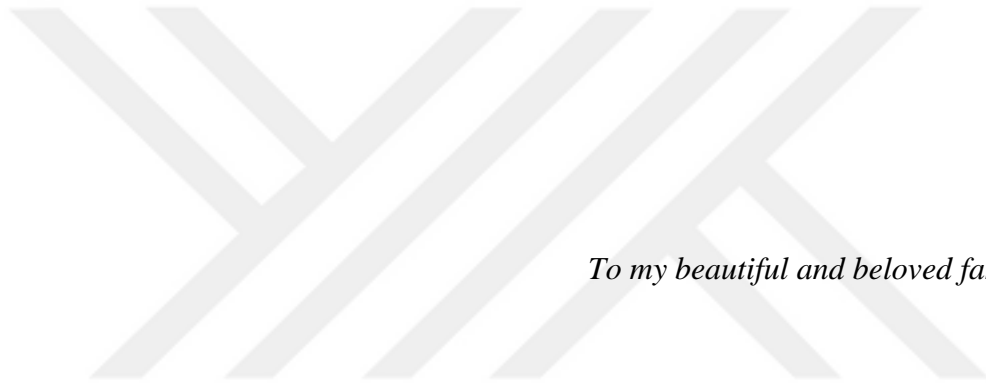
Anahtar Kelimeler: Müzik Duygusu Tanıma, Müzik Bilgisi Çıkarımı, Makine Öğrenmesi, Özellik Seçimi, Çok-Modlu Analiz

ÖZET

Müzik duygusu tanıma, müzik bilgisi çıkarım bilimsel topluluğunun yeni gelişmekte olan bir alanıdır ve aslında, duygular üzerinden yapılan müzik aramaları, web kullanıcıları tarafından kullanılan en önemli tercihlerden biridir.

Dünya dijitale giderken, Last.fm gibi çevrimiçi veritabanlarındaki müzik içerikleri katlanarak genişlemesi, içeriklerin yönetilmesi ve güncel tutulması için önemli bir manuel çaba gerektiriyor. Bu nedenle, kullanıcıların duygusal durumuna göre kişiselleştirilebilecek ileri ve esnek arama mekanizmalarına olan talep son yıllarda artan ilgi görmektedir.

Bu tezde, metinsel bazlı özelliklerin yanısıra müzikten türetilen sessel niteliklerle beslenen çeşitli sınıflandırılma modelleri sunarak, müzik duygu tanıma problemini ele almaya odaklanan bir çerçeve tasarlamıştır. Bu çalışmada, tempo, akustiklik ve enerji gibi ses özelliklerinin duygusal rolünü ve, iki farklı yaklaşımla, TF-IDF ve Word2Vec, elde edilen metinsel özelliklerin etkisini, hem denetimli hem de yarı denetimli tasarımlarla, dört araştırma deneyi altında ele aldık. Ayrıca, müzikten türetilen sessel özellikleri, içeriğe duyarlı verilerden gelen özelliklerle birleştirerek, çok modlu bir yaklaşım önerdik. Yüksek performanslı, otomatik bir duygu sınıflandırma sistemi oluşturmayı başarmak adına, 1500'den fazla etiketli şarkı sözü ve 2.5 milyondan fazla Türkçe belgenin bulunduğu etiketlenmemiş büyük veriyi içeren temel bir gerçek veri seti oluşturduk. Analitik modeller Python kullanılarak çapraz doğrulanmış veriler üzerinde birkaç farklı algoritma benimseyerek gerçekleştirildi. Deneylerin bir sonucu olarak, sadece ses özellikleri kullanılırken elde edilen en iyi performans %44,2 iken, metinsel özelliklerin kullanılmasıyla, sırasıyla denetimli ve yarı denetimli öğrenme paradigmaları dikkate alındığında, % 46,3 ve % 51,3 doğruluk puanları ile gelişmiş bir performans gözlenmiştir. Son olarak, sessel ve metinsel özelliklerin birleşimiyle oluşturulan bütünsel bir özellik seti yaratmış olsak da, bu yaklaşımın sınıflandırma performansı için önemli bir gelişme göstermediği gözlemlendi.



To my beautiful and beloved family...

ACKNOWLEDGEMENTS

I would like to thank Assoc. Prof. Abdullah Daşcı for his valuable support and mentoring in my thesis process. I consider myself a fortunate student who worked under the supervision of Assist. Prof. Ahmet Onur Durahim and want to deeply thank him for his precious guidance. Also, I must express my gratitude to Barış Çimen, his continuous support and academic wisdom assisted me during the course of this research.

I am thankful to my family, Şafak Bayındır, Mete Gökcalp and Mehmet Gökcalp for their endless support, patience, and guidance throughout my all steps. Also, I would like to thank Ateş Bey, he is always there for me. They always believe in me and encourage me all the time; I am lucky and happy to have them.

Besides, I would like to thank Ekin Akarçay, Sefa Özpınar, Ahmet Yakun, and Said Yılmaz for their contribution, support, and friendship. I thank all Business Analytics students for their kindly helps. And also, thank Osman Öncü for his agile support.

Finally, I give my deep thanks to Oğuzhan Sütçınar, completing this research would have been more painful without his support.

TABLE OF CONTENTS

ABSTRACT.....	i
ÖZET	iii
CHAPTER 1 - INTRODUCTION.....	1
1.1 Motivation, Contributions & Approach.....	4
1.1.1 Emotion Recognition	4
1.1.2 Feature Selection and Extraction	4
1.1.3 Creation of the Ground-truth Data and Emotion Annotation	5
1.1.4 Predictive Model Building using Machine Learning.....	5
1.2 Thesis Structure	5
CHAPTER 2 - LITERATURE REVIEW.....	7
Part-I: Psychology of Music: A Triangle encompassing Music, Emotion, and Human.....	7
2.1 Music and Emotion: Context & Overview	9
2.1.1 Definition of Emotion	9
2.1.2 Different Types of Emotion: Source of Emotion across the literature	10
2.1.3 Which Emotion Does Music Typically Evoke?.....	11
2.1.4 Subjectivity of Emotions.....	12
2.1.5 Musical Emotion Representation	13
2.1.5.1 Categorical Models	13
2.1.5.2 Dimensional Models	15

Part-II: Predictive Modelling of Emotion in Music	18
2.2 Framework for Music Emotion Recognition	19
2.2.1 Human Annotation	20
2.2.2 Emotion Recognition from Music through Information Retrieval	22
2.2.2.1 Audio Information Retrieval: Content-Based Feature Extraction.....	22
2.2.2.2 Lyric Information Retrieval: Contextual Feature Extraction	25
2.2.3 Emotion Recognition Using Features from Multiple Source.....	31
2.3 Emotion based Analysis and Classification of Music	41
2.3.1 Model Building by using Audio Features	41
2.3.2 Model Building by using Textual Features.....	43
2.3.3 Semi-supervised Learning by using Word Embeddings.....	44
CHAPTER 3 - METHODOLOGY	49
3.1 Dataset Acquisition.....	52
3.2 Selection of Emotion Categories and Annotation Process	53
3.3 Feature Selection and Extraction	57
3.3.1 Audio Feature Selection.....	57
3.3.2 Lyric Feature Extraction	61
3.3.2.1 Preprocessing and Data Cleaning.....	63
3.3.2.2 Textual Feature Extraction Process.....	65
3.4 Model Building and Testing	66

3.5 Evaluation.....	78
CHAPTER 4 - DISCUSSION & CONCLUSION.....	80
4.1 Research Framework Overview & Managerial Implications	80
4.2 Limitations & Future Works.....	83
BIBLIOGRAPHY	85



LIST OF FIGURES

Figure 2.1: Hevner's model (Hevner, 1936).....	14
Figure 2. 2: MIREX - The five clusters and respective subcategories.....	15
Figure 2. 3: Illustration of Core Affect Space.....	16
Figure 2. 4: Russel’s Circumplex Model	17
Figure 2. 5: GEMS-9 Emotion Classification	18
Figure 2. 7 Word Representation in Vector Space.....	46
Figure 3. 1 Analysis Flow Diagram	51
Figure 3. 2 A partial example for the labeled songs.....	55
Figure 3. 3 A portion from the labeled song data- After normalization.....	56
Figure 3. 4 A song lyric example – original version.....	62
Figure 3. 5 The lyric example after preprocessing without stemmed	64
Figure 3. 6 The stemmed lyric example	64
Figure 3. 7 The song data-set part.....	69
Figure 3. 8: A song example: Audio features-emotion tag matching	70
Figure 3. 9: A song example from lyric-emotion matching.....	71
Figure 3. 10: A song example with emotional tag, lyrics, and audio feature space.....	75

LIST OF TABLES

Table 2. 1 Subsequent MER & MIR Research Examples from the Literature	36
Table 3. 1 Tags with Sub-categories	54
Table 3. 2 Summary of ground truth data collection	56
Table 3. 3 Spotify Audio Feature Set and Feature Explanations	59
Table 3. 4: Music Audio Feature Analysis Performance Results	72
Table 3. 5: Music Lyric Feature (TF-IDF) Analysis Performance Results	73
Table 3. 6: Performance Results for Semi-Supervised Analysis using Word2Vec features ..	76
Table 3. 7: Performance Results for Semi-Supervised Multi-Modal Analysis	77

LIST OF EQUATIONS

Equation 3. 1 Accuracy Score	67
Equation 3. 2 Precision Score	67
Equation 3. 3 Recall (Sensitivity)	68
Equation 3. 4 F1 Score	68

LIST OF ABBREVIATIONS

AMG – All Music Guide

API - Application Programming Interface

BOW – Bag of Words

CBOW – Continuous Bag of Words

CCA – Canonical Correlation Analysis

GEMS - Geneva Emotional Music Scale

GMMs - Gaussian Mixture Models

GSSL – Graph-based Semi-Supervisor

HMM – Hidden Markov Model

IR – Information Retrieval

k-NN – k-nearest neighbors

LDA - Latent Dirichlet Allocation

LSA - Latent Semantic Analysis

MER – Music Emotion Regression

MIDI – Musical Instrument Digital Interfece

MIR – Music Information Retrieval

MIREX - Music Information Research Evaluation eXchange

MNB – Multinomial Naïve Bayes

MSE – Mean Square Error

NER – Name Entity Recognition

NB - Naïve Bayes

NLP - Natural Language Processing

NN – Neural Network

SVC – Support Vector Classifier

SVM - Support Vector Machine

POS - Part of Speech

PLSA - Probabilistic Latent Semantic Analysis

PSA - partial syntactic analysis

RF – Random Forest

RMSE – Root Mean Square Error

TF-IDF - Term Frequency-Inverse Document Frequency

V-A – Valence-Arousal

CHAPTER 1

INTRODUCTION

While the world goes into digital, extensive music collections are being created and become easily accessible. Thereby, the time and activities connecting music have found much more place in human life, and even people have started to involve music in their daily routines, such as eating, driving, and exercising (Tekwani, 2017). Also, in society, the emotional tendency of listeners has been manipulated by music, and affective responses to music have been evidenced in everyday life, such as background music in advertisements, in transportations during travel, and in restaurants (Duggal et al., 2014). Briefly, music is everywhere.

In scientific respect, music was described as “*a universal, human, dynamic, multi-purpose sound signaling system*” by Dr. Williamson, who is psychology lecturer at Goldsmith's College, London Music has been evaluated as universal because traditionally, almost every culture has its folkloric music. Drums and flutes have been found as primary instruments dating back thousands of years. Moreover, music is multi-purpose so that it can be used for identifying something, or it can encourage a crowd for bringing them together, or it can be employed for emotional trigger (Temple, 2015). Besides, Artist Stephanie Przybylek, who is also a designer and educator defined music as a combination of coordinated sound or sounds employed to convey a range of emotions and experiences (Przybylek, 2016).

In previous researches with the conventional approach, musical information has been extracted or organized accordingly the reference information, which depending on metadata-

based knowledge such as the name of the composer and the title of the work. In the area of Music Information Retrieval¹ (MIR), a significant amount of research has been devoted to some standard search structures and retrieval categories, such as genre, title, or artist, which can be easily found common ground, and quantified to a correct answer.

Even though this primary information will remain crucial, information retrieval, which depends on these attributes, is not satisfactory. Also, since musical emotion identification is still at the beginning of its journey in information science, the user-centered classification, which is based on predicting the emotional effect of music, still has a potential to discover in order to reach agreed-upon answers.

On the other hand, the vast music collections have also emerged a significant challenge on searching, retrieving, and organizing musical content; yet, the computational understanding of emotion perceived through music has gained interests in order to deal with content-based requests, such as recommendation, recognition, and identification. Consequently, a considerable amount of studies regarding the emotional effects of music has been designed recently, and many of them have discovered that emotion is an essential determinant in music information organization and detection (Song et al., 2012; Li & Ogihara, 2004; Panda et al., 2013). For example, in one of the earliest research, Pratt (1952) has summarized music as the language of emotion defended that evaluated music according to its emotional impressions, is a natural categorization process for human beings. After that, the connection and relationship between music and emotion were synthesized by Juslin and Laukka (2004), who declare that emotions are one of the primary impulses for music listening behavior.

Unfortunately, music listeners still face many hindrances while searching proper music for a specific emotion, and the requirement of innovative and contemporary retrieval and classification tools for music is maturing more evident (Meyers, 2007). Therefore, music listeners demand new channels to access their music.

The work displayed here is a music emotion recognition approach that renders the opportunity for listening to particular music in desired emotion, and consequently, it allows

¹ <https://musicinformationretrieval.com/index.html>

generating playlists with context awareness and helps users to organize their music collections, which lead to experience music in an inspiring way.

How can accurate predictive models of emotions perceived in music be created is the main question that we attempt to investigate it. In this respect, this thesis focuses on the investigation of

- Recognizing and predicting emotional affect driven from songs with the help of the annotation process, which contributes to human-centric perception for having a precise understanding of how can emotions and music be interpreted in the human mind,
- Retrieving different information from music through using multiple inputs, such as audio and textual features, and exploring the relationship between emotions and musical attributes,
- Proposing automatic music emotion classification approaches by employing supervised and unsupervised machine learning techniques and considering the emotional responses of humans to music, namely music psychology,
- Generating well-performed supervised models by using different algorithms and utilizing the extracted and analyzed audio features, as well as the appropriate textual metadata separately and also within a multimodal approach,
- Creating well-performed semi-supervised models by utilizing both the lyrical data from the songs and the big Turkish data collected from diverse public sources, including Turkish Wikipedia².

² <https://tr.wikipedia.org/wiki/Anasayfa>

1.1 Motivation, Contributions & Approach

Even though many variances can be seen regarding the approaches in the literature, this research offers an understanding of emotions in music, and the principles relating to machine learning through gathering different domains like music psychology and computational science under the same roof.

1.1.1 Emotion Recognition

In order to classify music with respect to emotion, first of all, we tried to create a precise understanding of how emotions and music are depicted in the human mind by considering the relation of music and emotion in the previous studies from various domains, that have been performed throughout the past century.

There have been many different representations and interpretations of human emotion and its relation to music. In the literature, emotions derived from music have been examined mainly under two approaches, such as categorical and dimensional. After all considerations, we observed that the categorical approaches have been more commonly used for emotional modeling, and generated better results in musical applications.

Therefore, in this research, the categorical model of emotion was implemented with four primary emotion categories as happy, sad, angry, and relaxed. These categories were chosen since they are related to basic emotions, which have been described in psychological theories, and also they encompass all quadrants of the Valence-Arousal space, which has been designed for capturing the perceived emotions and is therefore suited for the task of emotion prediction in songs.

1.1.2 Feature Selection and Extraction

After the emotional model resolution, the next step was to ascertain how does this model relate to musical attributes. In this research, we utilized the state-of-the-art textual and audio traits extracted from the music. Furthermore, a combination of lyrical and musical features was used for assessing the consolidated impact of these two mutually complementary components of a song. We aimed to reach appropriate representations of the songs before addressing them to the classification tasks.

1.1.3 Creation of the Ground-truth Data and Emotion Annotation

First of all, a database consisting of over 1500 song tracks and lyrics was compiled. The lyric data was cleaned and organized before moving further to the feature extraction process by employing text-mining algorithms. To be able to map the extracted attributes of songs onto the relevant emotional space, the songs were labeled into four emotional categories by four human annotators from diverse backgrounds. Furthermore, we utilized a big dataset with over 2.5 million Turkish texts, which was collected through three web sources to be able to generate a semi-supervised approach for emotion prediction. As far as we observed, this amount of data has not been used any relevant researches in Turkish literature.

1.1.4 Predictive Model Building using Machine Learning

In consideration of automatic emotion recognition from music, various MIR and MER researches have been done. Several machine learning algorithms such as Gaussian mixture models (Lu et al., 2006), support vector machines (Hu et al., 2009; Bischoff et al., 2009), neural networks (Feng et al., 2003) have been performed by using music attributes and emotion labels as model inputs.

One of the motivations behind this study is being able to provide an understanding of the association between emotion and musical features from various domains with the help of several machine learning algorithms. In this research, six different machine learning algorithms, which are support vector machines (SVM) with linear kernel, called SVC method, Linear SVC method, Multinomial Naïve Bayes, Random Forest classifier, Decision Tree classifier, and also Logistic Regression method were employed on the cross-validating data throughout the different experiments.

1.2 Thesis Structure

The literature background of this thesis is granted in Chapter 2 under three sub-sections. In the first section, we explore music psychology concerning human perception and the relation between music and emotion. The concept of emotion is clarified by examining the contextual views on emotion. Besides, the reality of human subjectivity in the literature is issued. Additionally, we explain the representations of musical emotion, namely emotional models. In the second section, previous works regarding emotion recognition from music are searched by

considering both emotion labeling approaches and information retrieval methods. In the last section, model designing and building phases of previous relevant researches are examined to observe how can music be classified according to emotion. As well as single-source, multisource supervised, unsupervised, and semi-supervised approaches are observed.

In Chapter 3, the design and implementation of the emotion classification system are outlined under four sub-sections. Ground-truth data collection and organization processes are revealed in the first section. In the second session, we describe emotional labels and model selection process. Besides, the annotation process regarding human perception of musical emotion is pointed out. In the third section, we present feature selection and extraction methods by utilizing both audio and lyrical sources. Also, data cleaning and pre-process are employed before textual information retrieval and explained detailly. Finally, in the last section, the predictive model building processes, which consist of training and testing phases, are designed and demonstrated under four different research experiments. In Experiment-1 and Experiment-2, audio and textual features are individually used, respectively. In Experiment-3, a semi-supervised approach is followed by using a word embedding method. In Experiment-4, we design a multimodal approach by combining audio and the selected textual features. After presenting the models' performances under different metrics, the chapter is concluded by the assessment of the model performances and the evaluation of the outcomes.

Finally, in Chapter 4, the overall framework is discussed and summarized. Besides, the limitations we met during this thesis, and some research insights are provided.

While considering all structure, in this thesis, we aim to introduce a prediction framework for providing a more human-like and comprehensive prediction of emotion, that capture the emotions the same way we as humans do, through building several machine learning models under four diverse and competitive research environments.

CHAPTER 2

LITERATURE REVIEW

In this chapter, several conceptual frameworks and methods representing the background knowledge of previous research on music and emotion were introduced concerning their pertinence to this project.

Part-I: Psychology of Music: A Triangle encompassing Music, Emotion, and Human

According to a straightforward dictionary definition, music is described as instrumental or vocal sounds consolidated to present harmony, beauty, and expression of emotion. Besides, it is evaluated as a means of expression that humankind has evolved over the centuries to connect people by evoking a common feeling in them (Kim et al., 2010). As social and psychological aspects are the preminent functions of music, it cannot be evaluated independently of any affective interaction in human life.

In both academia and the industry, researchers and scientists from cross-disciplines have been studying what music can express and how the human mind perceives and interprets music in order to find a music model fed by different features and human cognition. Music information

retrieval (MIR) researchers and music psychologists have been investigating the emotional effects of music and associations between emotions and music since at least the 19th century (Gabrielsson & Lindström, 2001). However, a gap emerged among the music studies in the past because studies from different disciplines focused on diverse aspects of emotion in music; yet, the fundamental presence of music in people's emotional state has been confirmed by further studies on music mood (Capurso et al., 1952). Moreover, additional indications of the emotional influence of music on human behavior have been presented by research from various study areas such as music therapy and social-psychological investigations involving the effects of music on social behavior (Fried & Berkowitz, 1979), and consumer research (North & Hargreaves, 1997).

