Prediction of Emotion Stimulated by Music

C. V. Nanayakkara

December 2015

Prediction of Emotion Stimulated by Music

Charini Vimansha Nanayakkara
Index No: 11001984
University of Colombo School of Computing

Supervisor: Dr. H. A. Caldera

December 2015

Submitted in partial fulfillment of the requirements of the
B.Sc. in Computer Science 4th Year Individual Project (SCS4001)

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name:

……

Signature of Candidate Date:

This is to certify that this dissertation is based on the work of

Ms. Charini Vimansha Nanayakkara

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Supervisor Name:

……

Signature of Supervisor Date:

Abstract

Existence of a mechanism for predicting musically induced emotions is of value to the music therapy domain, while being a study of considerable interest. Despite numerous studies having been conducted in this area, certain limitations as the lack of a music specific emotion model has direly affected the conclusions of those studies. This research has focused on creation of a music specific emotion model, whereas an array of experiments have been conducted to realize a fair music emotion prediction model. The best prediction model was realized to be audio feature based music emotion prediction incorporating oversampling and Random Forest algorithm.

Acknowledgement

I express my sincere and heart felt gratitude to Dr. H. A. Caldera, senior lecturer at University of Colombo School of Computing for providing his invaluable advice on conducting the research project successfully. As an amateur researcher, the guidance he provided me with was non-trivial in completing the research. Thus I acknowledge him with immense gratitude for his commitment as a supervisor.

Support by my friends and family is immensely appreciated, whereas researchers whose work has been cited in this survey are acknowledged with gratitude.

Table of Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Definition
MER	Music Emotion Recognition
SVM	Support Vector Machine
MFCC	Mel Frequency Cepstral Coefficient
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
AMC	Audio Music-mood Classication
GEMS	Geneva Emotional Music Scale
BOW	Bag of Words
POS	Part of Speech
TF-IDF	Term Frequency – Inverse Document Frequency
KNN	K Nearest Neighbor
LSA	Latent Semantic Analysis
BR	Binary Relevance
LP	Label Power-set
MLKNN	Multi Label K Nearest Neighbor
RAKEL	Random K Label-sets
EFFC	Early Fusion by Feature Concatenation
LFLC	Late Fusion by Linear Combination
LFSM	Late Fusion by Subtask Merging
LMD	Language Model Difference
API	Application Programming Interface
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
ROC	Receiver Operating Characteristic
SMO	Sequential Minimal Optimization

RMSE	Root Mean Squared Error
OOBE	Out of Bag Error
SMOTE	Synthetic Minority Oversampling Technique
PC	Personal Computer
ARFF	Attribute Relation File Format
CSV	Comma Separated Values

Chapter 1 - Introduction

Music and rhythm find their way into the secret places of the soul

-Pluto (Philosopher and Mathematician in Classical Greece)-

Music is a form of art that has been embraced by the human kind since the very beginning of evolution. The capability people possess to intuitively interpret the notion conveyed by a music piece, even in the absence of words, is possibly accountable to the omnipresence of music in our lives since birth. In fact, research attest to the fact that sensitivity to music is expressed even during prenatal stage, whereas the capability of perceiving emotion in music develops since infancy [1]. Figure 1.1[1] depicts an experiment conducted to learn how infants react to music.

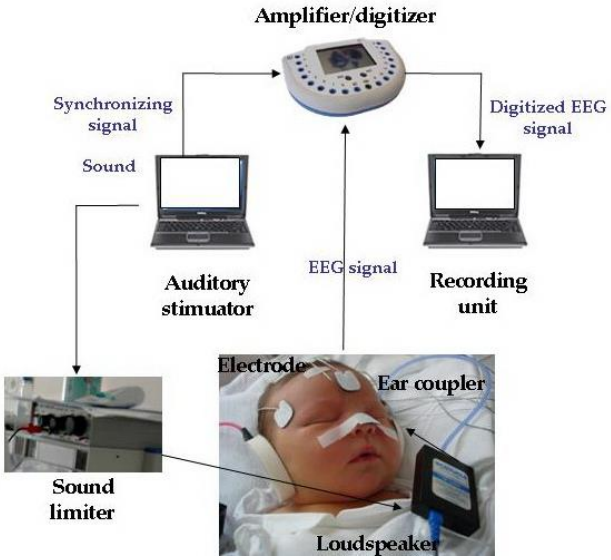

Figure 1. 1: An experiment conducted to learn how newborns react to music

In a research conducted for discovering methods adopted by persons for self-regulation of feelings, it was realized that the third most common and second most effective method was listening to music [2]. The capability of music to incur emotions is considered the primary reason behind it being considered an invaluable art [3].

[1] http://www.mcg.uva.nl/newborns/

Such extensive evidence concerning the impact music depicts on emotional aspects have motivated researchers to conduct number of studies on the matter during recent years.

1.1. Music and Emotion: A Background Review

Prior to pondering on the mystery surrounding music and emotion, it would be worthy to provide a concise clarification of the keywords concerning the research domain; *Emotion* and *Music*. The reader is provided with a basic understanding of how philosophers and psychologists have attempted to define the term emotion and how it relates to an aesthetic subject as music. This section discusses the contextual and prerequisite information essential to understand the main body of the thesis.

1.1.1. What is Emotion?

Afore discussing the definition of the term 'emotion', it would be interesting to learn why people associate the term emotion with music rather than mood. Despite many using these terms interchangeably, psychologists make a clear distinction between the two. Among the many characteristics which differentiate moods from emotion, the time factor plays a vital role. Moods comparatively tend to last much longer than emotions. However, the primary reason for emotions to be associated with music rather than moods is since people have the capability of specifying an object or event which evoked a certain emotion in them. On the contrary, the causal of mood remains vague more often than not [4] [5]. If music was considered an object, it would be difficult to associate mood with it for this matter.

With the confusion between mood and emotion resolved, a clarification of the term 'emotion' is required. Despite this being a universal phenomenon, there's no consensus as to the definition we could provide with to elaborate this term. Many philosophers and psychologists however, have made attempts at presenting their own theories regarding emotion. Five such renowned theories are as follows.

James – Lange Theory

According to this theory, people experience emotion due to the physiological reactions caused by external stimuli [5]. Thus, the order of occurrence of actions involved would be as described in Figure 1.2.

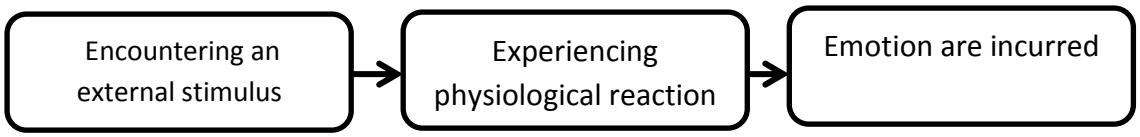

Figure 1. 2: James - Lange theory

For instance, if the fear of seeing a cobra is to be explained based on this theory, once a person sees the cobra, physiological symptoms as trembling and increase in heartbeat rate occurs. One would then interpret this to be fear. Simply put, it states that fear occurs due to the trembling and not vice versa. This theory is widely questioned since similar physical reaction as rapid heart rate subsequent to exercise doesn't incur fear in individuals [6].

Cannon – Bard Theory

This is also known as the thalami theory and opposes the former view. Cannon argued that despite certain physiological reactions one may not feel emotions, as in the former example of physical exertion. People would not have been able to identify distinct emotions if physical reactions were the only means by which to distinguish them, since emotions as fear and anger cause similar physiological responses [5]. Furthermore, acquiring emotions through analysis of physical reactions seems unlikely since people often feel emotion too rapidly for it to be a sole product of physical state. According to Cannon - Bard theory, physiological reaction and emotion both are caused simultaneously when a stimulus is encountered. When the brain receives a message based on the external event, a feeling and physical response is resulted [7].

Schacter-Singer Theory

This theory focuses on the cognitive aspect of emotion and is also known as the two-factor theory [7]. Unlike in James – Lange Theory, the person doesn't come to a conclusion of what to feel based merely on physical reactions, but takes into consideration the stimulus or nature of situation as well [8]. This fact has been proven with their infamous experiment using epinephrine (stimulant of the sympathetic system), where the participants who were injected with it interpreted what they were feeling as either anger or euphoria, based on what they were witnessing. In the former scenario, the participants were seeing an actor behaving angrily, whereas in the latter, they were witnessing a silly, euphoric act [5]. This theory further suggests that similar physiological response could affect different emotions in people, based on situation.

Cognitive Appraisal Theory

Despite there being former attempts at defining a relationship between emotion and cognition, Richard Lazarus is widely accepted as the pioneer of this theory since he was the first to claim that appraisal alone was sufficient for causing emotion [7]. According to this theory, thought occurs subsequent to experiencing a stimulus, following which emotion and physiological responses occur simultaneously [9]. The appraisal mechanism itself may be made on "simple sensory level or on the level of complex, conscious reasoning" as Lazarus has stated [10]. Whatever the mechanism of reasoning, cognition plays the fundamental role in effecting emotion in people according to this theory.