Despite the idea of music retrieval regarding emotion is an entirely new domain, the researchers of the musical expressivity survey have demonstrated that "emotions" are selected as the most frequent option with 100% rate followed by "psychological tension/relaxation" and "physical aspects" which have 89% and 88% rate respectively (Patrick et al., 2004). Besides, music information behavior researchers have distinguished emotion as an essential aspect adopted by people in music exploration and organization, and therefore, Music Emotion Recognition (MER) has received growing attention (Panda et al., 2013a).

According to the research on Last.fm³ which is one of the most prominent music websites, emotion labels bonded to music records by online users has come up as the third most preferred social tag after genre and locale (Lamere, 2008). Moreover, a recent neuroscience investigation has revealed the permanence of a natural connection between emotion and music by showing music influences brain structures, which are acknowledged to be crucially responsible for emotions (Koelsch, 2014).

Consequently, music identification, retrieval, and organization by emotion has gained increasing awareness over time (Juslin & Sloboda, 2010; Eerola & Vuoskoski, 2013), and the affective character of the music, often referred to as music emotion or mood, has been recently identified as an essential determinant and considered a reasonable way in accessing and organizing music information (Hu, 2010).

³<http://www.last.fm/>

In light of this information, it can be said that an accurate judgment of how music is experienced and how emotions are embodied in the human mind and also in computational systems is essential to be able to design analyses and classification practices.

2.1 Music and Emotion: Contextual Overview

In this part, the main contextual characters consisting of the emotion definition, types, and models are discussed. First of all, the definition of the term "emotion" is examined. Then, different types of emotions, such as expressed or perceived emotions as well as the sources of emotion, are presented. Besides, which emotion types can be induced or felt by music are addressed. Next, the subjectivity cognition in music is evaluated, especially regarding social or cultural issues in the previous backgrounds. Finally, we end up this section by presenting the different emotion representations in music research across literature, which has been mainly diverged on the categorical and the dimensional models.

2.1.1 Definition of Emotion

Describing the concept of emotion is not straightforward. Fehr and Russell explained the toughness as "*Everybody knows what an emotion is until you ask them a definition*" (Fehr & Russel, 1984). Although there are several ways to define emotions, it can be defined as a psychological and mental state of mind correlated with several thoughts, behaviors, and feelings (Martinazo, 2010) resulting in comparatively powerful and brief reactions to goal-relevant variations in the environment (Patrick et al., 2004).

Previous studies have used both of the terms emotion and mood to refer the affective perception (Eerola & Vuoskoski, 2013). According to Ekman (2003), the relation between emotions and moods is bidirectional since a mood can activate particular emotions; yet, highly dense emotional experience may lead to the emergence of a determined mood. Even though emotion and mood have been used interchangeably, there are main distinctions that should be clarified. As Meyer depicted in his study, which is one of the essential studies analyzing the meaning of emotion in music, emotion is temporary and short-lived, whereas mood is relatively stable and lasts longer (Meyer, 1956). This opinion was supported by the following studies for nearly half a century (Juslin & Sloboda, 2001). An emotion habitually arises from known causes, while a mood often arises from unknown reasons. For instance, listening to a particular

song leads to joy or anger that may come up after an unpleasant discussion, whereas people may feel depressed or wake up sad without having a specific described reason (Malherio, 2016).

Research on music information retrieval has not always laid out the distinction between these terms (Watson & Mandry, 2012), while psychologists have often emphasized the difference (Yang & Chen, 2012a). Although both mood and emotion have been used to imply to the affective nature of music, the mood is generally preferred in MIR research (Lu et al., 2006; Mandel et al., 2006; Hu & Downie, 2007), while emotion is more widespread in music psychology (Juslin & Sloboda, 2001; Meyer, 1956; Juslin et al., 2006), while

Nevertheless, in this study, “emotion” was employed instead of mood since human perceptions of music are appraised in limited time and under known conditions.

2.1.2 Different Types of Emotion: Source of Emotion across the literature

Even though all music may not convey a particular and robust emotion, as Juslin and Sloboda stated, “*Some emotional experience is probably the main reason behind most people’s engagement with music.*” (Juslin & Sloboda, 2001). There can be several ways where music may evoke emotions, and the sources of it have been a topic of discussion in the literature.

Since Meyer, there have been two divergent opinions for the music meaning, which are absolutist and referentialist views. The absolutist view defends the idea that “musical meaning lies exclusively within the context of the work itself,” whereas the referentialist claim “musical meanings refer to the extra-musical world of concepts, actions, emotional states, and character.” (Juslin & Sloboda, 2001). Afterward, Juslin and Sloboda used and developed Meyer’s statement by claiming that the existence of two contradictory emotion sources. While intrinsic emotion is fed by the structural character of the music, extrinsic emotion is triggered out of music (Meyer, 1956).

In another study, Russel investigated how listeners respond to music by dividing the emotional sources as emotion(s) induced and expressed by music (Russel, 1980). Likewise, Gabrielsson (2002) examined the source of emotion into three distinct categories, such as expressed, perceived, and induced (felt) emotions.

While the performer triggers expressed emotion through communication to the listeners (Gabrielsson & Juslin, 1996), both perceived and induced emotions are connected to the listeners' emotional responses, and both are dependent on social interaction among the personal, situational, and musical factors (Gabrielsson, 2002). Juslin and Luakka (2004) also analyzed the differentiation between inductions and perceptions of emotion and explained that perceived emotion is evaluated as the human perception through the expressed emotion in music, while induced emotion stands for the feelings in response to the music. Furthermore, in another comprehensive literature review, it has been shown that the perceived emotion is mostly preferred in MIR research since the situational factors of listening relatively less influence it (Yang & Chen, 2012a).

In consideration of the literature review, in this study, perceived emotion was selected as the focused source of emotion in music.

2.1.3 Which Emotion Does Music Typically Evoke?

Researchers carried out studies investigating whether all emotions perceived or expressed by music in the same way or is there a differentiation on emotion levels triggered by music.

In one of the earliest examinations, the basic emotions were found as better communicators than complex emotions since basic emotions have more distinctive and expressive characteristics (Juslin, 1997). In their research, Juslin and Sloboda (2001), claimed that basic emotional expressions could be related to the fundamental basis of life, such as loss (sadness), cooperation (happiness), and competition (anger), and thus, communicative aspects of the emotions could be better.

Scherer and Oshinsky (1977) researched universal recognition ability of basic emotions through facial expression and showed that each basic emotions might have also been connected with the vocal character. In another investigation, Hunter et al. (2010) claimed that people correlate sadness with a slow tempo and happiness with a fast tempo because of the human tendency that the emotion results from vocal expressions via acoustic signals like tempo.

Juslin and Lindström (2003) included complex emotions into various music pieces performed by nine professional musicians to examine the recognition level of complex

emotions. The result of the study showed the musicians could not communicate emotions to listeners as well as they did with basic emotions. Further studies also showed that perceived emotion from music could vary within basic emotions. Sadness and happiness can be conveyed well and recognized comfortably in music (Mohn et al., 2010), whereas anger and fear seem relatively harder to detect (Kallinen & Ravaja, 2006).

2.1.4 Subjectivity of Emotions

Regardless of the emotion types portrayed in the previous section, one of the main challenges in MER studies can be pointed out as the subjective and ambiguous construct of emotion (Yang & Chen, 2012).

Because emotion perception evoked by a song is inherently subjective and is influenced by many factors, people can perceive varied emotions when listening to even the same song (Panda et al., 2013b). Numerous constituents might impact how emotion is perceived or expressed, such as social and cultural background (Koska et al., 2013), personality (Vuoskoski & Eerola, 2011), age (Morrison et al., 2008), and musical expertise (Castro & Lima, 2014). Besides, the listener's musical preferences and familiarity with the music (Jargreaves & North, 1997) may make it hard to obtain consensus. Furthermore, different emotions can be perceived along with the same song (Malherio, 2016).

On the other hand, Sloboda and Juslin (2001) defended the existence of uniform effects of emotion amongst different people, and toward their research, they showed that not all emotion types have the same level of the agreement, yet listeners' judgments on the music's emotional expression are usually constant, i.e., uniform. In the same year, Becker claimed that emotional receptions to music are a universal phenomenon and supported the idea by indicating anthropological research. Furthermore, psychological studies demonstrated that emotional subjectivity is not enough biased to restrict constituting reliable classification models (Laurier & Herrera, 2009).

In 2015, Chen and colleagues (2015) investigated the effect of personality traits in music retrieval problem by building a similarity-based music search system in aspects of genre, acoustic, and emotion. They used Pearson's correlation test to examine the relationship between preferred music and personality traits. The result displayed that when it comes to song selection,

although people with different personalities do behave differently, there is no reliable correlation between personality traits and the preferred music aspects in similarity search.

Consequently, when considering the previous research, it can be said that the perceived emotion from music can vary from person to person; yet, music can express a particular emotion reliably when there is a certain level of agreement among listeners.

2.1.5 Musical Emotion Representation

Throughout the literature, studies on both Music Emotion Recognition (MER) and psychology have laid out various models providing insight into how emotions are represented and interpreted within the human mind. Although there still is no universally accepted emotion representation because of the subjective and ambiguous nature of emotion, two main approaches to emotional modeling, namely categorical and dimensional models, have dominated the field even today. Even though each model type helps to convey a unique aspect of human emotion, the main distinction between the two models is that categorical models embody perceived emotion as a set of discrete categories or several descriptors identified by adjectives (Feng et al., 2003), whereas dimensional models classify emotions along several axes, such as discrete adjectives or as continuous values (Russel, 1980).

2.1.5.1 Categorical Models

The categorical model, which consists of several distinct classes, produces a simple way to select and categorize emotion (Juslin & Laukka, 2004), and it has been mostly used for goal-oriented situations like the study of perceived emotion (Eerola & Vuoskoski, 2013). This model defends that people experience emotions as diverse and main categories (Yang & Chen, 2012a). The most known and foremost approach in this representation is Paul Ekman's basic emotion model encompassing the limited set of innate and universal basic emotions such as happiness, sadness, anger, fear, and disgust (Ekman, 1992).

One of the earliest, yet still the best-known model has been Hevner's adjective circle of eight designed as a grouped list of adjectives (emotions), instead of using single words (Hevner, 2003). Hevner's list is composed of 67 different adjectives, organized into 8 different groups in a circular way, that is shown in the following figure, Figure 2.1. The adjectives inside each cluster have a very close meaning, which is used to describe the same emotional state, and

meaning closeness between adjectives is more prominent than from adjectives from distant clusters (Malherio, 2016). This model has been adopted and redefined by further studies; for instance, Schubert (2003) created a similar circle with 46 words into nine main emotion clusters.

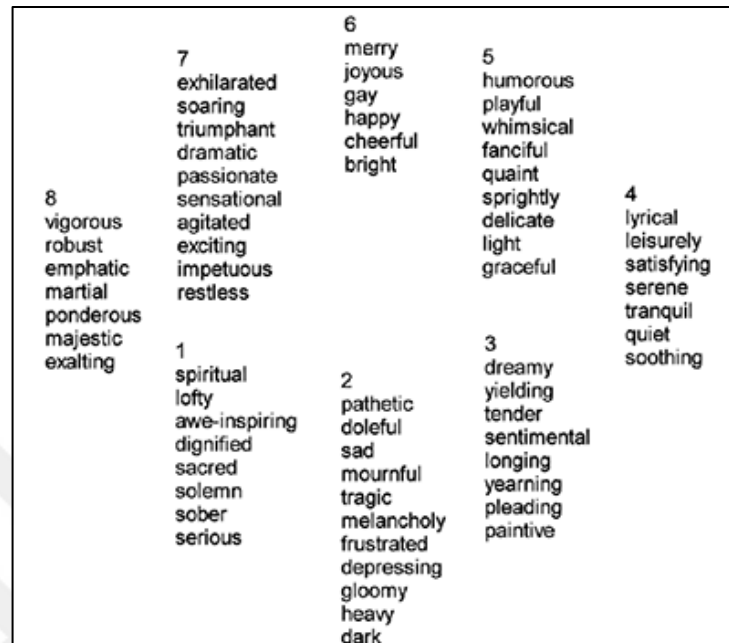


Figure 2.1: Hevner's model (Hevner, 1936)

During the studies, several emotion taxonomies have been emerged with various sets of emotions (Juslin & Sloboda, 2001; Hu & Lee, 2012; Yang et al., 2012). Besides, five clusters generated by Hu and Downie (2007) have gained prevalence in different domains of Music Information Retrieval (MIR) researches, such as music emotion recognition (MER), similarity, and music recommendation (Yang et al., 2012; Singhi & Brown, 2014). Furthermore, the five clusters and respective subcategories, depicted in Figure 2.2, were employed for audio mood classification in Music Information Retrieval Evaluation eXchange⁴ (MIREX), which is the framework employed by the MIR community for the formal evaluation of algorithms and systems (Downie, 2008).

⁴MIREX is a formal evaluation framework regulated and maintained by the International Music Information Retrieval Systems Evaluation Laboratory, IMIRSEL.

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

Figure 2. 2: MIREX - The five clusters and respective subcategories

Even though studies based on music and emotion have dominantly employed the categorical representations, some issues also exist since nonexistence of consensus on category numbers and subjective preference of humans for describing even the same emotion (Yang & Chen, 2012a; Yang & Chen, 2012b; Schuller et al., 2010)

2.1.5.2 Dimensional Models

A dimensional approach classifies emotions along several and independent axes in an affective space. In the literature, dimensional models showed differentiation mostly according to axes number as two or three, and also as being continuous or discrete (Mehrabian, 1996).

The typical dimensional model represents emotions within two main dimensions. Russell's valence-arousal model (1980) and Thayer's energy-stress model (1989), which represent emotions using a Cartesian space composed of the two emotional dimensions, are the most well-known models in this field.

In Russell's two-dimensional Valence-Arousal (V-A) space, which also known as the core affect space in psychology (Russell, 2003), valence stands for the polarity of emotion (negative and positive affective states, i.e., pleasantness), whereas arousal represents activation that is also known as energy or intensity (Russel, 1980). This fundamental model broadly used in several MER studies (Juslin & Sloboda, 2001; Laurier & Herrera, 2009), has shown that V-A Model provides a reliable way for people to measure emotion into two distinct dimensions (Yang & Chen, 2012b; Schuller et al., 2010; Schubert, 2014; Egermann et al., 2015).

Saari and Eerola (2014) have also suggested a third axis defining the potency or dominance of emotion to demonstrate the disparity among submissive and dominant emotions

(Mehrabian, 1996; Tellegen et al., 1999). Although the third dimension has been introduced as underlying elements of inclination in music (Bigand et al., 2005; Zentner et al., 2008), for the sake of integrity, this dimension was not generally employed in most of the MER investigations.

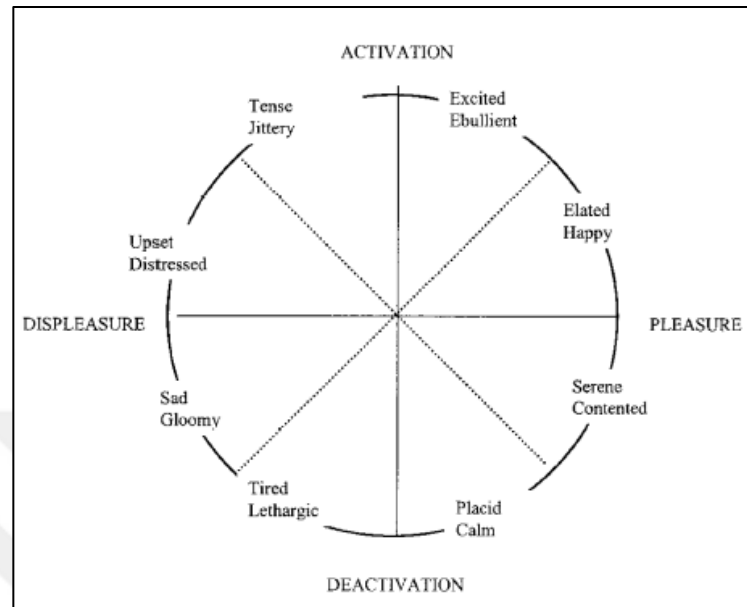


Figure 2. 3: Illustration of Core Affect Space

Moreover, dimensional models can be examined as being either discrete or continuous (Malherio, 2016). In discrete models, emotion tags have been used to depict different emotions in the distinct region of the emotional plane. The most famous examples for the discrete model are Russel's circumplex model, which is the two-dimensional model with four main emotional areas and 28 emotion-denoting adjectives (Russel, 1980), and also the adjective circle proposed by Kate Hevner, in which 67 tags are mapped to the respective quadrant (Hevner, 2003).

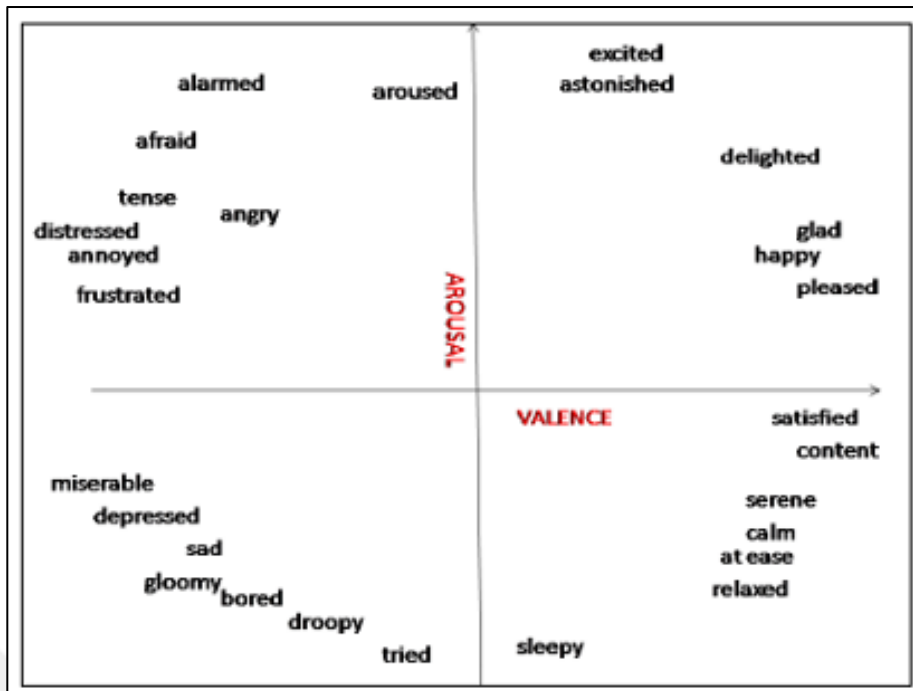


Figure 2. 4: Russel's Circumplex Model

Several researchers have utilized a subset of Russell's taxonomy in their studies. Hu et al. (2010) attested that Russell's space exhibits comparative similarities or distances within moods by distance. For occurrence, angry and calm as well as happy and sad are at opposite places, yet, for instance, happy and glad are close to each other (Hu & Downie, 2010a).

On the other hand, in continuous models, there are no specific emotional tags; instead, each point of the plane represents a different emotion (Yang et al., 2008a).

Even though the dimensional model has been widely used in literature, it has also been criticized for lack of clearness and differentiation among emotions having close neighbors. Also, some studies have shown that using the third dimension can increase ambiguity, yet some crucial aspects of emotion can be obscured in a two-dimensional representation. For example, fear and anger are resolutely located in the valence-arousal plane, but they have opposing supremacy (Yang et al., 2008b).

Apart from categorical and dimensional representation of emotion, the "Geneva Emotional Music Scale" (GEMS), which is a specially designed model to capture emotions induced by music, has been proposed (Zentner et al., 2008). In a later study, Rahul et al. (2014)

refined the GEMS model as (GEMS-9), which consists of nine primary emotions originating from 45 emotion labels. However, since GEMS only examine the emotion provoked by music and there exists no approved version in different languages, further investigation is necessary for the ever-increasing use of the model.

Emotional category	Explanation	Emotional category	Explanation
Calmness	Relaxation, serenity, meditativeness	Solemnity	Feeling of transcendence, inspiration, Thrills
Tenderness	Sensuality, affect, feeling of love	Power	Feeling strong, heroic, triumphant, energetic
Sadness	Depressed, sorrowful	Joyful activation	Feels like dancing, bouncy feeling, animated, amused
Tension	Nervous, impatient, irritated	Nostalgia	Dreamy, melancholic, sentimental feelings
Amazement	Feeling of wonder and happiness		

Figure 2. 5: GEMS-9 Emotion Classification

In this study, discrete dimensional representation of emotions with four emotional categories was employed because adopting from a mutually exclusive set of emotions has revealed an advantage for music emotion recognition through differentiating one emotion to another (Lu et al., 2010). Four primary emotions, such as happy, sad, calm, and relaxed, which have universal usage and cover all quadrants of the two-dimensional emotional model, were decided before starting the annotation process.

Part-II: Predictive Modelling of Emotion in Music

With the evolution of technology, the Internet has become a significant source of accessing information, which has resulted in an explosion of easily-accessible and vast digital music collections over the past decade (Song, 2016). Digitalization has also triggered the studies on MIR over automated systems regarding organizing and searching for music and related data (Kim et al., 2010). However, as the number of musical content proceeds to explode, the essence of musical experience has transformed at a primary level, and conventional ways of investigating and retrieving musical information on bibliographic knowledge, such as composer name, song title, and track play counts, have become no longer sufficient (Yang & Chen,

2012a). Thereby, music listeners and the researchers have started to seek for new and more innovative ways to access and organize music, and the efficiency necessity on music information retrieval and classification has become more and more prominent (Juslin & Sloboda, 2010).

Besides that, previous researches confirmed the fact that since music's preeminent functions are psychological and social, the most useful retrieval indexes should depend on four types of information, such as the genre, style, similarity, and emotion (Huron, 2000). Accordingly, a great deal of studies on music information behavior, which are not just from music psychology and cognition (as described in the above section), but also in machine learning, computer science, and signal processing, (Schubert, 2014), have identified emotions as an essential criterion for music retrieval and organization (Casey et al., 2008; Friberg, 2008). Likewise, a significant number of researches has been moved out on MER systems (Yang & Chen, 2012b).

So far, the cognitive aspects of music, as well as the emotional responses and representations, so-called music psychology, across the literature have been examined.

In the next section, we offer an examination of different MIR investigations in music theory, which contain the striking music features' extraction and the analysis of such features through the application of various machine learning techniques.

2.2 Framework for Music Emotion Recognition

Music theory is challenged to make observations and accordingly, acquainted judgments about the extraction of prominent music traits and the utilization of such traits.

Emotion identification can be inspected as a multilabel or multiclass classification, or as a regression enigma, in which each music composition is annotated with a collection of emotions (Kim et al., 2010), and a considerable number of researches with various experiments have been done on predictive emotional model creation (Yang & Chen, 2012a; Barthelet et al., 2012). Although the studies have diversified aspects changing according to the aim of the research, the accessible sources or emotional representations, the primary distinction among investigations have mainly been created through the feature selection and extraction processes by operating

various sources with or without human involvement and using different algorithms, methods, and techniques.

There have been numerous research strategies using the features from the singular source such as audio, lyrics, or crowdsourced tags. Furthermore, bimodal approaches like using both audio and lyrics, and also, multimodal approaches consolidating audio, lyrics, and tags have been applied in the previous researches.

Regardless of the employed taxonomy, collection of objective data, namely “ground-truth data” is generally the first and one of the most crucial steps for reaching necessary information to be able to apply analytics on (Malherio, 2016). In this respect, even though different approaches, such as data collection games and social-tags have been used (Kim et al., 2010), one of the most prevalent ways to generate a ground truth dataset is still manual labeling (Yang & Chen, 2012b; Schuller et al., 2010; Saari., 2015).

2.2.1 Human Annotation

The agile extension in compact digital devices and Internet technology have shaped music accessible practically everywhere, which has altered the cosmos of music experience and the ways of exploring and listening to music. Music discovery web services, such as AllMusic Guide (AMG)⁵, iTunes⁶, Last.FM, Pandora⁷, Spotify⁸, and YouTube⁹ have replaced traditional ways to access music (Casey et al., 2008). Although these platforms have extensive music catalogs and most of the musical content is effortlessly obtainable on the platforms, the lack of ground truth data set, and emotion labels have been retained as a particularly challenging problem for Music-IR systems mainly because of the copyright issues (Kim et al., 2010). Regardless of the employed MER taxonomy, since the collection and annotation of ground truth data is the foremost step for investigation of emotion in music, different approaches have been followed towards the retrieving information from these collections, as well as manage them in the field of MIR.

⁵ <http://www.allmusic.com/>

⁶ <https://www.apple.com/music/>

⁷ <http://www.pandora.com/>

⁸ <https://www.spotify.com/>

⁹ <https://www.youtube.com/>

Manual annotation is a commonly preferred way for creating a ground truth data set, which is generally applied by collecting emotional content information in music through a survey (Saari., 2015). Even though this is an expensive process in terms of human labor and financial cost, most researches have believed that this method enables better control regarding ambiguity (Yang et al., 2008b). For instance, Turnbull et al. (2008) collected the CAL500 data set of labeled music consisting of 500 songs, which was manually annotated into 18 emotional categories by a minimum of three non-expert inspectors. Similarly, in another MIR study, another publicly available dataset was also generated by three expert listeners through using six emotions (Trohidis et al., 2008).

A second approach considering the direct collection of human-annotated information (e.g., semantic tags) about music, involves social tagging. Music discovery and recommendation platforms, such as AllMusic and Last.FM have been utilized in some of the previous researches since they enabled to provide social tags through a text box in the interface of audio player (Levy & Sandler., 2009; Bischoff et al.,2009).

Panda et al. (2013) have suggested a methodology for the production of a multi-modal music emotion dataset by practicing the emotion labels in the MIREX mood classification task and utilizing the AllMusic database. Likewise, Song (2016) adopted social tags from Last.FM in order to create music emotion dataset with popular Western songs.

On the other hand, Duggal et al. (2014) created a website for labeling the songs into a maximum of 3 emotions. They generated an emotional profile for each song only if the song reaches a certain threshold level. Corresponding to manual annotation, using social tag can be interpreted a more comfortable and faster way to collect the ground truth data to create a useful resource for the Music-IR community. However, several problems defecting the reliability of the annotation quality also exist, such as data sparsity due to the cold-start problem, popularity bias, and malicious tagging (Lamere & Celma, 2007). In consequence, the discussion on the best way for reaching qualified emotion annotations considering a large number of songs, still exist.

Lastly, collaborative games on the web, so-called Games with a Purpose (GWAP) is another preferred method for the collection of music data and the ground truth labels. For instance, Kim et al. (2008) have presented MoodSwings, which is an online and collaborative

game for emotions annotation on songs. The game aims to record dynamic (per-second) mood ratings of multiple players within the two-dimensional Arousal-Valence space by using 30-second music clips. Yang and Chen (2012) have utilized another online multiplayer game called Listen Game, which was initially designed by Turnbull and his colleagues in 2008. In the game, players are asked to select both of the best and worst options, which describes the emotion of song by offering a list of semantically related words. Final scores of each player are decided by calculating the amount of agreement between the players' preferences and the decisions of all other players. Even though the method seems more practical for the annotation process, it was designed as suitable mostly for short-term, 30 seconds tracks, audio clips.

2.2.2 Emotion Recognition from Music through Information Retrieval

For effective music retrieval and music emotion recognition, musical feature selection for model inputs has been one of the crucial aspects of creating variations among previous research approaches. While some studies focused on solely one type of input extracted from music like audio or lyrical features, some of them exploited multimodal approaches embracing features from more than one structure such as a combination of audio and lyrics inputs, and also, annotators' tags as well for obtaining more accurate and reliable mood classifiers.

2.2.2.1 Audio Information Retrieval: Content-Based Feature Extraction

Since at least the 19th century, researchers have been studying to answer how does the human mind interpret and experience music (Gabrielsson & Lindström, 2001). The problem was more actively addressed in the 20th century through an investigation of the relationship between emotional judgments of listeners and particular musical parameters such as rhythm, mode, harmony, and tempo (Friberg, 2008). For instance, happy music has been commonly associated with a major mode, simple and consonant harmony, whereas sad music has been generally correlated with a minor mode, complex and dissonant harmonies (Panda et al., 2013a). On the other hand, some previous researches revealed that the same feature can reflect a similar manner for more than one emotional expression. For example, a fast tempo can reflect both happiness and anger (Juslin & Sloboda, 2001). However, there is a general assessment saying that emotional perception of music is derived mainly from the audio itself since the contextual information of music pieces may be inadequate or missing completely, such as for

newly composed music (Koelsch, 2014). Therefore, several researchers have also studied the hidden associations between musical characteristics and emotions over the years.

As far as the knowledge in the literature background, the first MER paper consisting of a method for sentiment analysis with audio features was published by Katayose and his colleagues in 1988. In this study, audio music principles such as harmony, rhythm, and melody, which were derived from the orchestral piano music records, were adopted to predict the emotion with heuristic customs (Katayose et al., 1988).

Even though Music-IR has been directed towards the enhanced usage of audio and acoustic features, and although some investigations have focused on revealing the most informative musical features for emotion recognition and classification, no single predominant feature has been generated in the literature. Sloboda and Juslin (2001) have proved the existence of some correlation between emotion and musical attributes, such as rhythm, pitch, tempo, mode, dynamics, and harmony. Friberg (2008) has prepared the following features as relevant for music and emotion, such as melody, harmony, timbre, pitch, timing, articulation, rhythm, and dynamics. However, some musical attributes ordinarily correlated with emotion was not reflected on that list such as mode, loudness (Katayose et al., 1988). Additionally, Eerola and his colleagues (2009) have revealed a particular subset of informative audio features for emotion recognition, which consists of a wide range of musical attributes, such as harmony, dynamics, timbre, and rhythm.

Despite the existence of various research, Lu and his colleagues (2006) proposed one of the first and most comprehensive studies by examining a categorical view of emotion. In this research, Thayer's model was used to represent emotions into four distinct quadrants, and three different musical features were extracted, which are intensity, timbre, and rhythm. Furthermore, several feature extraction toolboxes such as Marsyas¹⁰, Music Analysis, Retrieval, and Synthesis for Audio Signals, MIRtoolbox¹¹, and PsySound¹² have been developed for classification of musical signals through extracting audio features (Eerola et al., 2009). However, it is essential to note that audio features producing by these tools are not the same and show variation. For example, while the Marsyas tool extracts audio features such as melody

¹⁰ <http://marsyas.info/>

¹¹ <https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox>

¹² <http://psysound.org/>

spectrum (Beveridge et al., 2008; Tzanetakis & Cook, 2000), MIRtoolbox provides a set of features from the statistics of frame-level features.

The research has been done by Feng et al. (2003) can be given as one of the earliest MER studies utilized audio signals. In that study, only two musical parameters, which are tempo and articulation, were extracted as input features in order for classification of songs into four categorical emotion, that are happy, sad, anger, and fear. Although Feng achieved an average precision by 67%, only 23 pieces were used during the test phase. Because of the limited number of the test corpus as well as extracted features, unfortunately, the study cannot provide enough evidence of generality. Yang et al. (2008) proposed one of the first researches using a continuous model on emotion recognition through music signals. In this work, each music clip was matched with a point in Russell's valence-arousal (V-A) plane, and PsySound and Marsyas tools were utilized for audio information retrieval process to extract musical attributes, such as loudness, level, dissonance, pitch, and timbral features. Panda and Paiva (2011) also used the Yang's dataset, which consists of 194 excerpts from different genres and extracted audio features through using the Marsyas, PsySound, and MIR toolbox. As a result of this study, they achieved 35.6% and 63% valence and arousal prediction accuracy, respectively.

As audio decoding of musical features have been provided by some Web-services such as EchoNest¹³ and Spotify, the way of extracting audio information has also been evolved, and such web services have been used as a base for autodetection of emotion in music (Lehtiniemi & Ojala, 2013). Panda et al. (2013) proposed an approach by combining melodic and standard audio features in dimensional MER researches. In that study, EchoNest browser was used to extract 458 standard features and 98 melodic features out of 189 audio clips, and they showed that combining standard audio with melodic features improved performance results from 63.2% and 35.2% to 67.4 and 40.6% for arousal and valence prediction, respectively. In another study, Tekwani (2017) tried to find an answer for whether an audio content model can capture the particular attributes, which make a song sad or happy, in the same way as humans do, and for that purpose they utilized the Million Song Dataset¹⁴ (MSD) created by LabROSA at Columbia University in association with Echo Nest. 7396 songs, which were hand-labeled as happy and sad, and the musical audio attributes, such as Speechiness, Danceability, Energy, Acousticness,

¹³ <http://the.echonest.com/>

¹⁴ <http://millionsongdataset.com/>

and Instrumentalness were extracted through using the Spotify API¹⁵ for building a classification model. The research findings showed that danceability, energy, speechiness, and the number of beats are important features since they correlate the emotional perceptions of humans while interpreting music.

2.2.2.2 Lyric Information Retrieval: Contextual Feature Extraction

The annual Music Information Research Evaluation eXchange (MIREX) is a community-based framework evaluating Music-IR systems and algorithms for finding solutions to the audio music mood and genre classification since 2007 (Hu & Downie, 2007). Even though operating systems in this division have shown development over the years by using only acoustic features, utilizing solely audio features for emotion classification has reached a limit because of the undeniable presence of the semantic gap between the object feature level and the human cognitive level of emotion perception (Yang et al., 2008b). Indeed, several psychological studies have also confirmed that part of the semantic information of songs resides exclusively in the lyrics, and thus lyrics can provide a more precise and accurate expression of emotion (Logan et al., 2004). Namely, lyrics can contain and reveal proper emotional information that is not encapsulated in the audio (Besson et al., 2011). In the survey, which was prepared by Juslin and Laukka (2004) regarding everyday listening habits, lyrics have been chosen by 29% of the participants as the foundation of their judgments regarding their musical perception.

Lyric-based approaches have been found particularly tricky since feature extraction, and emotional labeling designs of lyrics are non-trivial, primarily when regarding the complexities associated with disambiguating affect from the text. Even though there was a paucity of researches, which utilize textual inputs for emotion detection, when compared to the other areas such as facial, speech, and audio emotion detection, emotion detection from text has gained increasing attention in recent years (Binali et al., 2010). Moreover, studies, which utilize lyrics by representing each word as a vector, and each text as a vector of features, have appeared (Song, 2016).

¹⁵ <https://developer.spotify.com/documentation/web-api/>

The most popular features extracted from the text can be classified into mainly three categories, such as content-based features with and without typical Natural Language Processing (NLP) transformations (e.g., stemming, Part-of-Speech Tags - POS tags, stopword elimination), text stylistic features based on the style of the written text, and linguistic features based on lexicons (Hu, 2010).

In MIR researches, the most preferred features in text analysis (and consequently, in lyric analysis) has been the content-based features, namely the bag-of-words, BOW, (Xia et al., 2008; Yang & Chen, 2012b; Lu et al., 2010). In this representation approach, texts, i.e., lyrics, are described as a set of words, namely bags, with various dimensions, such as unigrams, bigrams, and trigrams, which represents the counts of the word cloud. While the number of text features depicts the dimension of the text, the content of the text is determined according to the frequencies of the features within the text (Mulins, 2008). Even this approach can be employed directly, a set of transformation such as stemming and stopword removal have been generally applied to the subject after the tokenization of the original text to improve classification accuracy. While stemming transforms each word into their root, i.e., stemmed version, elimination of stopword, which also called function words, helps to remove non-discriminative words such as 'the' from the corpus (Malherio, 2016). In a study, Hu et al. (2010) used bag-of-words (BOW) features in various representations, such as unigram, bigram, trigram and they have indicated that higher-order BOW traits have captured more of the semantics through adopting combinations of unigram, bigram, and trigram tokens performed more reliable than single n-grams. In another research, the authors analyzed traditional bag-of-words features, and their combinations, as well as three feature representation models, which were absolute term frequency, Boolean, and TF-IDF weighting (Leman et al., 2005). Their outcomes confirmed that the combination of unigram, bigram, and trigram tokens with TF-IDF weighting provided the most dependable model performance, which indicates that higher-order BOW features can be more valuable for emotion categorization.

Even though BOW model has been one of the most widely used models in the literature, it requires a high dimensional space to represent the document and does not consider the semantic relationship between terms. Therefore, the order and relations between words are ignored, and unfortunately, it leads to relatively poor categorization accuracy (Menga et al., 2011). Favorably, there are other representations reflecting extensions of the BOW model, such

as methods focusing on phrases instead of single words, and others take advantage of the hierarchical nature of the text. Zaanen et al. (2010) presented a paper regarding the lingual parts of the music in an automatic mood classification system. In the research, user-tagged moods were used to create a collection of lyrics, and metrics such as term frequencies and TF-IDF values were used in order to measure the relevance of words into different mood classes.

Term Frequency-Inverse Document Frequency (TF-IDF) representation of a document is a reweighted version of a BOW approach, which considers how rare a word when concerning a text and the overall collection the text within. In this approach, the importance of a term increases proportionally to its occurrence in a document; but this is compensated by the occurrence of the term in the entire corpus, which helps to filter out commonly used terms. Thereby, the TF-IDF vector model enables to assign more weight to the terms which frequently exist in the subject text, i.e., a song; but, not in the overall collection, namely corpus. Consequently, a valid combination between popularity (IDF) and specificity (TF) is obtained (Sebastiani et al., 2002).

TF-IDF score computed as the multiplication of two measures. For instance, considering the i^{th} word in the j^{th} lyric

Term Frequency will be the number of times word “ i ” appears in document “ j ,” normalized by the document’s length:

$$TF_{i,j} = \frac{|\text{word } i \text{ appears in lyric } j|}{|\text{lyric } j|} \quad (2.1)$$

Inverse Document Frequency will be a measure of the general importance of the word in the corpus by showing how rare is the term among all document set:

$$IDF_i = \log \left(\frac{\text{total number of lyrics}}{|\text{lyrics containing word } i|} \right) \quad (2.2)$$

Consequently, the TF-IDF for word i in lyric j will be calculated as:

$$TF - IDF_{i,j} = (TF_{i,j} \times IDF_i) \quad (2.3)$$

Zaenen and Kanters (2010) presented mood classification system for music by utilizing the TF-IDF metric on lyrics. In the study, the TF-IDF was used to calculate the words' relevance for identified moods, and high TF-IDF values expose powerful word's relevance to the mood. As a conclusion of the research, they confirmed that TF-IDF can be practiced efficiently to distinguish words which typically represent emotional aspects of lyrics.

POS, part of speech, tags also have been commonly used as content-based features, which are typically accompanied by a BOW analysis in the literature (Tzanetakis & Cook, 2000; Meyers, 2007). In this approach, words are separated according to grammatical classes, such as nouns, adjectives, and verbs. Wang et al. (2011) presented a music emotion classification system for Chinese songs, which were based on the lingual part of music by using TF-IDF and rhyme. In this study, they adopted Thayer's arousal-valence emotion plane with four emotion classes, such as happy, angry, sad, and relax, and thereby they created a combined approach by taking the part of speech (POS) into consideration. As a conclusion of the research, they reached 77% accuracy and claimed that both of the features, as well as the combined approach, are useful to build a classification model.

Another feature has been mostly practiced in the literature is Text Stylistic Features, which reflect the stylistic aspects of the language. For example, Hu and Downie (2010) evaluated text statistics by considering the unique words' number, the number of lines, and the number of interjections, such as "yeah" or "hey," as well as distinctive punctuation marks such as "!" within each text in their corpus. In another research, they compared all textual features, as well as the audio features in order to show the cases in which lyrics outperform audio considering mood classification (Hu & Downie, 2010a). As the outcome of this research, they found out the performance of text stylistic features are the worst among all features, except some emotional categories, such as hopeful and exciting.

At last but not least, various language packages were expanded to present semantic meanings in different emotional aspects by utilizing linguistic text features, which are based on psycholinguistic resources. Some of those lexicons measure words in several dimensions. For instance, Affective Norms for English Words, ANEW, (Bradley & Lang, 1999), and WordNet (Soergel, 1998), have been implemented for estimating the emotion values from texts in three dimensions, such as arousal (excited- calm), valence (pleasant- unpleasant), and dominance (dominated- submissive), and the documents are scored by averaging the individual ratings of

words. Other lexicons, such as General Inquirer (GI) or Linguistic Inquiry and Word Count (LIWC) have been used to label affective or psychological states of each word. For example, in GI, happiness was associated with a category, which consists of adjective tags, such as "Emotion," "Pleasure," "Positive" and "Psychological wellbeing" (Hu & Downie, 2010a).

Besides that, lexicon-based methods have also been used in some earlier studies of lyric analysis for languages other than English. For example, Cho and Lee (2014) used a manually built lexicon in the Korean language to extract emotion vectors for the recognition process. In another study, Logan and Salomon (2001) have focused on evaluating artist similarities of the songs by utilizing lyrics, and the categorized stems taken from news and lyrics.

Other particular favorite textual feature analysis approach is Probabilistic Latent Semantic Analysis, PLSA, (Saari & Eerola, 2014; Logan et al., 2004). In their research, Laurier and his colleagues (2008) employed TF-IDF weighting and applied PLSA in order to decrease dimensionality on the data representation. In the outcome of the research, even no significant improvement was observed, dimensionality reduction allowed better flexibility on their model.