Facial-Feedback Theory of Emotion

This theory is somewhat on par with the James – Lange theory, in stating that physiological reactions do often have a direct impact on emotions. Apart from William James, Charles Darwin too maintained this fact on the grounds that, in certain instances, a specific facial expression may cause an emotion and not merely vice versa. As an example, it's stated that a person who's forced to smile may actually feel happier than he does when frowning [7].

When evaluating all these theories, it could be concluded that humans haven't yet succeeded in providing with a universal, unanimous definition of emotion. The eventuality of a survey conducted to clarify the meaning of emotion, with the participation of 33 renowned experts in emotion from various countries, was there being no consensus [11]. This could be partially accounted to the fact that the object which elicits emotions is invaluable in determining which emotions are evoked and their intensity [12]. For instance the sadness evoked by a song would rarely be as intense as the sadness one may experience at the demise of a loved one. Hence, it's worthy to study aesthetic emotions (of aesthetic context; related to works of art as music) as a separate discipline, rather than associating it with the holistic view of emotion [12]. Thus, the latter subsection is dedicated to clarifying the relationship between music and emotion.

1.1.2. What is Music and how is it Related to Emotion?

Music, according to the definition provided by WordNet, is “an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner” [13]. Despite existence of organized sound not being a sufficient condition for a composition to be called music [3], it undoubtedly is a significant characteristic of it. Among the many forms of music pieces, the one most listened to in the contemporary world is ‘song’, which comprises of lyrics apart from melody [3]. Thus, if the claim that music does elicit emotion is assumed to be true, both these characteristics of music could be perceived to contribute to the causal of it.

Even though many philosophical questions pertain as to why and how music evokes emotion, the irrefutability of it has motivated many researchers to focus on this area during recent years [12], [14], [15]. However, this does not necessarily imply that emotions are effected in all individuals by music [12]. It merely reflects that it’s worthy to conduct research in this domain since the majority of people are emotionally affected by music.

Two philosophers who have made commendable contribution in the music emotion domain are Stephen Davies and Jerrold Levinson. Stephen Davies claims that emotional response to music is due to ‘emotional contagion’ or ‘mirroring’ of the emotions conveyed by music [16]. Levinson states that savoring the emotions expressed by music could be equaled to ‘wine tasting’, where the bitterness is savored as much as the sweetness, which would explain why certain people enjoy listening to sad music [12]. However, both agree on the fact that most people believe music to be expressive of emotion [3]. The broad views on the music emotion theory could be categorized primarily into two, as the Cognivists’ approach and Emotivists’ approach. The former is on par with Davies’ theory in stating that music merely displays certain dynamic characters of emotions, thus making people believe that music piece is expressive of some emotion. It doesn’t necessarily imply that this emotion would be felt by the listener. On the contrary, Emotivists argue that music does actually have the potential of evoking emotions in people [3] [17] [18].

Three primary emotion categories have been identified to be associated with the context of music [19].

Expressed Emotion

This relates to the emotions a performer or composer wishes to communicate to listeners through a song. Even though one may intuitively feel that expressed emotion is the same as what a person would feel when listening to music, it has been empirically proven that it's not always the case [20].

Felt (evoked) Emotion

Emotion actually felt by the listener when listening to a song. This may differ from what effect he assumes a piece of music is trying to express and is the reason why certain people are not saddened by what he assumes to be sad music [21].

Perceived Emotion

Relates to the emotions a listener perceives or assumes as being expressed by a music composition. Despite this being highlighted as the main focus in MIR research [22], music related study based on felt emotion too have been conducted [12]. However, it's been proven that the capability of realizing which emotions are conveyed by music is correlated with the emotional intelligence of individuals [23].

Music elicits emotion in a more subtle manner than in the general perspective [24], where emotions are more felt than acted upon [12]. Hence, the term 'emotion' is regarded in the more restricted sense of 'feeling' in this context, where emotion wouldn't necessarily result in any visible physical reaction [3] [12].

Among the four aspects; structural features, listener features, performance features and contextual features; which are said to be decisive of what emotions an individual gets when listening to music [17], the former two have been used in most research work. While structural features comprise of audio and lyrics, listener features concern age, cultural background, etc. of the listener. If merely the song is considered to be responsible for the feelings perceived in a person, it should be caused by these structural features which are the audible characteristics of music. Thus, many research work [14] [25] [26] in the Music Emotion Recognition (MER) domain have considered lyrics and/or audio for identifying emotions perceivable when listening to a music piece.

Lyrics Features of Music

Lyrics are the words appearing in a song. Where lyrics are concerned, it could be intuitively supposed that the semantic meaning of lyrics may affect causal of emotion when listening to music [3]. Research [14] further attests to this fact since lyric based music emotion classification attempted in this research shows that words such as 'death', 'evil', 'hell', 'pain' are associated with the emotion category *angry*, whereas *not-angry* emotion class depicts close relation with terms as 'love', 'heart' and 'need'. Hence the semantic meaning of lyrics plays a vital role in effecting emotions in music listeners. However a non-trivial assumption made when utilizing lyrics to detect emotions caused by music, is that the listener understands what the words convey.

Audio Features of Music

Audio features relate to the sound waves produced when performing a musical composition. For instrumental compositions[2] this consists of the sound produced by musical instruments. If the composition is a song, the sound generated by the vocal chords of the singer too is included.

Despite the ability to presume that lyric features impact emotion due to its semantic meaning, it's less evident how audio features may contribute to it. However, one cannot deny the impact audio features seemingly have on emotions, since instrumental music too evokes feelings in people [3].

Unlike lyrics, audio cannot be represented textually. Thus, certain physical characteristics of audio are measured with equipment, which are used to represent sound which exist in the form of waves. Such physical characteristics could be depicted as values which are utilized for audio based music analysis tasks [14] [15] [26]. Scientific study of the correlation between physical features of sound and physiology is known as psychoacoustics [27]. Among the psychoacoustic groups specified in research [27], loudness, pitch (depicted by Figures 1.3 and 1.4 respectively[3]) and timbre (depicted by Figure 1.5[4]) are more focused upon in the context of music [22] [28]. Physical features of audio, such as Spectral Centroid, Spectral Flux and Mel Frequency Cepstral Coefficient (MFCC), which are derived using audio feature retrieval techniques, are directly associated with psychoacoustics. Thus they are often among the many audio features used in the MER domain.

[2] Instrumental music is music devoid of any lyrics and vocal characteristics. Sound is produced merely by musical instruments.
[3] http://www.yamahaproaudio.com/global/en/training_support/better_sound/part1/01/
[4] http://www.musictheoryis.com/timbre/