Consequently, when considering the previous researches in the literature, it can be said that some particular word representation approaches revealing highly accurate outcomes have been commonly preferred and employed, such as bag-of-words (BOW), part-of-speech (POS), Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA). However, there are also some limitations regarding those approaches, such as high dimensionality, which leads to neglectation of the similarity between features and data sparsity (Bengio et al., 2001). Two factors mainly cause data sparsity. The first reason is the absence of a large-scale labeled training data, which restricts to build supervised models and causes a biased estimation. Secondly, natural language words are Zipf distributed¹⁶. Namely, most of the words resemble a few times within the corpus, or they can be out of the textual corpus (Guo et al., 2014).

Conventionally, supervised lexicalized Natural Language Processing (NLP) methods get a word, and then converted it into a feature vector by using a one-hot encoding (Turian et al., 2010). In this representation, the feature vector possesses the same size of the vocabulary, and solely one dimension is on; but unfortunately, the one-hot representation of a word cannot

¹⁶ <https://www.statisticshowto.datasciencecentral.com/zeta-distribution-zipf/>

handle with data sparsity problem, which leads to a sparse estimation of the model parameters for the word, that are rare or absent in the labeled training data. To overcome the restrictions, and to discover more effective and generalized representations, researches have studied on semi-supervised techniques for inducing word features by exploiting the numerous unlabeled data. As a contemporary NLP architecture, this technique helps the utilization of word embedding, which is a dense, continuous, and low-dimensional vector representations of words, which enables a similar representation for the words with similar meaning (Guo et al., 2014).

Word embeddings can be assessed as a class of techniques, in which each word is represented as real-valued vectors within a pre-defined vector space. Each and every word is mapped to one vector, and the vector values are learned in a way that resembles a neural network, and therefore the technique is often evaluated into the field of deep learning (Brownlee, 2017). The origin of word embeddings was created in order to develop better language modeling (Bengio et al., 2001). Word embedding has the simultaneous learning ability from the distributed representation of each word, namely the similarity between words along with the probability function for the word sequences were denoted with the representations. Similar words are supposed to be distributed close to one another in the vector space.

When comparing to BOW approach, word embeddings;

- Take place in an unsupervised learning paradigm having the capability to learn from large-volume unlabeled data through context-predicting models, such as neural network models, and spectral techniques like canonical correlation analysis,
- use pre-defined vector space, which leads to having a fixed number of dimensions (features) regardless of any increase in the count of unique words, and thus, it can deal with the curse of dimensionality; whereas dimension raises parallel with unique word count in BOW method,
- can build a semantic relationship between words since the closeness of word vectors is correlated with similarity in meaning; whereas BOW cannot reveal semantic relationships among words since it has binary word representation with two options, such as (0,1), which is based on word frequency

Previous works have practiced this representation and have confirmed the effectiveness of the word embedding features in several tasks, such as named entity recognition (NER), and thereby many investigators have benefited of word vectors to simplify and improve NLP applications (Collobert et al., 2011; Turian et al., 2010; Collobert & Weston, 2008). The detail usage of this representation will be explained further in the modeling section.

2.2.3 Emotion Recognition Using Features from Multiple Source

Even though previous investigations have proclaimed conflicting results for audio and lyrics analyses regarding Music-IR tasks (i.e., lyrics-based method outperforms audio-based method or vice versa), studies in the literature exhibited that both lyric-based and audio-based methods have accomplished satisfying outcomes.

On the other hand, some of the previous studies displayed that language and music complement each other in many different ways. For instance, while the music appears to induce emotions more intensely than ordinary speech, it does not reflect semantic meaning as language does (Mihalcea & Strapparava, 2012). The consideration has directed many kinds of researches towards multimodal approaches, namely combining features from different domains in order to enhance emotion recognition in music classification. Applying consolidated analysis of audio and facial expressions were the earliest attempts regarding emotion classification by using multimodal approaches (Cohn & Katz, 1998; Zeng et al., 2009), which have prompted to the usage of multimodal studies in other Music-IR classification tasks by mostly using the combination of audio and lyrics or audio and tags (Kim et al., 2010).

Considering emotion classification researches, Yang, and Lee (2004) generated the first study on combinations of features from text and audio domains with 145 song clips. One of 11 emotional classes depending on PANAS labels, were used for hand-labeling each clip. As the outcome, they saw that the addition of textual features develops the performance but not significantly. Yang et al. (2008) presented a bi-modal study by incorporating audio and lyrical features extracted from 1240 Chinese pop songs. The scholars intended to examine the utilization of lyrics, which potentially have valuable semantic knowledge, to defeat a probable emotion classification limit caused by the usage of audio features alone. For that purpose, 30-second fragments extracted from the middle of per song were used for the audio analyzing part, and BOW approach was employed for the text analysis. In the result, acoustic features alone

performed 46.6% classification accuracy, whereas combining audio with lyrics yielded 57.1% accuracy by increasing the model performance around 21%.

Laurier et al. (2008) assumed that even though the emotional aspect of songs can be reflected through musical features; a relevant emotional knowledge also can be conveyed by the lyrics, and they presented a hybrid classification model by combining lyrics and audio features into a single vector space, that allowed to use all features within one classifier. Music Information Retrieval and Natural Language Processing techniques were used to examine each feature both independently and also in a combined version. The outcome of the research showed that the model performance improved for happy and sad quadrants by 5% when compared to using solely audio features, but the accuracy did not change for relaxed and angry quadrants.

Hu et al. (2009) consolidated audio and lyrics for emotion recognition into 18 emotion categories. BOW approach with TF-IDF weighting operated for lyrics features after the stemming process, and lyrics traits are precisely blended with 63 audio traits before classifier training. The outcome revealed that although the multimodal approach improves the performance in identifying 13 out of the 18 mood categories, some emotion categories showed better performance without the feature combination. For instance, audio alone performs better for upbeat, desire, and happy, whereas lyrics perform the best accuracy for grief when it was used individually.

In 2010, Hu and Downie studied on the importance of lyrics in music mood classification by evaluating and comparing a wide range of text features, such as linguistic and text stylistic features, and then the best lyric features were combined with the features extracted from music audio (Hu & Downie, 2010a). The study's results displayed that combining lyrics and audio outperformed to the usage of each feature alone. Additionally, the examination of learning curves indicated that the hybrid system, which consists of both audio and lyric, needed fewer training samples to achieve the same or better classification accuracies. In the same year, they have made an extended version of their previous study while working with 5,296 songs for classification of those songs into 18 individual emotion categories (Hu & Downie, 2010b). In that study, the emotional classes were retrieved from the listeners' tags taken from Last.FM, by following multimodal approach with combining audio and lyrics features. As an interesting insight, the researchers observed that audio and lyrics have their particular benefits in the

specific mood classes. While lyric attributes fairly outperformed audio spectral features in seven emotion categories, such as cheerful, hopeful, exciting, romantic, anxious, angry, and aggressive; the audio features were more valuable in determining emotions in the 3rd quadrant of the valence-arousal space, such as calm.

McVicar et al. (2012) claimed that the predetermined emotion of a song inspires the musician for using certain audio features regarding harmony, timbres, and rhythmic characteristics, as well as the choice of lyrics. Therefore, they proposed an unsupervised learning approach by combining audio and lyrics features in order to identify common characteristics between them through computing the Pearson's correlation coefficient between each lyric and audio traits in V-A space. The outcome proofed the existence of some of the statistically significant correlation; yet, the absolute correlation value cannot exceed 0.2.

Mihalcea and Strapparava (2012) examined the connection between the musical and linguistic inputs and their affective role over 100 popular songs. They used Musical Instrument Digital Interface (MIDI) tracks of the songs to extract musical features, such as pitch, timbre, and intensity instead of employing audio signal analysis; while the textual features were extracted by BOW method to derive tokens from the corpus. In that research, crowdsourcing was used for data annotation to classify the song into six primary emotions of Ekman, and multilabel approach was followed. The experiment set was divided into three phases, such as focusing usefulness only the textual features, only the musical features, and the joint approach combining both domain features on the emotion classification task. The result showed that the joint model caused a reduction in the error rate by 31.2% when concerning the classifier using only the musical features, and by 2.9% when concerning the classifier using only the textual features. Consequently, through comparative experiments, they displayed that emotion recognition can be performed using either textual or musical features, and textual and musical features can be combined for reaching a developed accuracy.

Consolidating tags and audio features is another favored multimodal approach in Music-IR researchers. Turnbull et al. (2009) created the CAL500 data set, which consists of audio analysis, and semantic information from web documents, which are social tags, in order to examine tag classification. In this research, several algorithms were compared, such as kernel combination SVM, calibrated score averaging, and RankBoost. The research's outcome displayed that multimodal designs perform much better than unimodal approaches.

Using ground truth mood labels from AllMusic Guide, Bischoff et al. (2009) designed two experiments on the combination of emotion tags and audio signals to achieve a better result on emotion recognition systems. For each track, social tags collected from Last.FM and multi-dimensional audio feature vectors were produced. After that, Naive Bayes classifier and SVM classifier were trained for the social tags and audio vectors respectively, and a simple weighted combination approach was employed to create the joint model. In the first experiment, they used this approach to predict one of the five mood categories employed from MIREX, while in the second study, the approach is utilized for the prediction of the V-A model's quadrants. The outcome demonstrated that tag features were more informative than audio, while the multimodal strategy exhibits better performance in both experiments.

More recently, Schuller et al. (2011) analyzed regression of musical mood in continuous dimensional space by consolidating of audio, lyrics, and tags on a set of 2.648 UK pop songs. Another multimodal approach for the music emotion recognition (MER) field was introduced by Panda et al. (2013) through combining information from audio, MIDI files, and lyrics. The dataset was employed from the AllMusic database and organized into five emotional tags proposed by MIREX Mood Classification Task. As each song may have more than one label, the final emotion for each song was assigned according to the most outnumbered label. Emotion assignments were processed according to Paul Ekman's model. For feature extraction process, Marsyas, MIR Toolbox, and PsySound Audio frameworks were used to extract various audio features, and MIDI Toolbox¹⁷ utilized to extract MIDI features. Lastly, textual features were extracted by using Jlyrics, which is a common lyric analysis framework implemented by Java, as well as using an NLP technique based on WordNet. Finally, several supervised learning algorithms were used to test classification accuracy as support vector machines (SVM), decision trees, Naïve Bayes, and K-Nearest Neighbors by using MATLAB and Weka. Study results depicted that lyrical features performed worse accuracy compared to audio and MIDI features, but using the combined features developed the results significantly (Hu, 2010).

Duggal et al. (2014) attempted to predict the emotions derived from songs as a multilabel classification problem through the combination of musical and lyrical features. For this examination, 183 songs were gathered from different genres, and the annotation process

¹⁷ <https://github.com/miditoolbox/>

was conducted by online users from different professions according to felt emotion(s). For feature extraction process, topic modeling was practiced by employing Latent Dirichlet Allocation (LDA) for textual inputs, while a set of high-level musical features including Acousticness, Danceability, and Instrumentalsness were extracted by Spotify API, which is a web-based API, was used to extract audio features from the web. The result of the study demonstrated that the combined features approach performed a better result by 8.9% than acoustic-only classification and by 9.4% than lyrics-only classification.

Consequently, the multimodal approach depending on a combination of different features has motivated many of MIR researches since this approach may lead to improvement on recognition of some emotions conveyed by music, and hence, may constitute a better classification system. Even though numerous studies exhibited relative performance gains and complementary results through a combination of features from different domains, it should be considered that such joined strategies upon the content classification based on the emotion of music have been studied only for the past few years. Thus, there still are contradictory judgments on which feature(s) can be more beneficial when considering emotional recognition and prediction in music.

The following table, Table 2.1, displays the detailed examinations across several kinds of research in the literature while considering the various basis.

Table 2. 1 Subsequent MER & MIR Research Examples from the Literature

Reference	Data	Emotional Models & Annotation	Feature Selection	Methods/ Algorithms	Outcomes
Yang et al. (2008)	Database of 1240 Chinese pop songs	Russel's/Thayer's V-A model Songs were hand-labeled into 4 emotion classes: happy, angry, sad, and relaxing	Low-level audio features by Marsyas and PsySound Text-mining by BOW approach Lyrics are preprocessed with IR operations such as stopword removal, stemming, and tokenization	Support vector machine (SVM) is adopted	Enhanced the classification accuracy of valence from 46.6% to 57.1%
Laurier et al. (2008)	1000 songs, with a uniform distribution over the four quadrants	Russel's/Thayer's V-A model with 4 quadrants: happy, angry, sad, and relaxing Songs were labeled using Last.FM tags & lexical database WordNet-Affect used to eliminate synonyms	Timbral, rhythmic, tonal, and temporal audio features Lyric similarity, Language Model Differences (LMD), and Latent Semantic Analysis (LSA) applied for lyric features	SVM Random Forest Logistic Regression	5% improvement for "happy" & "sad" quadrants No significant chance for "angry" & "relaxed" quadrants

Reference	Data	Emotional Models & Annotation	Feature Selection	Methods/ Algorithms	Outcomes
Hu et al. (2009)	Nearly 3000 songs	18 emotion classes from Russel's model were used with respect to the social tags from Last.fm WordNet-Affect used for filtering	63 spectrum-derived audio features BOW approach with TF-IDF weighting	SVM classifier	Performance improved 13 of the 18 classes, but single feature usage depicts the best result for 5 categories (where lyrics alone outperforms audio & vice versa)
Bischoff et al. (2009)	1192 songs	5 MIREX mood clusters V-A model	240-dimensional audio feature vectors social tags collected from Last.fm used as textual features	Naive Bayes Classifier for Social Tags SVM Classifier for audio data	Multimodal approach overperformed considering single features usage
Hu & Downie (2010)	5,296 popular songs	18 emotion classes from Russel's model Tags derived from Last.fm Multilabel approach employed	63 spectral Audio features by the MARSYAS Three textual features as basic, linguistic, stylistic features Features based on N-grams of Content Words (stopwords eliminated) No stemming applied	Features Based on General Inquirer & Features Based on ANEW and WordNet SVM classifier	Lyrics outperformed audio traits, for seven categories "romantic," "angry," "cheerful," "aggressive," "anxious," "hopeful," and "exciting." Audio performed better than all lyric feature types in "calm" mood category.

Reference	Data	Emotional Models & Annotation	Feature Selection	Methods/ Algorithms	Outcomes
McVicar & Freeman (2011)	Million Song dataset & MusiXmatch lyrics database	Russel's V-A model	65 spectral, percussive, harmonic and structural features extracted by EchoNest API BOW approach with TF-IDF weighting	Canonical Correlation Analysis (CCA) between lyrics & audio features	There exist weak but highly significant correlations between lyrical and audio features.
McVicar & Bie (2012)	Million Song Dataset with the MusiXmatch lyrics data ANEW dataset	Russel's V-A model Social tags from Last.FM	Audio features extracted from the million-song dataset BOW approach with TF-IDF weighting	Pairwise 2-dimensional CCA and 3-dimensional CCA between the tag space, lyrics & audio representations	Correlations exist between audio, lyrical and tag features
Mihalcea & Strapparava (2012)	MIDI Files of 100 popular songs	The six basic emotions proposed by Ekman: Anger, Disgust, Fear, Joy, Sadness, Surprise. Crowdsourcing for data annotation from Amazon Mechanical Turk service	Song level, Line level & Note level audio features Textual Features (1) unigram Features obtained by BOW representation (2) lexicon features	Linear regression with Pearson correlation index	Multi-modal approach improved accuracy over usage of singular features up to 31%

Reference	Data	Emotional Models & Annotation	Feature Selection	Methods/ Algorithms	Outcomes
Panda et al. (2013)	AllMusic database & MIDI Files as 30-second mp3 tracks	5 emotion clusters defined in MIREX. Ekman's emotional model	177 standard features and 98 melodic features from Marsyas, MIR Toolbox and PsySound, 19 structural & semantic textual features extracted by Jlyrics framework, Synesketch framework based on WordNet	C4.5 Decision Tree, Support Vector Machines (SVM), Naïve Bayes, & K-NN	Employing the multi-modal approach developed the study's outcome from 44.3% to 61.1%
Duggal et al. (2014)	183 unique songs	GEMS-9 Emotional Model	Spotify API used to extract 13 high-level musical Topic modeling by employing Latent Dirichlet Allocation	SVM classifier	Combined features approach exceeds that of acoustic-only classification by 8.9% and that of lyrics-only classification by 9.4%.
Tekwani (2017)	7396 songs from the Million Song Dataset (MSD) created by LabROSA	Happy or sad from Russell's 2D representation of valence and arousal	Features used for Audio signal analysis: Low level like MFCC & Timbral & Pitch features & Descriptive features from Spotify like energy, acousticness,	Principal Component Analysis (PCA) & Recursive Feature Elimination (RFECV) with a Random Forest Classifier & XGBoost, Gradient Boosting	Best valuable features: danceability, energy, speechiness and the number of beats Highest accuracy is 75.52 % with a Gradient Boosting Classifier

Reference	Data	Emotional Models & Annotation	Feature Selection	Methods/ Algorithms	Outcomes
			speechiness, danceability	Classifier, ADABOOST, SVMs & Naive Bayes classifier	
Y. An et al. (2017)	<p>Four different audiovisual emotion datasets:</p> <p>4552 songs with Chinese + English Lyrics &</p> <p>3316 songs with only Chinese Lyrics</p>	<p>Two different labeling style: Thayer emotion model with three emotion categories: contentment, depression & exuberance</p> <p>Valance dimension with positive and negative labels</p>	<p>Information containing the singer, music name, lyrics & the category of the music, were crawled by Python's Scrapy-a framework</p>	<p>To classify music by emotion based on lyrics, Naive Bayes algorithm was used</p>	<p>The highest final accuracy was approximately 68%.</p>
Miroslav et al. (2017)	<p>Music dataset provided for the MediaEval's EiM task</p> <p>431 audio excerpts for training & 58 of full songs for evaluation set</p>	<p>Russel's V-A emotional Space, Labeling into the range of [-1, 1] for both arousal and valence</p>	<p>Baseline audio feature-set including MFCCs, spectral features, flux, centroid Raw audio feature, the Mel band features extracted by the Librosa Python library</p>	<p>Convolutional (CNN) & Recurrent Neural Networks (RNN) methods</p>	<p>The best result reported on this dataset was reached by RMSE of 0.2 for arousal & 0.27 for valence</p>

2.3 Emotion-based Analysis and Classification of Music

Determining the appropriate and relevant machine learning algorithm is another significant part of building predictive models through learning from, and making a prediction considering the data. The aim and use of designing a predictive model of emotions are essential when selecting which stimuli to include in the modeling framework (Song, 2016). As No Free Lunch Theorem¹⁸ defends, in predictive modeling, there is no one algorithm works best for all problems and outperforms the others, which lead to variation in literature according to the main aim of studies.

2.3.1 Model Building by using Audio Features

In the literature, even various algorithms were employed for audio modeling, such as SVM (Schuller et al., 2010; Song, 2016), regression (Eerola et al., 2009), k-NN (Saari & Eerola, 2014), neural network (Kim et al., 2008), Gaussian Mixture Models (Lu et al., 2006), and random forest (Vuoskoski & Eerola, 2011). Audio modeling can be summarized up under two main topics, such as categorical emotion classification and parametric emotion regression (Kim et al., 2010).

In one of the earliest MER investigation on audio signals, Feng et al. (2003) used two musical attributes, and 23 music pieces to classify music into four emotional categories by applying neural networks, which resulted in recall and precision score of 66% and 67% respectively. In the same year, Li and Ogihara employed acoustic traits, such as timbre and rhythm of 499 clips from several genres to train support vector machines (SVMs) in order to classify music into one of 13 mood categories. As the conclusion of this research, they attained 45% accuracy performance.

Lu et al. (2006) utilized 800 classical music clips extracted from a data set of 250 music pieces to generate a model of emotion by using acoustic features, such as intensity, timbre, and rhythm. The emotion was identified with Gaussian Mixture Models (GMMs) for the four principal quadrants on the V-A space. Although the algorithm reached 85% accuracy, this outcome was regarded with caution because the multiple clips' extraction process from the

¹⁸ <http://www.no-free-lunch.org/>

same song records was not explained adequately. In 2007, with the first-time usage of audio features for music emotion classification in MIREX, Tzanetakis achieved 61.5% accuracy performance by employing an SVM classifier fed by the features, such as MFCC, spectral shape, centroid, and roll-off (Tzanetakis, 2009).

Korhonen and his colleagues have introduced a methodology that applies system-identification techniques on six classical music pieces to represent the music's emotional content as a function of time and musical features by using Russell's V-A model and launching MER as a continuous intricacy (Korhonen et al., 2006). In the conclusion of this study, the average R2 statistic found as 21.9% for valence, and 78.4% for arousal. Additionally, Yang et al. (2008) have evaluated emotion recognition from music as a continuous modeling problem (Yang et al., 2008a). Each music piece was mapped to a point in the V-A plane, and several classification techniques were applied on the dataset of 189 audio clips by utilizing only standard audio features. The best-attained results regarding the R2 metric were 28.1% for valence and 58.3% for arousal. Yang's dataset also studied by Panda and Paiva (2011). In their study, MIR toolbox, Marsyas and PsySound were used to extract both standard and melodic audio features, and as a result, 63% and 35.6% accuracy prediction was produced for arousal and valence, respectively.

In order to build personalized emotion classifier, Mostafavi et al. (2014) practiced on 100 audio clips originated from numerous film and video game sounds and extracted audio features by using MIRtoolbox. A set of emotion classifiers were trained by using the extracted features, which have been tagged by volunteers, and several classification algorithms, which are SVM, k-NN, Random Forest, and C4.5 were developed to detect the ideal method. Even though SVM showed the lowest accuracy score among other algorithms, SVM, as well as Random Forests, delivered the best average F-Score indicating a higher recall and precision scores by 90%.

In 2017, Tekwani tackled music mood classification from an audio signal perspective by classifying music as happy or sad through audio content analysis. In this investigation, 7396 songs were hand-labeled into two distinct categories. Spotify API was used for extraction of some audio features, such as Speechiness, Danceability, Energy, Acousticness, and the performance of different algorithms, such as Random Forest, XGBoost, Gradient Boosting, AdaBoost, Extra Trees, SVM, Gaussian Naive Bayes, and K-NN were evaluated and compared.

The result of the experiment displayed that ensemble classifiers like GBoost, Gradient Boosting Classifier, AdaBoost, and Random Forests performed better than SVM and Naive Bayes classifiers with the highest accuracy 75.52 % by a Gradient Boosting Classifier.

2.3.2 Model Building by using Textual Features

Kim and Kwon (2011) studied lyrics-based emotion classification using feature selection by partial syntactic analysis (PSA). In their study, they defended that it is challenging to classify emotions accurately by adopting the existing music emotion classification methods using mostly the audio features associated with music melodies, such as tempo, rhythm, tune, and musical note, but lyrics can exhibit stronger relation with emotion. Namely, songs make listeners feel emotionally different according to the lyrical contents, even when melodies are similar. Therefore, the researchers utilized the emotion features extracted from 425 random Korean-language songs. Then, they employed supervised learning methods, including NB (Naive Bayes), which is the most representative probability model and expects robust independence among learning features; SVM (Support Vector Machine), which reveals the best when to classify data by difference, and; HMM (Hidden Markov Model), which exhibit the information on time flow, to classify the emotions of song lyrics. The outcome of the research showed that SVM performed better than other proposed lyric attribute-based systems with the accuracy rates of 58.8% and 53.6% considering the emotion category division in 8 and 25 emotions, respectively.

Chi et al. (2011) built research on 600 pop song dataset, which mood rated by 246 participants, to evaluate the contribution power of the lyrics as well as the audio regarding overall valance and arousal mood ratings of each song. The study was designed under three section according to the utilized features, such as lyric only, audio-only, and the combination of both. The linear regression model was employed to build a statistical analysis, and the research outcome revealed that lyric text feature achieves a higher accuracy (82%) than audio features (75%) with respect to valence rating, whereas audio performs a bit better-considering arousal rating.

Teja (2016) attempted to find the underlying mood of albums in order to recognize and recommend similar albums to users, while using five emotion categories, such as happy, sad, anger, grief, and romantic. He employed topic modeling and used Latent Dirichlet Allocation

(LDA) which follows BOW approach by reflecting each word as a token and N-Grams algorithms, which are similar with LDA method except the usage of the N-words combination for topic assignment. As a classification method, Naive Bayes classifier was selected, and the classifier was trained by using a word list consisting of the most frequently occurred positive and negative words. Then, each word from the album lyrics was classified either as positive or negative according to the songs' polarity. At the end of the research, 89.4% accuracy has revealed.

2.3.3 Semi-supervised Learning by using Word Embeddings

Lyric-based classification of music can be assessed as a text classification problem, which is the main research area for natural language processing (Qi, 2018). As can be observed in the previous researches, this domain generally has been formulated as a supervised learning problem through establishing classification and regression algorithms proven to enable reliable outcomes, such as support vector machines (SVMs) and Naive Bayes.

Because of some severe limitations on the previously employed approaches, like data sparsity which explained in the previous section, with the progress of machine learning techniques in recent years, investigators have also attempted to generate more complex models, such as convolutional neural networks, which develop the potentiality of training a much broader dataset with outperformed classification accuracy (Kim, 2014; Senac et al., 2017). Likewise, some researches have displayed that neural network-based language models perform better than N-gram models (Schwenk, 2007; Mikolov, 2011).

To be able to utilize neural network algorithms for text classification tasks, the input array of words should be transformed into an array of vectors, so-called a matrix, which is designated a word embedding in natural language process (NLP). The word embedding selection may influence neural algorithm performance. While the single simplest word embedding, which uses an arbitrary random vector for each word, has depicted satisfactory results in many researches, Word2Vec state-of-the-art in this area, which is neurally-trained word representation seizing the semantic relationship between words (Qi, 2018).

In 2013, Mikolov et al. (2013) have published the Word2Vec toolkit, which is the mostly employed pre-trained word embedding model in the literature, that eventually made word

embedding state-of-the-art in NLP. Besides Word2Vec, different pre-trained word embedding models also exist, such as doc2vec, GloVe, and fastText.

Before to examine Word2Vec model particularly, it is essential to understand what does the pre-trained model, which is the concept introduced by Collobert and Weston in 2008, imply. In this study, the researches leveraged from unlabeled data in NLP task to deal with both costly character of the labeling process, and the abundant nature of unlabeled data by designing a single convolutional neural network architecture (Collobert & Weston, 2008). The network encompasses a package of language processing predictions including POS Tags, named entity tags, chunks, semantically related words, and the probability, which makes the given phrase valuable for both semantic and grammatical manners by utilizing relevant language models. The entire network is trained concurrently by supervised algorithms, which proceeds on the labeled data, except of the language model, which was learned from the entire Wikipedia website, namely unlabeled corpora, through approaching the system as an unsupervised task. Thereby, they have presented a semi-supervised approach for NLP through jointly training supervised methods on the labeled data, as well as unsupervised tasks on the unlabeled data.

Utilization of word embedding, and pre-trained models have gained popularity in the literature along with the publication of Mikolov, which originates the Word2Vec technique in 2013. Mikolov et al. (2013) have introduced methods to scale the vector representation quality with the aim of not only similar words tend to be close to each other, but also the words could reflect the similarity in multiple degrees. They performed simple algebraic operations on the words by using a word offset technique, and the result exhibited that the similarity of word representations proceeds beyond simple syntactic regularities. For instance, vector (" King") – vector ("Man") + vector ("Woman") appears in a vector that is closest to the vector representation of the word Queen (Mikolov, 2011). Thereby, the semantic relationship between those words has been represented correctly, and the computation of high dimensional word vectors from a much bigger data set has become achievable by reducing computation complexity.

Consequently, Word2Vec takes a text corpus as input and creates feature vectors, which are distributed the numerical representation of word features in return correspond to the word in the corpus. Besides, it is capable to group the vectors of similar words in the vector space without human intervention by training words against other words, which make neighbor them

in the input corpus. For doing that, there are two algorithms known as the continuous bag of words (CBOW) and Skip-gram. While CBOW uses context to predict a target word, skip-gram utilizes a word to predict a context-depicted in Figure 2.7 referenced from "Efficient Estimation of Word Representations in Vector Space" prepared by Mikolov and his colleagues.

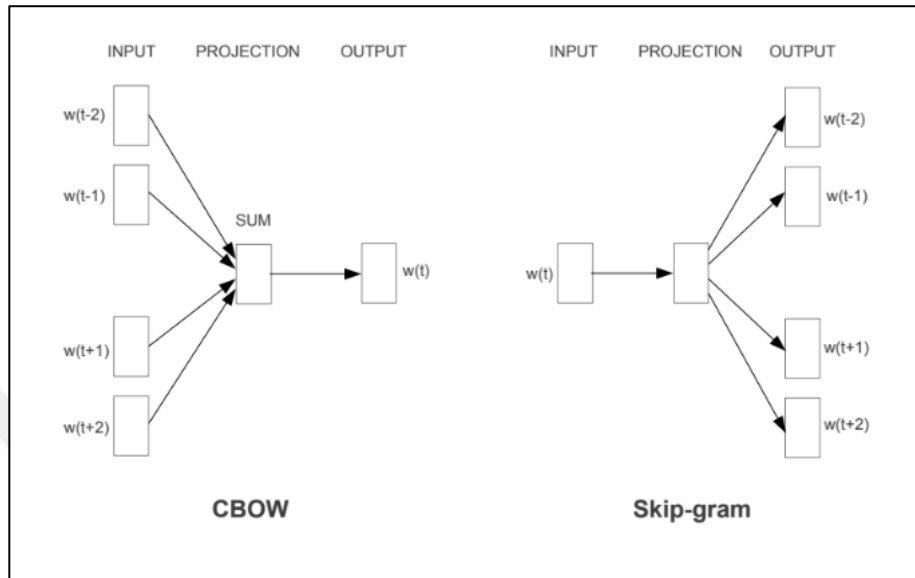


Figure 2. 6 Word Representation in Vector Space

In Music Information Retrieval, semi-supervised learning has been used firstly in 2004 by Li and his colleagues to recognize similar artists by utilizing both lyrics and acoustic data (Mikolov et al., 2013). They extracted content-based features including spectral and timbral attributes, and as lyric features, BOW with TF-IDF weight, POS tags, lexical and stylistic features were extracted, as well as the function words by using a small set of data consisting of 56 songs. The similarity among artists was determined subjectively by examining similar artists page of All Music Guide to be able to have a ground truth data and clusters, and a large number of neighbors of them were selected considering this relation. In this research, a small set of labeled samples was used for the creation of seed labeling in order to build classifier models, which improve themselves by utilizing unlabeled data. Support Vector Machine was used to train a supervised classifier, which distinguishes each cluster from the rest. Besides, a co-updating approach was used, which depends on the usage of labeled samples to train a classifier, and then uses this classifier to predict labels for the unlabeled data. The experiment result

showed that a small number of labeled samples with multiple sources can be used to build an automatic similarity recognition system.

In another research, Wu et al. (2013) attempted to solve a genre-specific MER problem by proposing a new model, which is called "SMART" by using two kinds of auxiliary data, which are unlabeled audio data and social tags. They followed the assumptions defending that songs, that have similar contents tend to have similar emotional labels (Li et al., 2004), and emotion of each song is similar to its neighbor (Chapelle et al., 2006). This research has proved that even though the labeled number of songs is limited, emotion prediction and assignment is possible by propagating supervision knowledge from labeled to unlabeled data. In this research, the Million song data set were used, as well as a large amount of Pop music data, which consists of several real-world datasets created for evaluation purpose. For feature extraction phase, several audio features, including rhythm, loudness, and timbre were gathered, while the social tag data was represented by two different methods, which are the weighted summation of tags' emotion values for each song, and BOW with TF-IDF weighting, by clustering tags into 11 categories. The proposed method, SMART, compared to both graph-based semi-supervised learning (GSSL) method and support vector regression (SVR) method with tag refinement. Different experiment designs were created through using audio and tag featured individually, and also with their several combinations. The emotion predictions of all methods were evaluated by Mean Squared Error (MSE). The study result showed that SMART method trained with only 10 labeled instances, is as capable as support vector regression trained with 750 labeled songs. Consequently, the researchers have proven that a limited amount of labeled data indeed can be used to estimate a large amount of unlabeled data.

In one of the more recent researches, Qi (2018) studied music classification based on textual corpora by implementing two approaches, such as TF-IDF, that relates frequency analysis and Word2Vec, which uses a convolutional neural network algorithm. Both implementations were performed by using Python Scikit-learn library. In the first approach, a word frequency-based model was employed by using Multinomial Naïve Bayes classifier performing on TF-IDF vectorization of songs. The training and testing sets were created randomly as the test set has 10% of the overall data averagely. In the result, a model was created with accuracy shifts around 60%. In the second approach, a smaller set of data was trained with Word2Vec representation, which employs a neural network to fine-tune the word embeddings

while training. The tests of the model on the randomly selected data set displayed the accuracy performance changing mostly between 0.65 and 0.80.



CHAPTER 3

METHODOLOGY

In this chapter, the methodology followed in this research is explained.

First of all, we give the details of the collection process of the data consisting of 1500 songs in total. Subsequently, the selection of the emotion categories and the emotion annotation process are explained. After that, we illustrate the feature selection and extraction processes, while concerning the generation of both audio and lyric feature vectors, that are valuable inputs to build emotion classification models. In this part, we also explain the data preprocessing methods we used, which prepare our corpus for the detailed analyses. Finally, the model-building processes consisting of four different classification experiments are clarified. In the first two experiments, we utilize audio and textual features extracted from music individually, and various supervised approaches are employed by utilizing the labeled song data. Besides that, we attempt to generate semi-supervised models through using both labeled lyric data and unlabeled big data, which are explained in the third and fourth experiments, where bi-modal and multimodal approaches are applied, respectively.

Figure 3.1 displays an overview of the process flows for our proposed emotion detection system. The study starts with the song data collection process, consisting of the song lyrics, the song metadata for audio information retrieval task, and the tracks of the songs, to further the research into the emotion annotation phase, in which the research participants labeled the songs into respective emotional categories. As the next step, textual features are extracted using two

different approaches, which are TF-IDF and Word2Vec, whereas audio features are gathered by information retrieval from Spotify.

The flow continues through the model building step regarding four experimental approaches. Model 1 and Model 2, that were symbolized as “M1” and “M2” in the diagram, use the different attributes of the songs from a single resource and build several models on the labeled data. Model 3 and Model 4, which are “M3” and “M4” respectively, utilize both labeled song data and big unlabeled data to design and compare bimodal and multimodal machine learning approaches, respectively. While Model 3 utilizes textual features derived by Word2Vec method; Model 4 uses a merged feature set, which consists of both audio and Word2Vec textual features.



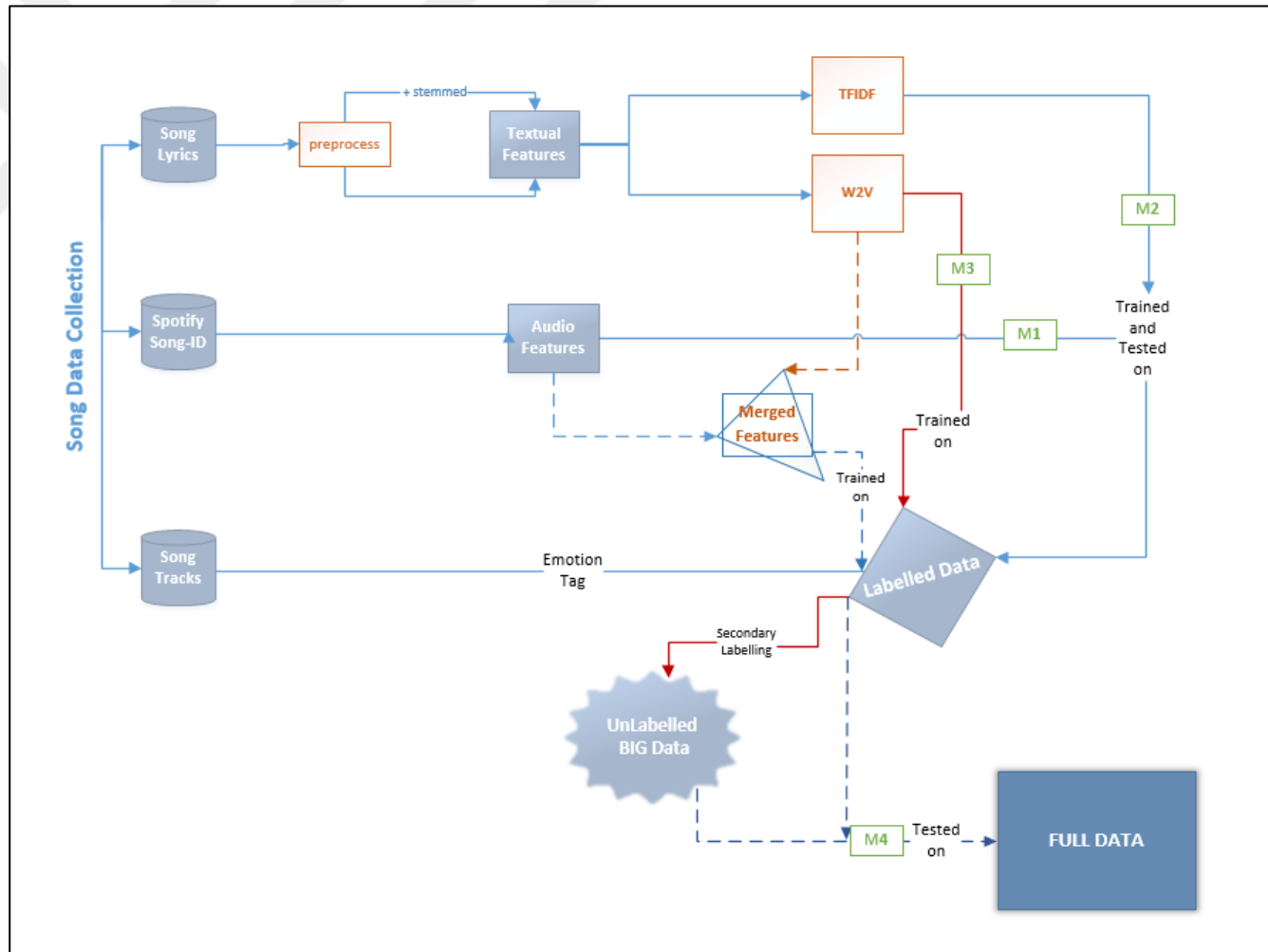


Figure 3. 1 Analysis Flow Diagram

3.1 Dataset Acquisition

For the aim of developing an automated emotion recognition system from music, the first step we took was the creation of the primary resource utilized by human annotators, before the feature extraction process. Therefore, in order to create our ground-truth song dataset, first of all, 127 famous Turkish singers were determined from the several music pages, such as DreamTurk and D&R webpage, while considering their widespread fame and increasing popularity, especially in the last five years. After that, we chose averagely 10 to 15 songs from each music artist, and the first version of the corresponding list of song data was generated.

As the second step, we reached music tracks of the selected song dataset to be able to create content data for the annotation process, which requires human participation to extract perceived emotion. According to 2016 Music Consumer Insight Report, that is based on a global survey conducted by Ipsos across 13 countries, YouTube has been chosen as the most used music service and the streaming platform, and 82% of all participants presented YouTube as the main reason for their website usage (Kim et al., 2010). Relying on this information, in this research, mp3 tracks of the related songs were accessed and downloaded by using a YouTube converter in order to generate music archive for the labeling process. For this collection process, we paid attention to reach high-quality song tracks, which do not include any additional parts, such as advertisement videos. Thereby, the best possible music records were supplied to the annotators to eliminate any record distractions during the listening process.

After that, the lyrics of each respective songs were collected from various websites to be able to constitute mathematical models of emotion expressed by textual information. Since a unique database consisting of all songs was not found, the lyrics were gathered from several online music platforms by using both custom code and manual effort. Then, each lyric was organized as individual documents under a shared folder to create the corpus data before the feature extraction process, which make us be able to see the context effect of the songs on emotion detection.

After the collection process, we wanted to be sure that the assembled data was convenient enough to apply a model on it. Thus, we performed some elimination over the collected data to be able to increase the reliability of the secondary context data, namely the

lyrics of songs. In that respect, the songs which belong to Turkish artists; yet, consisting of phrases performed in another language were removed. Also, some of them were adapted version of the original songs such as remix and cover versions of the tracks. Therefore, we eliminated such songs displaying the mentioned attributes in order to have a robust dataset considering annotation and feature extraction processes.

Lastly, Spotify¹⁹ Song-ID of the remaining songs, that is the unique identifier for each song on the Spotify platform, was utilized to further the audio feature extraction process. The tracks, which cannot be found on Spotify, were removed from the final set. Consequently, the data cleaning process left 1500 different songs belonging to various genres, which are mainly Pop and Rock, as the input source for the modeling framework.