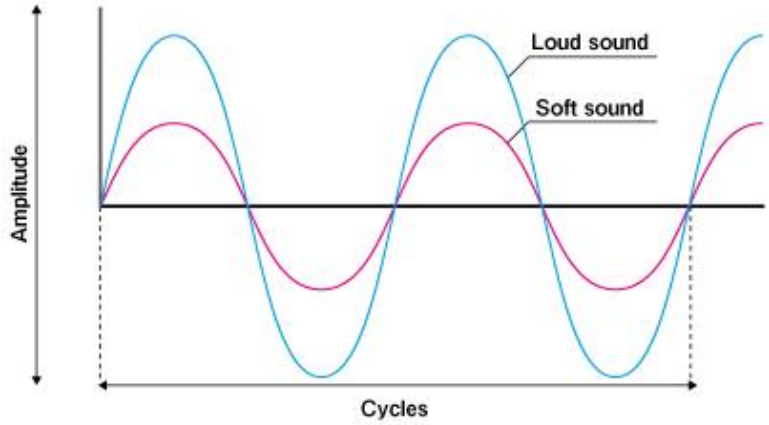

Figure 1. 3: Loudness of Sound Wave

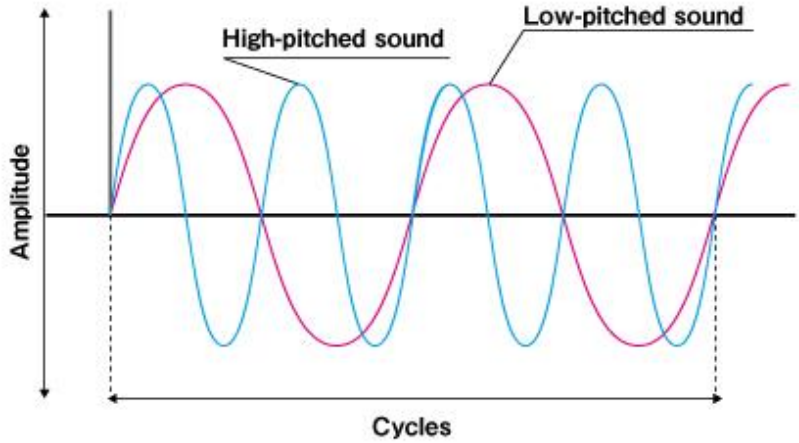

Figure 1. 4: Pitch of sound wave

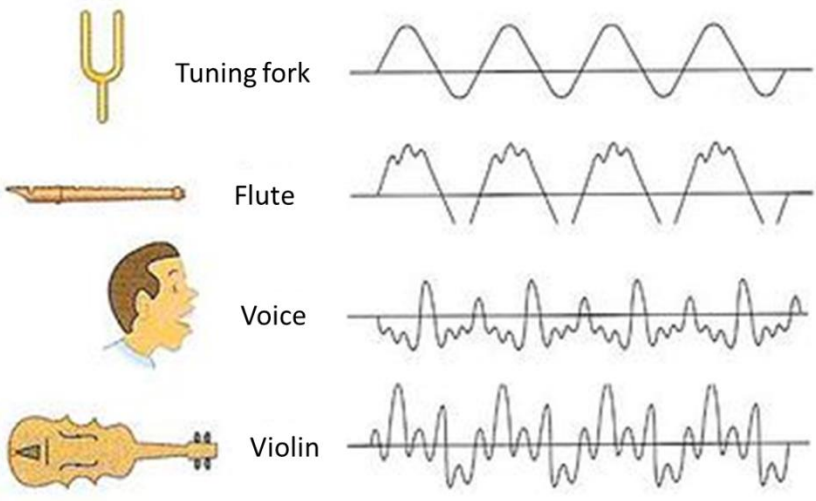

Figure 1. 5: Timbre of sound wave

1.2. Motivation

According to a study by Huron, the most useful retrieval indexes of music are social and psychological functions as emotion and stylistic [22]. Hence, MER, a branch of Music Information Retrieval (MIR) is an area of study of considerable significance in the music industry [22]. Furthermore, music emotion detection provides the basis for numerous applications as song selection in mobile devices, music recommendation systems, TV/ radio programs and music therapy [29]. Music has shown immense value as a therapeutic tool due to its capability of artificially eliciting emotions in the listeners [30]. Patients suffering from autism, cancer, Parkinson, Alzheimer's disease, depression, etc. are often provided with music therapy. Figure 1.6[5] depicts the results of such treatment conducted for elderly persons. Prior to selecting music pieces for treatment it would be much beneficial to learn the kinds of emotion they tend to evoke in people. Such extensive applicability of music emotion recognition combined with the personal liking for the music field, provided motivation for conducting a research on this topic.

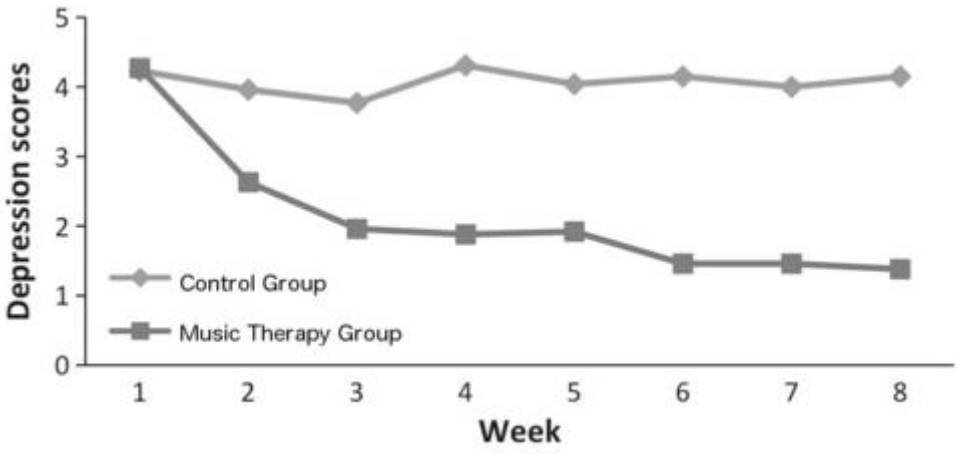

Figure 1. 6: Results of long term music therapy in the elderly

[5] http://neurowiki2013.wikidot.com/individual:benefits-of-music-on-mood-disorders

1.3. Research Question

The primary research problem addressed in this thesis is **evaluating capability of automatically predicting the emotions expressed by a music piece, with a fair level of accuracy**. Emotions perceived by an individual when listening to a certain song, is referred to as 'emotions expressed by a music piece' in this study. This research problem would be addressed by modelling several data driven approaches for prediction of musically induced emotion and evaluating them to ascertain their level of acceptability. Two research questions are necessary to be addressed to realize a solution for this primary research question.