3.2 Selection of Emotion Categories and Annotation Process

Regarding the employed emotion representation in the previous investigations, it can be said that emotions perceived and reflected through music, have been one of the much-debated subjects in MIR20, music psychology, and also MER research domains. When a considerable amount of work has been examined, we decided to adopt Russel's circumplex model in this research, which represents emotions as distinct categories on valence-arousal dimensional space. This representation exposed a mutually exclusive emotion set on the emotion space, which led to better separation between diverse emotional categories, and thus, it has appeared as one of the most comprehensive representations among the various emotion models.

Four primary emotion categories that as "joyful," "sad," "relaxed," and "angry" were chosen as emotion tags considering their universal root and coverage. We believe, those four categories embrace the four distinct parts of the two-dimensional plane. In valance respect, "Joyful" and "Relaxed" tags reflect quite positive moods, whereas "Angry" and "Sad" categories have relatively negative emotional valence. On the other perspective, "Joyful" and "Angry" tags have a higher arousal level when compared to "Relaxed" and "Sad" respectively since they generally exhibit more active emotional judgments on the perception of people.

¹⁹ <https://developer.spotify.com/documentation/web-api/reference/tracks/get-track/>

Table 3.1 displays the emotion tags, which have the relevant sub-options under each label in order to not to restrict the judgments of annotators on perceived emotions through music.

Table 3. 1 Tags with Sub-categories

JOYFUL	RELAXED	SAD	ANGRY
Happy	Calm	Sorrowful	Aggressive
Delighted	Meditativeness	Miserable	Anxious
Excited	Serenity	Melancholic	Nervous
Aroused	Satisfied	Depressed	Fearful
Astonished	Glad	Bored	Annoyed

As this study intended to constitute an automatic classification system by utilizing both audio and lyric features extracted from music, as well as human cognition on music, the annotation process was one of the most crucial steps in our research design. As explained in the previous section, after removing redundant sources, we have come up with 1500 Turkish songs were prepared to be labeled.

Before the annotation process, a number was assigned to each song to create an anonymous data frame, and the order of songs was changed randomly. Namely, since we aimed to have a more reliable labeling process, the song data were supplied to the participants in random orders without song or artist information depicted in order to prevent listening prejudice resulting from previous acknowledge and also hearing songs which belong to the same artist over and over again. After that, the primary datasets were divided into three sub-sections, and the sub-sets were shared with each annotator in the respective order. Besides, before sharing the data for labeling, a roadmap for the annotation process was given to each participant, and they were asked to listen at least 60 seconds of each song to be able to construct more comprehensive emotional perspective on songs. Eventually, the songs were hand labeled into at least one of the four emotion categories by four human annotators who are all undergrad

Turkish students, whose ages ranged from 20 to 28; yet, coming from the different educational, professional and socio-economic backgrounds.

In the annotation process, annotators selected the perceived emotion by assigning "1" for the corresponding emotional category. However, there was no strict limitation on the number of assigned labels to the songs. Namely, annotators were free to select more than one emotional category if they cannot perceive one dominant emotion through the song.

SONGS	EMOTION			
	JOYFUL	RELAXED	SAD	ANGRY
1438			1	
1263		1		
1187	1			
1202	1	1		
1357				1
1309			1	

Figure 3. 2 A partial example for the labeled songs

Despite that, in the end, there were only a few songs labeled by maximum two different emotions, such as joyful and relaxed, and almost all songs were mapped with one particular emotion regardless of the annotator. An example of the annotated songs can be seen in Figure 3.2.

After all labeled data was collected from each participant, one united labeled song dataset was created, and this set was normalized by the sum of all labels considering to all emotions will be equal "1" for each song. Thereby, each song has had a weighted probabilistic score reflecting each emotion categories, as can be seen in Figure 3.3.

Song_ID	Artist	Song_Name	Track_EL_JOYFUL	Track_EL_RELAXED	Track_EL_SAD	Track_EL_ANGRY
3b6DyYGSUWO1Tmq2Plx3e7	Can Bonomo	Meczip	0.75	0.00	0.00	0.25
5FrsFlteBq3DuloCv5fKAp	Can Bonomo	Resmini Görünce	0.25	0.25	0.25	0.25
7zW2jqvdpCNJscJP1rvWON	Can Bonomo	Şaşkın	0.50	0.00	0.25	0.25
5sI6AXOjGddXnlSoeCQ57A	Candan Erçetin	Bahane	0.80	0.20	0.00	0.00
4VJ2qEoMOXYyhNPHBVH8xs	Candan Erçetin	İster Sallan Gez	1.00	0.00	0.00	0.00
05cxbMplv4rliL45Ealube	Candan Erçetin	Kırık Kalpler Durağında	0.00	0.25	0.75	0.00

Figure 3. 3 A portion from the labeled song data- After normalization

When the descriptive analysis was performed on the labeled data considering the emotional agreement of the participant, we observed that 14% of the songs were perceived precisely under the same emotion category by all annotators, whereas at least 2 of the participants were agreed on the emotional tag by 91% rate regarding 1500 songs. Besides, "sad" came out as the most agreed upon emotion category with 59% agreement rate by all participants, while "angry" created quite adverse outcomes, and exposed as the least agreed on emotional category.

Ultimately, we removed the emotionally confusing songs since it is difficult to determine which emotion category they belong to, before moving on the feature extraction process. Hence, the noisy song data, which cannot go beyond a certain threshold, namely do not reflect any particular emotion were eliminated from the annotated data, which correspond to almost 17% of the original labeled data, and thus, the dataset was reduced to 1246 songs in total. Table 3.2 exhibits the summary of ground truth data collection after all data annotation and elimination processes.

Table 3. 2 Summary of ground truth data collection

Emotion	Number of Songs
Joyful	344
Relaxed	284
Sad	549
Angry	69
Overall	1246

As a consequence, with the help of direct annotation process, we reached the labeled music content dataset, which is suitable for training and classification by the application of various machine learning algorithms.

3.3 Feature Selection and Extraction

Selection of both audio and lyric features which, are valuable model inputs, is a quite significant step to be able to automate the classification of songs into the selected emotion categories. To distinguish the features used in various supervised learning algorithms, first of all, we investigated previous works in the literature. As mentioned in the literature review section, several investigations performed the use of various features.

In this research, we collected audio music features belonging to the songs from Spotify through using a Web API. On the other hand, we applied text-mining for the lyric data in order to extract valuable text features for the model building process. The details of the feature collection and extraction process, and also, the result selected features were explained exclusively in the following parts.

3.3.1 Audio Feature Selection

In this research, the audio features, such as tempo, rhythm, energy, and acousticness of each selected music track were retrieved through using Spotify Developer Tools²¹. Spotify Web API endpoints, which depend on simple REST principles, render JSON metadata wherein music artists, tracks, and albums from the Spotify Data Catalogue.

Spotify is one of the most known music platform respecting audio feature collection, especially considering recent investigations. For instance, Tekwani (2017) studied on Million Song Dataset containing audio features and metadata for tracks, and they manually labeled more than 7000 songs, as either happy or sad. Besides, they fetched characteristic features like Energy, Danceability, Speechiness, and Acousticness by using Echo Nest API, which also is used as a part of Spotify's Web API (Gabrielsson & Lindström, 2001).

²¹ <https://developer.spotify.com/>

Further researches examined in the literature suggest that there is no unique dominant feature, but rather many acoustic features play a role regarding in determining the emotional content of the music. Even though, still there are some questions having no consensus on the answer, such as what aspects of the musical signal made people able to perceive emotions, and which features can be more valuable regarding emotion classification.

With all the consideration of the previous researches, in this study, we accessed 13 distinct audio features for each song track, including danceability, loudness, valence, and more by using Spotify Developer Platform. The Spotify audio track features consist of both high and low-level musical characteristics belonging to the songs. The high-level features comprise of several low-level features in a composite manner. For example, acousticness consists of tempo, rhythm, stability, beat strength, and overall regularity. Likewise, energy is constructed from timbre, onset rate, dynamic range, general entropy, and perceived loudness.

In this respect, the unique Spotify ID's (URI's) of more than 1500 tracks belonging to 127 artists were archived by manual collection process. This process required a bit effort; however, according to the best of our knowledge, it was one of the best and the popular methods to reach such features. Using the collected URI's and the Spotify API, we extracted the related data for each sample. Details regarding the audio features extracted from Spotify were explained in Table 3.3.

Table 3. 3 Spotify Audio Feature Set and Feature Explanations

Feature	Type	Feature Description
Acousticness	Float	Acoustic contents' confidence measure ranged from 0 to 1, considering whether the track is acoustic. Acousticness=1 expresses high confidence the record is acoustic.
Danceability	Float	Illustrates a confidence measure which represents how proper a track is for dancing depending on a musical elements' combination consisting of rhythm stability, overall regularity, tempo, and beat strength. Danceability= 0 means track is least danceable.
Duration_ms	Int	A song's length, i.e., duration (milliseconds).
Energy	Float	Includes a perceptual degree of intensity and activity, & ranges from 0 to 1. Perceptual features contributing to this trait incorporate timbre, onset rate, dynamic range, general entropy, and perceived loudness. e.g., fast, loud, and noisy stands for energetic music in general.
Instrumentalness	Float	A measure representing vocal existence in the track. The probability the track includes no vocal content increments accordingly the rise in instrumentalness score, and max value is 1.
Key	Int	Described as song's signature & uses standard pitch class notation. e.g., 0 = C, 2 = D.

Feature	Type	Feature Description
Liveness	Float	Distinguished whether a song was performed live or not by recognizing the audiences' existence. Lower liveness values depict a decreased probability that the track was not performed live. A value above 0.8 implements a sturdy possibility that the track is live.
Loudness	Float	Loudness is described as decibels (dB) & averaged over the entire track. Values commonly range within -60 and 0 dB.
Mode	Int	Displays tracks' modality (minor- major). Major is mapped by 1, and minor is 0.
Speechiness	Float	A measure of spoken words in a record. 1 is the highest value for speechiness and increases with raise in speech-like identification in a track. Rates within 0.33 and 0.66 express tracks that may hold both speech and music.
Tempo	Float	Illustrates in beats per minute (BPM) and related to the speed of a piece.
Time Signature	Int	Specify how many beats are within each bar (or measure).
Valence	Float	Describes the musical positiveness conveyed by a song, its range is 0-1. A measure from 0 to 1, which Records with low valence tone has a less positive perception.

In the previous section, we explained that a portion of the data eliminated since we cannot find their audio information on Spotify, even though the song tracks were reachable on YouTube. Besides, the songs with adapted versions were also removed from the collected data archive. At last, 13 music attributes for 1246 songs in total were archived for the audio modeling process.

3.3.2 Lyric Feature Extraction

Lyrics are vibrant sources and can produce valuable information regarding the emotions of songs. To be able to build a classification of the songs into four emotion categories by utilizing their lyrics, first of all, we extracted song lyrics from several online music databases, such as "allmusic.com", "songlyrics.com" and "musixmatch.com" with the help of Python's beautiful soup package²², which parses the websites for lyric collections. For those lyrics of the songs, that we cannot find, the Google search engine was used, and the remain lyrics were collected by manual effort.

An instance for a song lyric, before the implementation of any text preprocessing, was presented in the following figure – Figure 3.4.

²² <https://pypi.org/project/beautifulsoup4/>

```
Akar zaman, Yakar zaman  
Kirli temiz eksik tamam  
Yuvarlanıp yokuşlardan  
Duraklarda duruyorum.  
Duyduğumu bağırmadan  
Fazlasına sarılmadan  
Gece gündüz arasında  
Mekikleri dokuyorum..  
Yanmış içinden!!!  
Söylenmez, Dile gelmez  
Kuyunun dibinde  
Çalar eski bir şarkı  
Yanmış içinden!!!  
Merhemi yok, Bulunmaz.  
Gizden gölgeden  
Yürür yarası saklı  
Elimdeki kartlar bunlar  
Aynada gördüğüm kafam  
Başka kafamdaki adamdan  
Olsun, Devam ediyorum...
```

Figure 3. 4 A song lyric example – original version

In this study, we used Python, which is object-oriented and high-level programming language with dynamic semantics to text preprocessing, feature extraction, data analysis, and model building steps. First of all, the required libraries were imported, that are Pandas²³, NumPy²⁴, Collections²⁵, and Scikit-learn²⁶.

Pandas is a fundamental Python package for data science, which supports to manipulate and analyze data by allowing the creation of expressive and flexible data structures such as data frames storing the data in rectangular grids. NumPy is a primary package for scientific computing which contains a potent N-dimensional array object, and this feature was utilized to use stratified folds for accuracy testing in this research. Besides, Collections, which are Python containers, was used to store data collections, such as emotion distributions for songs. Lastly, Scikit-learn library was imported to be able to apply classification and regression algorithms.

²³ <https://pandas.pydata.org/>

²⁴ <https://www.numpy.org/>

²⁵ <https://docs.python.org/2/library/collections.html>

²⁶ <https://scikit-learn.org/stable/>

3.3.2.1 Preprocessing and Data Cleaning

The ambiguity and complexity intrinsic in human language is a significant restraint to prosperous computer understanding. Thus, dealing with such problems is one of the most critical tasks of any data related design, and so, some preprocessing tasks should be applied before moving on to feature extraction step to be able to have healthier classification outcome. We summarized some problems that we come across, and the methods we applied to deal with them.

Stop word/s Removal

Text documents ordinarily contain many function words, also known as stop words, which are not necessary to sense the general idea of the text. Since they carry limited meaning; they do not supply any significant value for modeling. In many information retrieval processes, such words are filtered out of the corpus in order to increase the relevance of the corpus and reduce the dimensionality to develop the model performance.

In this step, we created a stopword list by utilizing a list from GitHub, and thereby, 223 words in total were determined as non-valuable and eliminated from the corpus, such as "defa," "dahi," "herhangi," "pek," "şunu," "yoksa" etc. to further the lyric analysis.

Digits and Punctuation Removal

All the numerical data was also eliminated from the corpus. Besides, all punctuation such as "!" and ";" and also, all special characters were removed from the corpus by using Python's regular expression.

Tokenization

It is a method of converting a block of text into words or phrases called tokens by splitting the text according to specific characters, tabs, or spaces. For the tokenization process in this research, the lyrics under corpus were divided into words by taking advantage of spaces through using Python strip function in order to further data cleaning. For this process, Python split functionality were used. Then, all the tokens were transformed into the lower-case to deal with case-sensitivity issue.

The original version of the song lyric example after the mentioned required removal and tokenization processes were displayed in Figure 3.5.

```
'akar' 'zaman' 'yakar' 'zaman' 'kirli' 'temiz' 'eksik' 'tamam'  
'yuvarlanıp' 'yokuşlardan' 'duraklarda' 'duruyorum' 'duyduğumu'  
'bağırmadan' 'fazlasına' 'sarılmadan' 'gece' 'gündüz' 'arasında'  
'mekikleri' 'dokuyorum' 'yanmış' 'içinden' 'söylenmez' 'dile'  
'gelmez' 'kuyunun' 'dibinde' 'çalar' 'eski' 'şarkı' 'yanmış'  
'içinden' 'merhemi' 'yok' 'bulunmaz' 'gizden' 'gölgeden' 'yürür'  
'yarası' 'saklı' 'elimdeki' 'kartlar' 'aynada' 'gördüğüm' 'kafam'  
'başka' 'kafamdaki' 'adamdan' 'devam' 'ediyorum'
```

Figure 3. 5 The lyric example after preprocessing without stemmed

Stemming

Stemming is a process, which was used to group words with the same morphological-base into one class, namely reducing the words to their root (stem) version. For instance, "seni," "sana," and "senden" words were reduced to their root, which is "sen." For this process, the Spacy²⁷ stemmer was employed. When considering the previous researches, since the stemming process exhibited a mixed outcome in text-classification, we investigated both versions for choosing the set of words to incorporate the BOW set.

The following figure, Figure 3.6, displays the stemmed version of the song lyric after stopword removal and text preprocessing steps.

```
akar zaman yak zaman kir temiz eksik  
yuvarla yokuş durak dur duy bağır fazla  
sar gece gündüz ara mekik doku yan iç  
söyle dil gel kuyu dip çalar eski şarkı  
yan iç merhem yok bulun giz gölge yürü  
yara saklı el kart ayna gör kafa  
başka kafa adam devam et
```

Figure 3. 6 The stemmed lyric example

²⁷ <https://spacy.io/api/lemmatizer>

As can be observed, the tokens were replaced with their root-bases. For example, “yakar” was replaced to “yak,” “kirli” was replaced to “kir,” “elimdeki” was replaced to “el,” etc.

3.3.2.2 Textual Feature Extraction Process

In this study, the textual features derived from the song lyrics were extracted both using a frequency-based analysis, as well as a similarity-based approach.

For the frequency-based strategy, bag-of-words (BOW) features with TF-IDF weighting were extracted from the lyrics after completing stopwords removal by using both the original forms (non-stemmed) and stemmed version of lyrics. Since we attempted to develop a model by using a collection of lyrics and corresponding user tags, we preferred to utilize TF-IDF metric in order to represent the relative importance of specific words for a particular emotion category. Thereby, we aimed to estimate which emotion state is most relevant regarding the given lyrics, where the emotion is represented by the combined lyrics of all songs, namely the corpora, which have that particular emotion assigned.

As explained before, by employing this approach we not only considered the number of times specific word (w) appears in a particular song lyric (s), which reflects Term-frequency, $TF(w,s)$, but also, in how many documents, i.e., songs, the word appeared in were determined through inverse document frequency, $IDF(w,C)$, where C stands for the corpus size, in other words, number of the songs in total. Thereby, if the frequency of a word increases within the same song, the word importance is also improved, but the word importance is decreased if it occurs in other songs in the corpus. Consequently, high TF-IDF values symbolize the high relevance of the word for the respective emotion class.

By this approach, a feature vector was created for each document, i.e., song, in a V -dimensional vector space, where each vector corresponds to a point, and the vector dimension correlates with the number of words.

Term by document matrices, which are two-dimensional matrices, whose rows stand for the terms and columns represents the documents for each entry, (w, s) index, which is represented by a TF-IDF weight, was created through employing both original and stemmed

version of the tokens, and then lyrics were fed into supervised learning algorithms to generate corresponding emotion detection models.

For this process, we used Python TF-IDFVectorizer²⁸ and also, we utilized "ngram_range" parameter offered by Scikit-learn, which allows using the combination of n words to tune the model input further by assigning the lower and upper boundary of the range of n-values. So, we were able to use all values of n, in which $1 \leq n \leq 3$, through taking into account the combinations of unigram, bigram and trigram tokens, instead of just using singular words (unigrams). Thereby, we attempted to capture more of the word semantics, which may lead to boost performance by accessing higher-order BOW feature combinations

Although the TF-IDF approach is resourceful for extracting the lexical text features, it does not have the capability for capturing the semantics of words. Therefore, we also used Word2Vec, which is a word embedding model obtained from the hidden layer of a two-layered neural network, as the second approach in our research to be able to create textual features considering syntactic and semantic similarities. Word2Vec gets a large corpus of text as its input and generates a multi-dimensional vector space considering each unique word in the corpus, that was appointed a corresponding vector in the space. So, it generates a unique dense vector for each word, while investigating the appearance of other words around the particular word, which was discussed detailly in the literature review section. For this process, we adopted Python's Gensim library²⁹, which was designed to extract semantic topics from the documents automatically.

3.4 Predictive Model Building and Testing

For the classification model building and testing step, we operated Scikit-learn Python library. Four different experiments were designed through using different musical features, which are audio features extracted from Spotify, and textual inputs, such as TF-IDF features and Word2Vec features.

Moreover, we attempted to create a multimodal approach by combining audio features and the winner textual features. In order to receive better classification achievements, different

²⁸ https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

²⁹ <https://pypi.org/project/gensim/>

classification methods were utilized, such as support vector machines (SVM) with a linear kernel, which is the libsvm based implementation also called SVC method and Linear SVC method. While Linear SVC uses liblinear estimators, that is optimized for a linear hyperplane, SVC supports different kernels and does not penalize the intercept used for separation of different classes. Furthermore, in this research, we employed Multinomial Naïve Bayes, Random Forest classifier, Decision Tree classifier, and also Logistic Regression method in order to find the best-performed machine learning algorithms for emotion classification throughout the different experiments.

Besides the usage of the supervised learning approaches, also a semi-supervised machine learning approach was applied for the efficient usage of big unlabeled data without the hand-labeling effort, which over consumes time and human power.