The initial question necessary to be addressed is determining the distinct emotions which could be expressed by music in general. It's required to identify the classes of emotion which are elicited in people by music, since music may be expressive of merely some emotions but not all. Furthermore this is necessitated by the distinction of musically induced emotion from the general class of emotions [12] (i.e. musically induced sadness is not as bitter or intense as sadness caused by personal mishap).

The second question concerns identifying a model which has the capability of assigning a given music piece to emotion categories determined in the former phase. This model must have the capability of automatically identifying the emotions expressed by a song, by analyzing the structural characteristics (lyrics and audio features) of it.

1.4. Scope

Research is subject to several assumptions and constraints due to the incapability of covering the vast area belonging to the domain of music and emotion. Merely English songs would be considered for the research task for convenience of interpreting the lyrics and due to their vast availability. It is further assumed that the listeners have the capability of understanding English, since it may impact the emotions they perceive when listening to a song.

Merely structural features (audio and lyrics) of music are considered for the study whereas listener, performance and contextual features have been disregarded. This is based on the assumption that merely the audible characteristics of a song are responsible for conjuring emotions in people. It is further assumed that some of these structural features are correlated with musically expressed emotions, thus allowing an automated model to discriminate among emotions by studying the features of songs.

As elaborated under a latter section, tags people have associated with music would be used for determining emotions expressed by a music piece. Thus it's assumed that if the semantic meaning of a tag conveys emotion, it reflects the emotion one perceives when listening to that song (the song with which the tag is associated). Furthermore, in the attempt made for reducing dimensionality of this emotion space using clustering technique, it's assumed that emotions which frequently appear together in same songs belong to the same emotion cluster.

The research is further subject to the constraint of utilizing free and open source software for research requirements, whereas the research would be conducted in such a way, so that it's feasible to produce satisfactory results within given time frame.

1.5. Contribution

Despite a number of music emotion research having been conducted utilizing audio and lyric features, few studies have assessed the comparative and combined performance of these features for a single dataset. Analysis of both multi-label[6] and single-label[7] approaches using a single dataset too, has not formerly been attempted. Furthermore, a considerable number of research work lack focus on music specificity of the emotions they have associated with music pieces. Rather, the bulk of MER research have utilized generic emotion models and attempted to describe musically induced emotion within this constrained model. On the contrary, we have attempted to retrieve perceived emotion from the music context itself, using tags. A noteworthy aspect of the research is the magnitude of the dataset used for the research. All MER related researches we encountered have conducted experiments using datasets comprising less than 3000 data instances. Thus, this study has the advantage of its derivation of results being supported by a vast dataset, much larger than what many have utilized. While the knowledge required for determining the approach of conducting research was derived from the literature, the overall research method is distinct from former research. The research method has been determined following extensive analysis of existent approaches and combining the positive aspects of each. Thus this research is a unique and novel attempt in the MER domain.

[6] *Multi-label classification:* A single record is associated with several class attributes. In this research a song would be classified into several emotion categories when multi-label approach is used.

[7] *Single-label classification:* Only one class attribute is present. Could be further categorized as multi-class and binary classification, where class attribute could assume several distinct values in former and merely two values in latter.

1.6. Organization of the Dissertation

Subsequent chapters of the Dissertation are organized as follows. The second chapter discusses research conducted on music emotion prediction and the strengths and weaknesses of those studies. Third chapter provides with an overview of the methodology whereas chapter four has provided with an overview of the experimental setup, inclusive of description of algorithms attempted. Chapter five has dealt with feature engineering tasks conducted in research. The sixth chapter extensively describes the mechanism followed in creation of music specific emotion model, whereas the seventh chapter relates to discussion of results and analysis. Eventually the eighth chapter concludes the study with an overview of potential future work.

Chapter 2 - Literature Review

2.1. Introduction

The former chapter provided with a concise introduction to the domain of music and emotion, whereas the knowledge prerequisite of the reader to comprehend the research objective was delivered. It further focused on the research questions and scope, thus informing the reader which specific areas the thesis would cover.

This chapter discusses the approaches followed in former research work for determining the emotions expressed by music pieces. As discussed under section 1.3, two research questions require to be addressed for the purpose of solving the research problem. Thus, the survey has been divided into two distinct subchapters for convenience of the reader to comprehend how each question has been addressed in existent literature. The first research question concerns determining emotion categories music is capable of evoking in individuals. This question has been extensively discussed under section 2.2. The second question of identifying a model which has the capability of assigning a given music piece to emotion categories is discussed under section 2.3.