For each category, k-Fold cross-validation was adopted with various k values as 3,6, and 10 in order to receive the most reliable accuracy performances of these models, and to avoid overfitting. To analyze the utility of the various feature selection methods, we used the accuracy score, F1-score with four variants, precision score, and recall score as the performance measures, whose formulations given below.

$$\frac{\mathbf{TP + TN}}{\mathbf{TP + FP + TN + FN}}$$

Equation 3. 1: Accuracy Score

$$\frac{\mathbf{TP}}{\mathbf{TP + FP}}$$

Equation 3. 2: Precision Score

$$\frac{TP}{TP + FN}$$

Equation 3. 3: Recall (Sensitivity)

$$\frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

Equation 3. 4: F1 Score

3.4.1 Research Data

After all textual preprocessing phases, firstly, all lyric sets which have 1500 songs in total, was read by using Python. The final data consists of several columns, which are Song_ID, Artist and Song Name; Song Lyrics, which were preprocessed and depicted as list format; 13 Spotify Audio Features, and the probabilistic emotional tags, which were created through annotation process and organized accordingly. A portion from the research dataset was displayed in the following figure, Figure 3.7.

As next, the emotional categories were transformed into numeric values, which has the range from "0" to "3", i.e., Joyful=0, Relaxed=1, Sad=2, and Angry=3. Besides, "-1" was used for the songs found as emotionally confusing for removing the noisy data from the corpus. Thereby, the final corpus was prepared to be ready to further analysis.

Song_Lyrics	Danceability	Duration_ms	Energy	Instrumentalness	Key	Liveness	Loudness	Mode	Speechiness	Tempo	Time_signature	Valence	Acousticness	JOYFUL	RELAXED	SAD	ANGRY
['tek', 'güvencem', 'aşk', 'tek', 'tehlikem	0.561	196747	0.384	0.00276	10	0.128	-7877	0	0.027	100.05	4	0.177	0.825	0.50	0.25	0.25	0.00
['aptal', 'zannedip', 'teselli', 'ol', 'ya', '	0.426	295333	0.707	0	0	0.168	-6364	1	0.0351	144014	4	0.6	0.0803	0.80	0.00	0.20	0.00
['derin', 'uçurumlarda', 'açan', 'diken	0.528	209573	0.476	0	1	0.128	-8919	1	0.0308	104176	4	0.384	0.459	0.00	0.00	1.00	0.00
['geç', 'kalan', 'sendin', 'isteseydin', 'g	0.515	272440	0.516	0.00000288	10	0.163	-9256	0	0.0266	98025	4	0.541	0.457	0.00	0.00	0.80	0.20
['neyleyim', '3', 'günlük', 'yalan', 'düny	0.407	173120	0.116	0	0	0.187	-11262	1	0.0451	70716	3	0.312	0.881	0.00	0.00	1.00	0.00
['çantalar', 'elimde', 'kuşatıldım', 'surl	0.499	214280	0.83	0.0000566	8	0.0637	-5561	0	0.0738	110775	4	0.729	0.11	0.75	0.00	0.00	0.25
['üşüyordu', 'rüyaları', 'gerçekliği', 'ça	0.523	277387	0.729	0.000136	9	0.0944	-7195	0	0.0307	107952	4	0.334	0.256	0.00	0.25	0.25	0.50
['yordu', 'tüm', 'olanların', 'sorgusu', '	0.555	270493	0.855	0.000222	8	0.141	-6413	0	0.0471	109201	4	0.696	0.0182	0.00	1.00	0.00	0.00

Figure 3. 7 The song data-set part

3.4.2 Research Experiments

Experiment-I: Audio Feature Analysis

Thirteen different audio-related features, including tempo, danceability, energy, and acousticness were used as input values to classify 1246 labeled song data into four different emotion categories.

Figure 3.8 depicts a song example reflecting the audio features and feature values utilized as model inputs.

Audio Feature	Value
Danceability	0.481
Duration_ms	76387
Energy	0.271
Instrumentalness	0.732
Key	3
Liveness	0.122
Loudness	-17782
Mode	1
Speechiness	0.0407
Tempo	121873
Time_signature	4
Valence	0.0867
Acousticness	0.851

Figure 3. 8: A song example: Audio features-emotion tag matching

The various supervised algorithms were used to evaluate the performance of the audio features, such as SVC method, Linear SVC method, Random Forest classifier, Decision Tree classifier, and Logistic Regression. The performance outcome of this

experiment when considering CVsize=10 without stem, can be found in the following table, Table 3.4.

Experiment-II: Lyric Analysis using TF-IDF

In this design, 1246 labeled songs were used to evaluate the representative power of text using TF-IDF features with n ngram_range= (1,3), which allows us to combine unigram, bigram and trigram features altogether. Moreover, six different classifiers were trained and then tested on cross-validated data with k=3,6,10 values.

Figure 3.9 displays an example for a song lyric and the relative assigned tag considering for both original and stemmed versions of the words.

```
('akar zaman yakar zaman kirli temiz eksik yuvarlanıp yokuşlardan duraklarda duruyorum duyduğumu  
bağırmeden fazlasına sarılmadan gece gündüz arasında mekikleri dokuyorum yanmış içinden söylenmez  
dile gelmez kuyunun dibinde çalar eski şarkı yanmış içinden merhemi yok bulunmaz gizden gölgeden  
yürür yarası saklı elimdeki kartlar aynada gördüğüm kafam başka kafamdaki adamdan devam ediyorum', 1)
```

```
('akar zaman yak zaman kir temiz eksik yuvarla yokuş durak dur duy bağır fazla sar  
gece gündüz ara mekik doku yan iç söyle dil gel kuyu dip çalar eski şarkı yan iç  
merhem yok bulun giz gölge yürü yara saklı el kart ayna gör kafa başka kafa adam devam et', 1)
```

Figure 3. 9: A song example from lyric-emotion matching

In this design, after stopword/s removal and data preparation steps, both original and stemmed versions of the song lyrics were used as model inputs. However, there was no significant difference between the performance results when considering both datasets. Namely, the stemmed data did not show any particular improvement on the classification performance. The experiment's performance evaluation, considering the stem version with CVsize=10, can be seen in Table 3.5.

Table 3. 4: Music Audio Feature Analysis Performance Results

Algorithm	Accuracy Score	F1_Macro Score	F1_Micro Score	F1 Score	F1_Weighted Score	Precision Score	Recall Score
LogisticRegression	0.44	0.15	0.44	0.15	0.27	0.11	0.25
SVC	0.44	0.16	0.44	0.16	0.27	0.15	0.25
DecisionTreeClassifier	0.34	0.26	0.33	0.26	0.33	0.26	0.26
RandomForestClassifier	0.35	0.23	0.35	0.23	0.32	0.23	0.24
LinearSVC	0.33	0.12	0.33	0.12	0.17	0.08	0.25

Table 3. 5: Music Lyric Feature (TF-IDF) Analysis Performance Results

Algorithm	Accuracy Score	F1_Macro Score	F1_Micro Score	F1 Score	F1_Weighted Score	Precision Score	Recall Score
LogisticRegression	0.46	0.19	0.46	0.19	0.32	0.25	0.27
SVC	0.44	0.15	0.44	0.15	0.27	0.11	0.25
DecisionTreeClassifier	0.36	0.27	0.36	0.27	0.36	0.26	0.27
LinearSVC	0.46	0.28	0.46	0.28	0.41	0.29	0.31
MultinomialNB	0.44	0.15	0.44	0.15	0.27	0.11	0.25
RandomForestClassifier	0.42	0.25	0.42	0.25	0.36	0.26	0.27

Experiment-III: Lyric Analysis using Word2Vec

In this experiment, we attempted to design a semi-supervised approach by using the labeled song data set, and a vast amount of unlabeled data, which consists of more than 2.5 million documents in total, that were gathered from diverse public sources, including Turkish Wikipedia.

For this purpose, firstly, Word2Vec, unsupervised state-of-the-art model in the word embedding studies since it considers the similarity between words along with the probability function for word sequences, was utilized to extract textual-based features from the song data set regarding the semantic meanings of the words. After that, four different supervised learning algorithms, which are SVC, Linear SVC, Random Forest classifier, and Logistic Regression, were trained on the labeled data. Furthermore, the winner algorithm was employed to label the unlabeled data, and lastly, the classification performance was tested on the non-stemmed cross-validated data with $k=3,6,10$ sizes. The performance of each algorithm for $CVsize=10$ was exhibited in Table 3.6.

Experiment-IV: Multimodal Approach using Word2Vec and Audio Features

In this experiment, we aimed to create a multimodal design via through employing a semi-supervised learning approach. The labeled song data was used to train a supervised model used for labeling the big unlabeled data. Unlike the Experiment-III, in this research design, the combination of audio and textual features was used as a combined input set in order to develop classification models by utilizing four different supervised machine learning algorithms, such as SVC, Linear SVC, Random Forest classifier, and Logistic Regression. Each model was tested on the cross-validated big data with $k=3,6,10$ values, respectively.

Figure 3.10 reflects an example for a song lyric, the assigned emotional tag, and the musical features, respectively.

```
('zamanı geldi ağla gözlerim ağla giden geri gelmedi ağla gözlerim ağla gönül  
hasta günden güne solsa ağlamak zor ağla gözlerim ağla derdim sığmaz dağlara  
döndüm an bağlara vur başını taşlara ağlara gözlerim ağla aşk çember sardı  
benliğimi kabul et yenilgimi ağla gözlerim ağla', 2,  
array([ 4.56000e-01, 2.37587e+05, 2.50000e-01, 4.71000e-06,  
        2.00000e+00, 9.87000e-02, -1.23950e+01, 0.00000e+00,  
        4.04000e-02, 1.65083e+02, 4.00000e+00, 1.39000e-01,  
        9.16000e-01]))
```

Figure 3. 10: A song example with emotional tag, lyrics, and audio feature space

The performances of each algorithm regarding CVsize= 10, were displayed in Table-3.7.



Table 3. 6: Performance Results for Semi-Supervised Analysis using Word2Vec features

Algorithm	Accuracy Score	F1_Macro Score	F1_Micro Score	F1 Score	F1_Weighted Score	Precision Score	Recall Score
SVC	0.44	0.15	0.44	0.15	0.27	0.11	0.25
LogisticRegression	0.51	0.32	0.51	0.32	0.46	0.34	0.35
RandomForestClassifier	0.42	0.29	0.42	0.29	0.39	0.29	0.30
LinearSVC	0.50	0.34	0.50	0.34	0.46	0.36	0.35

Table 3. 7: Performance Results for Semi-Supervised Multi-Modal Analysis

Algorithm	Accuracy Score	F1_Macro Score	F1_Micro Score	F1 Score	F1_Weighted Score	Precision Score	Recall Score
RandomForestClassifier	0.42	0.29	0.42	0.29	0.39	0.29	0.30
LinearSVC	0.50	0.34	0.50	0.34	0.46	0.36	0.35
SVC	0.44	0.15	0.44	0.15	0.27	0.11	0.25
LogisticRegression	0.51	0.32	0.51	0.32	0.46	0.34	0.35

3.5 Evaluation

In Experiment-I, we designed a classification approach by utilizing the audio features solely as the model input to predict the perceived emotions derived from the music. In this design, the SVC model created the best classification performance with 44.2% accuracy rate, which was followed closely by Logistic Regression model with 44.06%. Besides, Decision Tree Algorithms showed the best performance concerning most of the F1 metrics, precision, and recall scores. In this design, Linear SVC was the worst performed model with 33.1% accuracy score.

The outcomes derived by Experiment-II and Experiment-III revealed that Logistic Regression and Linear SVC algorithms offered the best accuracy performances, when we only utilized the textual information regardless of the employed extraction method.

In Experiment-II, Linear SVC generated the highest performance score with 46.3% accuracy rate. Besides this algorithm outperformed the other models considering each performance metric. Interestingly, while Logistic Regression performed the second-best results for accuracy and F1-micro scores, Decision Tree displayed a better performance than Logistic Regression regarding precision and recall scores, as well as other F1 metrics.

On the other hand, in Experiment-III Logistic Regression created the best accurate model with 51.3% performance score, which also displayed the best outcomes considering the recall, F1-micro, and F1-weighted scores. Besides, Linear SVC algorithm performed the best results using the textual features regarding precision, F1-macro, and F1 scores. Additionally, the worst accuracy outcomes were generated by Decision Tree models, which was fed by TF-IDF features, with 36.2%, and Random Forest algorithm displayed the lowest accuracy score with 43.7% performance when fed by Word2Vec features.

Moreover, when considering all of the first three experiments, it can be said that the textual features outperformed the audio features without exception regarding emotion recognition from music. The outcomes of the experiments showed that textual features supplied more valuable inputs for the models, rather than musical attribution. Besides, utilizing a semi-

supervised approach in MER domain can improve the performance for all metrics when predicting the emotion from the given contents.

In Experiment IV, even though we attempted to generate a multimodal classifier by combining the audio and Word2Vec textual features for music emotion recognition; the study outcomes did not show any noteworthy differences when compared to Experiment III. This result depicted that the audio features did not bring any remarkable improvement into the classification design. In this approach, Logistic Regression performed the best accuracy in terms of emotional classification with 51% accuracy rate.

In consideration of all experimental research designs, the best performance results generated by SVC, Linear SVC, and Logistic Regression algorithms when the models fed by audio, TF-IDF, and Word2Vec features, respectively. Overall, while SVC and Logistic Regression models showed the most stable accuracy rates regardless of the given input, i.e., musical or textual features, noticeable performance variations were observed when employing Random Forest and Linear SVC algorithms. Linear SVC classifiers created the lowest accuracy scores when we used the audio inputs whereas; both of the models illustrated significantly improved performance when textual features were utilized.

CHAPTER 4

DISCUSSION & CONCLUSION

4.1 Research Framework Overview & Managerial Implications

We have proposed an emotion prediction system by adopting Russell's emotion classification model. One of our goals in this study was that making an in-depth examination about the significance of the various features derived from different resources on the performance of the MER system by evaluating their affective attribution on the songs. Besides, we attempted to find the best possible classification models not just for the audio and lyric dimensions individually, but also a combination of both. In that respect, a multi-modal approach was designed in a context of emotion recognition through combining information from distinct sources, which are audio, lyrics, and big textual corpus.

To accomplish the intent, we proposed a new ground truth dataset containing 1500 songs, which was manually annotated into the four emotion clusters defined in Russell's model. The annotation process was accomplished by four people from diverse demographics and also, different educational and socio-economic backgrounds. Besides, a comprehensive unlabeled dataset was collected to improve the performance of classification models through a semi-supervised approach.

Further, we have extracted and organized a comprehensive feature-set consisting of both musical and textual inputs. In the information retrieval design, a novel set of musical features were extracted from Spotify, as well as state of the art textual features, such as Word2Vec vectors and bag-of-words (BOW) features weighted by TF-IDF, which enables usage flexibility for unigrams, bigrams, and trigrams, undergone or not to a set of textual transformations, e.g., stemming and stop-word removal.

In the first two experiments, we constructed the best possible classifiers both for audio and lyrics attributes separately. As first, thirteen musical features are used as the model input, and next, TF-IDF scores for the words were utilized. Furthermore, a word-embedding approach with Word2Vec method was followed in the third experiment in order to reach the best possible textual features when considering the perception of emotion. In this study, we followed a semi-supervised approach by utilizing both labeled song data and big unlabeled text data which was used to generate word vectors. Finally, in the fourth experiment, we constructed a combined set of features consisting of the extracted audio features and the best performed lyrical attributes to train and test several classification models.

We employed six different algorithms throughout the experiments since the previous evaluations in the literature mostly depicted that these techniques exhibit the best performances. Furthermore, the classification results were cross-validated, and the performance outcomes created using seven metrics were reported to compare and evaluate the four approaches.

The obtained outcomes demonstrated that the proposed semi-supervised approach can be more resourceful when compared to emotion classification approaches, which depend on the usage of audio features solely. As a result of this study, we also showed that several classification models can be implemented accurately by using both musical and textual features; yet, the inclusion of the textual features improves the performance of the overall models.

The research outputs and insights can be utilized for business practices. Emotion classification can be applied to various situations from daily life activities of humans to marketing strategies of brands. With the contribution of digitalization, the emotional impact of

music can provide more insights into the physiological, psychological, and behavioral understanding of people and their reciprocal relationships with the environment and business.

The current music recommendation approaches are generally based on simple preferences and previous selections of users, yet people's patterns for music choice are usually related to the emotional status at the moment of that choice. Since most people continue to listen to music for its affective perceptiveness, individuals seek out more creative and personalized paths while encountering music, regardless of time and context. With consideration of all, a prominent part of the music industry has been started to invest in new recommendation systems by utilizing the reciprocal relationship between human dynamics, emotion, and music.

As artificial intelligence (AI) and deep learning continue improving, utilization of such affective computing approaches for emotion classification may reshape the music industry and services fundamentally by underpinning personalized musical recommendations. For example, the right music can be recommended for the moment considering dynamic personal attributes, which depends on physiological and psychological states of human, as well as situational determinants, more intuitively and consistently. Moreover, recommending and playing the music that matches the users' moods detected from his or her physiological signals, such as skin conductance, blood pressure, and facial cues can be benefited for real-time music selection and recommendation. Furthermore, the recent neuroimaging studies have shown that appraisal of a new musical piece has a neural relationship with precise parts of the human brain, which triggers motivation, pleasure, and reward mechanisms, and the brain's activation areas processing sound features, which associated with emotions and decision making. This connection between our emotional brain affected by music and decision making can be used to predict how much money a person would be willing to pay on an original music piece (Salimpoor, 2013; Koelsch, 2014). Music certainly has the power to stimulate strong emotions within us, and the personal attributes deriving from the neural correlations evoked by its emotional expressionism can be utilized and managed to reshape the marketing strategies of business brands through influencing the decision making processes of people regarding products even other than music. For instance, in a purchasing behavior study, it has been observed that playing French music as background sounds in a wine shop boosted wine sales for the wines produced from France (North and Hargreaves, 1997). Besides, it has been proved that the purchasing amount of people can be manipulated by modifying the genre or mood of

the background music. For example, Areni and Kim (1993) confirmed that hearing classical music makes customers tend to purchase more and direct their attention to more expensive products by effecting their price perception.

Additionally, music emotion recognition (MER) function may be utilized in a portable device, such as a phone application, and thereby the best music matching to the emotional state of the user or the environmental conditions where the person within can be recommended, and personal music collection can be created by more agile and efficient ways. In overall, advances in automatic music emotion recognition can be enhanced through increasing and utilizing human interaction with emotionally sensitive and sociable machines, which results in better music experiencing journey.

Each of these situations depends profoundly on the emotional content of the music and its impact on people's behavior. In consideration of all, it can be said that music takes place in a part of everyday life and it has vital power in influencing our emotions and so our cognitions and decisions. Consequently, when it comes to human-centric business perspective, it is possible to utilize music emotion recognition and recommendation approaches to improve instore music design of stores or places, which will make the purchasing experience of customers better and also increase sales volumes. Besides, the people-brand relationship can be developed by agile, real-time, and personalized advertising resulted from the utilization of emotional content and context of music.

4.2 Limitations & Future Works

Music is a complex caption to analyze since it consists of a multitude of independent and dependent parameters. In this research, a framework, which was built by the models of human emotions, was generated to classify music, while using its audio and lyrical contents. Russell's dimensional model of emotion was utilized regarding its congruence with music psychology. Although the results showed promise for the used framework, the highest classification rate of 51.3% was not eminently high. The level of the obtained accuracy rates can be attributed to certain limitations of the study.

One of the significant limitations in this work was the limited amount of labeled ground-truth dataset. Manually annotating music with emotion labels is an expensive and time-

consuming task, and naturally a highly subjective process. Having limited human and data sources, and time for the annotation phase was very restrictive since it affects the rest of the process dramatically. This framework could be re-generated into the improved version of its current version in several ways. Utilization of web-integrated social tagging media platform in the Turkish language, which enables much more user participation, can help to generate a more integrated and cumulative labeling process. Furthermore, listeners' comments can be extracted from YouTube and utilized to annotate the songs with the help of text mining. This approach eliminates the participant restrictions and helps to manage research's time more effectively.

Besides, the dependence of the model performances on the chosen songs and also the annotators can be determined through utilizing metaknowledge, such as song titles and demographic information like the gender of the singers and annotators. This approach may give a valuable perspective to experiments and improve classification accuracies by bringing a further standard on perceived emotion from the music since it is capable of discovering various attributes generated from songs, singers and also listeners, which helps to create more specialized and customized music collections for users.

Additionally, we have experienced that the multi-modal design did not bring any significant contribution to our classification performance. For the future researches, various and better input combination techniques, as well as different machine learning algorithms can be utilized to enhance the classification performance. Additionally, an extra annotation section can be designed in order to achieve a deeper perception upon lyric-based emotion classification by presenting the songs' lyrics solely for emotional labeling.