2.2. Identifying Emotion Categories

A majority of MER research conducted to date rely on generic emotion models, which do not necessarily focus on the context of music [22] [24]. Rather, those could be regarded as attempts of capturing the range of emotions people are capable of experiencing in its entirety. These generic emotion models could be primarily categorized as categorical and dimensional, where the former defines emotions with discrete values as adjectives and the latter defines emotions as continuous values represented along axes [22]. While Thayer's and Russell's models come under dimensional category, Farnsworth's and Tellegen-Watson-Clark models are identified as categorical [22].

2.2.1. Dimensional Models

Russel's Model

This model describes emotion in 2D space where x-axis represents *valance* and y-axis denotes level of *arousal*. Positive and negative depicts activation and deactivation with respect to arousal, whereas pleasure and displeasure are shown with respect to valance [14], [31]. To simplify the definition of these terms, positive arousal (or activation) depicts a state of alertness, increase in heart rate in the individual experiencing an emotion falling under this category. Likewise negative arousal (deactivation) conveys the opposite of this notion. Excitement would be an example for positive arousal whereas serenity would be an example for negative arousal. Positive valance (pleasure) reflects the attractiveness of a certain emotion. Emotions depicting positive valance are generally desired to be felt by individuals. On the contrary, negative valance (displeasure) relates to emotions we dislike to feel. Happiness is an example for positive valance while sadness is an example for the latter.

Thayer's Model

Is an extensively applied model [32] [33] [15] [25] which extends Russell's model by defining arousal axis as energetic or tense (stress) arousal [32] [31]. This is the same as Russel's model which combines valance with arousal. Figure 2.1 depicts Thayer's model separately using energy and stress as its dimensions, whereas Figure 2.2 represents both models in a single diagram.

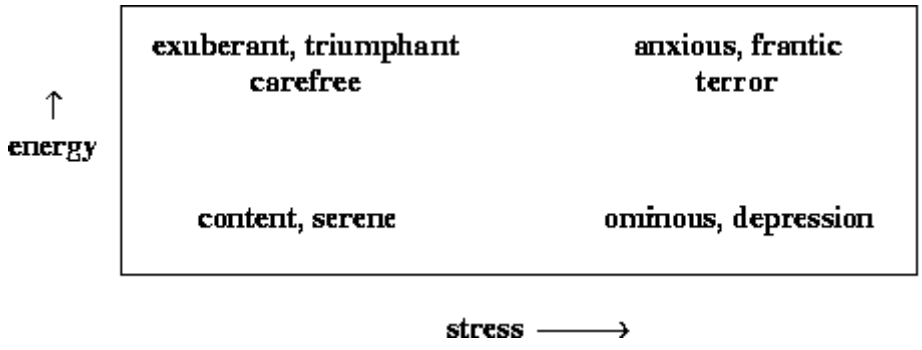

Figure 2. 1: Thayer's model

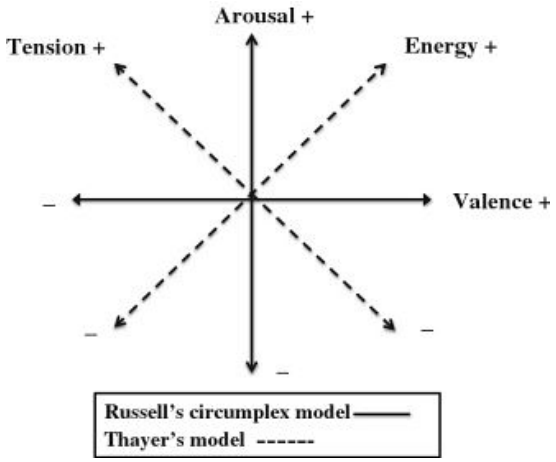

Figure 2. 2: Thayer's and Russel's models [31]

2.2.2. Categorical Models

Tellegen-Watson-Clark Model

This model associates adjectives with each class, hence making it a categorical model. Merely six of these classes have been incorporated in research [29]. Figure 2.3 depicts the Tellegen-Watson-Clark model where (-) represents the classes not utilized in [29].

Figure 2. 3: Tellegen-Watson-Clark model

Farnsworth's Model

This model has regrouped the original 8 clusters of Hevner's into 10 novel adjective clusters [34] [22]. As depicted in Figure 2.4, it has been further extended with classes marked with (+) in research [34]. Extension has been performed by a selected individual. This extended version has then been reduced to six super-groups as (A-B), (C-D), (E-L), (H-I-J), (G-K) and (F-M) [34]. However this extension and reduction are highly unacceptable since they have been performed by a single person, which biases the outcome to his perception.