In this research, we have proposed novel classification systems through association discovery across various contextual and conceptual music attributes and also utilizing several predictive model building approaches. We intend to have a more extensive understanding of the role of emotion and perception evoked by music. Besides, we have achieved to generate automatic emotion-based recognition and classification systems utilizing musical perception, information processing, and machine learning algorithms. As a result of various experiments conducted in this research, we have proven that music has an undeniable connection with emotion, and diverse musical attributes, and also human-centric perspective can be practiced analyzing and organize music across emotions.

BIBLIOGRAPHY

- Areni, C. S., & Kim, D. (1993). The influence of background music on shopping behavior: classical versus top-forty music in a wine store. *Advances in consumer research*, 20(1), 336-340.
- Barthet, M., Fazekas, G., & Sandler, M. (2012). Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. *In 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) Music and Emotions*, pages 19-22, June 2012.
- Becker, J. (2001). Anthropological perspectives on music and emotion. *In: P.N. Juslin, & J.A. Sloboda (Eds.), Music and Emotion: Theory and Research*. New York: Oxford University Press, 135–160.
- Bengio, Y., Ducharme, R., & Vincent, P. (2001). A neural probabilistic language model. *NIPS*.
- Besson, M., Chobert, J., & Marie, C. (2011). Transfer of Training between Music and Speech: Common Processing, Attention, and Memory. *Frontiers in psychology*. 2. 94. 10.3389/fpsyg.2011.00094.
- Beveridge, S., Knox, D., & Macdonald, R. (2012). Emotion Recognition in Western Popular Music: The Role of Melodic Structure.
- Bigand, E. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts, *Cognition, and Emotion*, vol. 19, no. 8, p.1113.
- Binali, H., Chen, W., & Vidyasagar, P. (2010). Computational approaches for emotion detection in text. 172 - 177. 10.1109/DEST.2010.5610650.

- Bischoff, K., Firan, C.S., Paiu, R., Nejdil, W., Laurier, C., & Sordo, M. (2009). Music mood and theme classification-a hybrid approach. *In Proc. of the Intl. Society for Music Information Retrieval Conf.*, Kobe, Japan.
- Bradley, M.M., & 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, University of Florida.
- Brownlee, J. (2017, October 11). *What Are Word Embeddings for Text?* Retrieved from <https://machinelearningmastery.com/what-are-word-embeddings/>
- Capurso, A., Fisichelli, V. R., Gilman, L., Gutheil, E. A., Wright, J. T., & Paperte, F. (1952). *Music and Your Emotions*. Liveright Publishing Corporation.
- Casey, M.A., Veltkamp R., Goto M., Leman M., Rhodes C., & Slaney M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4), 668-696.
- Castro, S. L. & Lima, C. F. (2014). Age and musical expertise in uence emotion recognition in music. *Music Perception: An Interdisciplinary Journal*, 32(2):125-142.
- Chapelle, O., Scholkopf, B., & Zien, A. (2006). *Semi-Supervised Learning*. MIT Press.
- Chen, P.I., Liu JY., & Yang YH. (2015). *Personal Factors in Music Preference and Similarity: User Study on the Role of Personality Traits*.
- Chi, C-Y., Wu, Y-S., Chu W-R., & Wu Jr, D. (2011). The Power of Words: Enhancing Music Mood Estimation with Textual Input of Lyric.
- Cohen Mostafavi, A., R, Z., & Wiczorkowska, A. (2014). From Personalized to Hierarchically Structured Classifiers for Retrieving Music by Mood. 8399. 231-245. 10.1007/978-3-319-08407-7_15.
- Cohn, J., & Katz, G. (1998). Bimodal expression of emotion by face and voice, *in ACM Intl. Multimedia Conf.*

- Cho, H., Kim, S., Lee, J., & Lee, J-S. (2014). Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews.
- Collobert, R. & Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. NEC Labs America.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research* 12 (Aug), 2493-2537.
- Downie, J. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, Vol. 29, No. 4, pp. 247–255.
- Duggal, R., Chakraverty, S., & Narang, J. (2014). Extracting Emotions in Songs Using Acoustic And Lyrical Features.
- Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models, in Proc. of the *Intl. Society for Music Information Conf.*, Kobe, Japan.
- Eerola, T., & Vuoskoski, J. K. (2013). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, 30(3):307-340.
- Egermann, H., Fernando, N., Chuen, L., & McAdams, S. (2015). Music induces universal emotion-related psychophysiological responses: Comparing Canadian listeners to Congolese Pygmies. *Frontiers in Psychology*, pp.1-9.
- Ekman, P. (1992). An argument for basic emotions, *Cognition, and Emotion*, vol. 6, pp. 169–200.
- Ekman, P. (2003). Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional Life. Times Books.

- Fehr, B., & Russel, J. (1984). Concept of Emotion viewed from a prototype perspective. *Journal of Experimental Psychology*, Washington, pp. 464-486.
- Feng, Y., Zhuang, Y., & Pan, Y. (2003). Popular Music Retrieval by Detecting Mood. *Proc. 26th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, vol. 2, no. 2, pp. 375–376.
- Friberg, A. (2008). Digital Audio Emotions: An Overview of Computer Analysis and Synthesis of Emotional Expression in Music, *DAFx*, pp.1-6.
- Fried, R., & Berkowitz, L. (1979). Music that charms. And can influence helpfulness. *Journal of Applied Social Psychology*,9, 199–208.
- Gabrielsson, A., & Juslin, P. (1996). Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music*, Vol. 24.
- Gabrielsson, A., & Lindström, E. (2001). The influence of musical structure on emotional expression, *Music and Emotion: Theory and Research*, pp.223–248.
- Gabrielsson, A. (2002). Emotion Perceived and Emotion Felt: Same or Different?. *Musicae Scientiae (special issue)*, *European Society for the Cognitive Sciences of Music (ESCOM)*.
- Guo, J., Che, W., Wang, H., & Liu, T. (2014). Revisiting Embedding Features for Simple Semi-supervised Learning. Research Center for Social Computing and Information Retrieval Harbin Institute of Technology, China ‡Baidu Inc., Beijing, China.
- Hevner, K., (1936). Experimental Studies of the Elements of Expression in Music. *American Journal of Psychology*, 48(2), pp. 246–268.
- Hu, X., & Downie, J. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. *In: 8th International Conference on Music Information Retrieval (ISMIR'07)*.
- Hu, X., Downie, J., & Ehmann, A. (2009). Lyric Text Mining in Music Mood Classification. *Proc. Int. Soc. Music Information Retrieval Conf.*411-416.

- Hu, X. (2010). Music and Mood: Where Theory and Reality Meet.
- Hu, X., & Downie, J.S. (2010a). When lyrics outperform audio for music mood classification: a feature analysis. *In Proceedings of ISMIR*, pages 1–6.
- Hu, X., & Downie, J.S. (2010b). Improving mood classification in music digital libraries by combining lyrics and audio. *In Proceedings of the 10th annual joint conference on Digital libraries*, pages 159–168. ACM.
- Hu, X., & Lee, J. H. (2012). A cross-cultural study of music mood perception between American and Chinese listeners. *In Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 535-540, Porto, Portugal.
- Hunter, P., Schellenberg, E.G., & Schimmack, U. (2010). Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 4(1):47-56.
- Huron, D. (2000). Perceptual and Cognitive Applications in Music Information Retrieval. *International Symposium on Music Information Retrieval*.
- Jargreaves, D. & North, A. (1997). *The Social Psychology of Music*. Oxford University Press, Oxford, UK.
- Juslin, P.N. (1997). Emotional communication in music performance: A functionalist perspective and some data. *Music Perception*, 14, 383–418.
- Juslin, P. N., & Sloboda, J. A. (2001). Psychological perspectives on music and emotion. *Music and Emotion: Theory and Research*. New York: Oxford University Press.
- Juslin, P.N., & Lindström, E. (2003). Musical expression of emotions: Modeling composed and performed features.
- Juslin, P. N., & Laukka, P. (2004). Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening, *Journal of New Music Research*, 33:3, 217-238, DOI: 10.1080/0929821042000317813.

- Juslin, P. N., Karlsson, J., Lindström E., Friberg, A., & Schoonderwaldt, E. (2006). Play it again with feeling: computer feedback in musical communication of emotions. *Journal of Experimental Psychology. Applied*, 12(1): 79-95.
- Juslin, P. N., & Sloboda, J. A. (2010). *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, New York, USA.
- Kallinen, K., & Ravaja, N. (2006). Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, 10(2):191-213.
- Katayose, H., Imai, M., & Inokuchi, S. (1988). Sentiment extraction in music, 9th International Conference on Pattern Recognition, Rome, Italy, pp. 1083-1087 vol.2. DOI: 10.1109/ICPR.1988.28447.
- Kim, Y., Schmidt, E., & Emelle, L. (2008). Moodswings: A collaborative game for music mood label collection, in Proc. Intl. Conf. on Music Information Retrieval, Philadelphia, PA.
- Kim, M., & Kwon, H.C. (2011). Lyrics-Based Emotion Classification Using Feature Selection by Partial Syntactic Analysis. 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence.
- Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., & Turnbull, D. (2010). Music emotion recognition. A state of the art review. In: 11th International Society of Music Information Retrieval (ISMIR), pp. 255 -266.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
- Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nature Reviews Neuroscience*, 15(3), 170-180.
- Korhonen, M.D., Clausi, D.A., & Jerniga, M.E. (2006). Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) Volume: 36, Issue: 3*.

- Kosta, K., Song, Y., Fazekas, G., & Sandler, M.B. (2013). A study of cultural dependence of perceived mood in Greek music. In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR), pp 317-322, Curitiba, Brazil.
- Lamere, P., & Celma, O. (2007). Music recommendation tutorial notes, ISMIR Tutorial.
- Lamere, P. (2008). Social tagging and music information retrieval. *J. New Music Res.* 37, 2, 101-114.
- Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688-693. IEEE.
- Laurier C., & Herrera, P. (2009). Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines. *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*. chap. 2, (pp. 9-32). IGI Global.
- Lehtiniemi A., & Ojala J. (2013). Evaluating MoodPic-A concept for collaborative mood music playlist creation. 17th International Conference on Information Visualisation (IV), London, UK, 15-18.
- Leman, M., Vermeulen, V., Voogdt, L., Moelants, D., & Lesare, M. (2005). Prediction of musical affect using a combination of acoustic structural cues. *Journal of New Music Research*, 34(1):3967.
- Levy, M., & Sandler, M. B. (2009). Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia*, 11(3):383-395.
- Li, T., & Ogihara, M. (2004). Music artist style identification by semi-supervised learning from both lyrics and content. *ACM Multimedia 2004 - proceedings of the 12th ACM International Conference on Multimedia*. 364-367. 10.1145/1027527.1027612.
- Logan, B. & Salomon, A. (2001). *A Music Similarity Function Based On Signal Analysis*. Compaq Computer Corporation Cambridge Research Laboratory One Cambridge Center Cambridge MA 02142 USA.

- Logan, B., Kositsky, A., & Moreno, P. (2004). Semantic analysis of song lyrics. 827 - 830
Vol.2. 10.1109/ICME.2004.1394328.
- Lu, L., Liu, D., & Zhang, H.J. (2006). Automatic Mood Detection and Tracking of Music
Audio Signals, *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no.1,
pp. 5-18.
- Lu, Q., Chen, X., Yang, D., & Wang, J. (2010). Boosting for multi-modal music emotion
classification. In *Proceedings of the 11th International Society for Music Information
Retrieval Conference (ISMIR)*, pages 105-110, Utrecht, Netherlands.
- Malherio, R.M.S. (2016). *Emotion-based Analysis and Classification of Music Lyrics*,
University of Coimbra, Doctor of Philosophy in Information Science and Technology.
- Mandel, M., Poliner, G., & Ellis, D. (2006). Support vector machine active learning for music
retrieval. *Multimedia Systems*, 12 (1): 3-13.
- Martinazo, B. (2010). *Um Método de Identificação de Emoções em Textos Curtos para
Português do Brasil*. MSc Thesis. Pontifícia Universidade Católica do Paraná.
- Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and
measuring individual differences in temperament. *Current Psychology: Developmental,
Learning, Personality, Social*, 14, 261-292.
- Menga, J., Hongfei Lin, H. & Yu, Y. (2011). A two-stage feature selection method for text
categorization. *Computers and Mathematics with Applications* 62 (2011) 2793–2800.
- Meyer, L. B. (1956). *Emotion and Meaning in Music*. Chicago: University of Chicago Press.
- Meyers, O.C. (2007). *A mood-based music classification and exploration system*, MSc thesis,
Massachusetts Institute of Technology.
- McVicar, M. & Bie, T.D. (2012). CCA and a Multi-way Extension for Investigating Common
Components between Audio, Lyrics, and Tags. 9th International Symposium on
Computer Music Modelling and Retrieval (CMMR 2012), Queen Mary University of
London.

- Mihalcea, R., & Strapparava, C. (2012). Lyrics, music, and emotions. 590-599.
- Mikolov, T., Deoras, T., Kombrink, S., Burget, L., & Cernocky, J. (2011). Empirical Evaluation and Combination of Advanced Language Modeling Techniques, In Proceedings of Interspeech.
- Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. 1-12.
- Mohn, C., Argstatter, H., & Wilker, F.W. (2010). Perception of six basic emotions in music. *Psychology of Music*, 39(4):503-517.
- Morrison, S. J., Demorest, S. M., & Stambaugh, L. A. (2008). Enculturation effects in music cognition: The role of age and music complexity. *Journal of Research in Music Education*, 56(2):118-129.
- Mulins, M. (2008). Information extraction in text mining. Computer Science Graduate Student Publications. 4. http://cedar.wvu.edu/computerscience_stupubs/4
- North, A.C., & Hargreaves, D.J. (1997). Music and consumer behavior. *The Social Psychology of Music*. Oxford: Oxford University Press, 268–289.
- North, A. C., Hargreaves, D. J., & McKendrick, J. (1997). In-store music affects product choice. *Nature*.
- Panda, R., & Paiva, R.P. (2011). Automatic Creation of Mood Playlists in the Thayer Plane: A Methodology and a Comparative Study. 8th Sound and Music Computing Conference, Padova, Italy.
- Panda, R., Bruno, R., & Pedro, P. R. (2013a). Dimensional Music Emotion Recognition: Combining Standard and Melodic Audio Features.
- Panda, R., Ricardo, M., Bruno, R., António, O., & Pedro, P. R. (2013b). Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis.

- Pratt, C.C. (1952). *Music as the language of emotion*. Oxford, England: The Library of Congress.
- Przybyłek, S.(2016). *What is music definition, terminology, and characteristics*. Retrieved from: <https://study.com/academy/lesson/what-is-music-definition-terminology-characteristics.html>.
- Qi, X.R. (2018). *Exploring Corpora-Based Music Classification: Classifying Japanese Popular Music using Lyrics*. Department of Computer Science Tufts University.
- Russell, J. A. (1980). A circumspect model of affect. *Journal of Psychology and Social Psychology*, vol. 39, no. 6, p. 1161.
- Russell, J.A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145.
- Saari, P., & Eerola, T. (2014). 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. (2015). *Music mood annotation using semantic computing and machine learning*. Ph.D. thesis, University of Jyväskylä.
- Salimpoor, V. N., van den Bosch, I., Kovacevic, N., McIntosh, A. R., Dagher, A., & Zatorre, R. J. (2013). Interactions between the nucleus accumbens and auditory cortices predict music reward value. *Science*, 340(6129), 216-219.
- Sebastiani, F.: Machine learning in automated text categorization. *ACM CSUR*, Vol. 34, No. 1 (2002) 1–47.
- Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1(4):331-346.
- Schubert, E. (2003). Update of the Hevner adjective checklist, *Perceptual and Motor Skills*, vol. 96, pp. 1117–1122.

- Schubert, E. (2014). Perceived emotion with continuous musical features. *Music Perception: An Interdisciplinary Journal*, 21(4):561-585.
- Schuller, B., Dorfner, J., & Rigoll, G. (2010). Determination of nonprototypical valence and arousal in popular music: Features and performances. *EURASIP Journal on Audio, Speech, and Music Processing*, pages 1-19.
- Schuller, B., Weninger, F., & Dorfner, J. (2011). Multi-modal non-prototypical music mood analysis in continuous space: reliability and performances. *In Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference*, pages 759–764.
- Schwenk, H. (2007). Continuous space language models. *Computer Speech and Language*, vol. 21.
- Senac, C., Pellegrini, T., Mouret, F., & Pinquier, J. (2017). Music feature maps with convolutional neural networks for music genre classification. *In Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI '17*, 19:1–19:5, Florence, Italy. ACM.
- Singhi, A., & Brown, D. G. (2014). On cultural, textual and experiential aspects of music mood. *In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 3-8, Taipei, Taiwan.
- Soergel, D. (1998). WordNet. An Electronic Lexical Database.
- Song, Y., Dixon, S., & Pearce, M. (2012). Evaluation of Musical Features for Emotion Classification. *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*.
- Song, Y. (2016). The Role of Emotion and Context in Musical Preference School of Electronic Engineering and Computer Science Queen Mary, University of London.
- Tao, L., & Ogihara, M. (2004). Detecting emotion in music. *In ISMIR*, volume 3, pages 239-240.

- Teja, D.K. (2016). Content-Based Music Recommender System B. Thomas Golisano College of Computing and Information Sciences Rochester Institute of Technology Rochester, New York.
- Tekwani, B. (2017). Music Mood Classification Using The Million Song Dataset.
- Tellegen, A., Watson, D., & Clark, L. A. (1999). On the dimensional and hierarchical structure of affect, *Psychological Science*, vol. 10, no. 4, pp. 297–303.
- Temple, I. (2015, April 23). *How to Improve Your Well-Being through Music – Soundfly*. Retrieved from <https://flypaper.soundfly.com/features/how-to-improve-your-well-being-through-music/>.
- Thayer, R. E. (1989). *The Biopsychology of Mood and Arousal*. Oxford University Press, New York, USA.
- Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multilabel classification of music into emotion, in Proc. Of the Intl. Conf. on Music Information Retrieval, Philadelphia, PA.
- Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. *ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Pages 384-394, Uppsala, Sweden.
- Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2.
- Tzanetakis, G., & Cook, P. (2000). MARSYAS: A framework for audio analysis. *Organized Sound*, 4(3), 169-175. DOI:10.1017/S1355771800003071.
- Tzanetakis, G. (2009). MARSYAS Submission to MIREX 2009. The University of Victoria, Austrian Computer Society (OCG).

- Vuoskoski, J. K. & Eerola, T. (2011). The role of mood and personality in the perception of emotions represented by music. *Cortex (Special Section on Music in the Brain): Research Report*, 47(9):1099-1106.
- Wang, X., Xiaoou, C., Yang, D., & Wu, Y. (2011). Music Emotion Classification of Chinese Songs based on Lyrics Using TF*IDF and Rhyme. *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*. 765-770.
- Watson, D., & Mandryk, R.L. (2012). An in-situ study of real-life listening context *Proceedings of the 9th Sound and Music Computing Conference*, Copenhagen, Denmark 11-14.
- Wu, B., Zhong, E., Hao Hu, D., Horner, A., & Yang, Q. (2013). SMART: Semi-Supervised Music Emotion Recognition with Social Tagging. 10.1137/1.9781611972832.31.
- Xia, Y., Wang, L., Wong, K., & Xu, X. (2008). Sentiment vector space model for lyric-based song sentiment classification, in *Proc. of the Association for Computational Linguistics. Columbus, Ohio, U.S.A: ACL-08*, pp. 133–136.
- Yang, D., & Lee, W.S. (2004). Disambiguating music emotion using software agents. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04)*, pages 52–58.
- Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H.H. (2008a). A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 448–457.
- Yang, Y.-H., Lin, Y.-C., Cheng, H.T., Liao, I.B., Ho, Y.C., & Chen, H. (2008b). Toward multi-modal music emotion classification. *Advances in Multimedia Information Processing-PCM 2008*, pages 70–79.
- Yang, Y., & Chen, H. (2011). Music Emotion Recognition. *Multimedia Computing, Communication and Intelligence Series, CRC Press, Taylor & Francis Group*.

- Yang, Y.-H., & Hu, X. (2012). Cross-cultural music mood classification: A comparison of English and Chinese songs. *In Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 19-24, Porto, Portugal.
- Yang, Y.H., & Chen, H.H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*. Vol. 3, No. 3.1-30.
- Zaanen, M.V. & Kanters, P. (2010). Automatic Mood Classification Using Tf*Idf Based on Lyrics.
- Zeng, Z., Pantic, M., Roisman, G., & Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58.
- Zentner, M., Grandjean, D., & Scherer, K. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement, *Emotion*, vol. 8, no. 4, pp. 494–521.