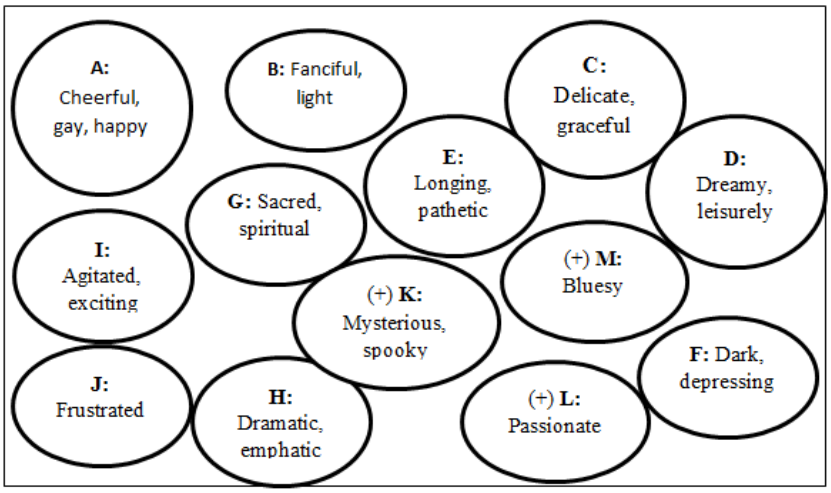

Figure 2. 4: Farnsworth's model

Even though these models have been originally formed by renowned psychologists, several studies [24] [12] have depicted that they are not the most effective methods of determining musically induced emotion. This is highly accounted to the restriction it imposes on the listener to define any emotion they experience within a constrained model. A significant limitation of dimensional models is the difficulty of making clear distinction between certain emotions with similar arousal, valance levels. For instance both anger and fear have positive arousal and negative valance, thus assigning them to the same emotion category [22]. Furthermore, categorical models are highly subjective to the designers' perception of what emotion is.

Researchers who have identified the limitations incurred by adopting these generic emotion models for identifying emotions related to the music context have deviated from basing their work on generic models. Following are former research attempts at discerning music specific emotions.

2.2.3. Model Based on Hierarchical Clustering

Research [35] submitted for 2007 MIREX AMC contest has identified five emotion clusters. These music specific emotion clusters have later been utilized in several other MER research tasks [26] [36] [37] as well. Dataset derived from AllMusicGuide.com (AMG) has been utilized for this task, since it comprises of 179 emotion labels which music experts have associated with songs in the dataset. Emotions associated with at least 50 albums and 50 songs have been retained for purpose of identifying emotion clusters. 40 emotion labels have thus been selected. This allows formation of two sub-datasets with album-emotion, song-emotion pairs. Clustering has been conducted as follows on each dataset independently.

Per each pair of emotion labels, the count of albums/ songs sharing that label has been calculated. Based on the count, Pearson's correlation per each pair of tags has been deduced. Subsequently, agglomerative hierarchical clustering has been conducted using Ward's criterion.

Following comparison of the clustering results for the two datasets, it has been realized that 29 tags get clustered identically in both scenarios. Five music-related emotion clusters as depicted in Table 2.1 have thus been identified in this study.

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Rowdy	Amiable/ Good natured	Literate	Witty	Volatile
Rousing		Wistful	Humorous	Fiery
Confident	Sweet	Bittersweet	Whimsical	Visceral
Boisterous	Fun	Autumnal	Wry	Aggressive
Passionate	Rollicking	Brooding	Campy	Tense/anxious
	Cheerful	Poignant	Quirky	Intense
			Silly	

Table 2. 1: Emotion grouping using hierarchical clustering [35]

2.2.4. Model Based on K-means Clustering

This approach has used tags people have associated with songs, from the music website, last.fm [38]. Of the many unique tags encountered in the website, 19 have been retained since they convey emotion. The research has used 2554 songs where a 19 dimensional vector has been maintained per song to record whether each tag was related to it or not. K-means clustering using hamming distance has been executed on this space, which has formed mutually exclusive clusters of the songs considered. Hence the tags of the songs in a cluster have been taken to represent an emotion category. The most accurate clusters have been formed when k=3 as the maximum silhouette value was acquired for 3 clusters. Figure 2.5 depicts the evaluation of these clusters using Principal Component Analysis (first two components). Emotions depicting highest similarity from each cluster have been placed close to one another on the two dimensional plane.

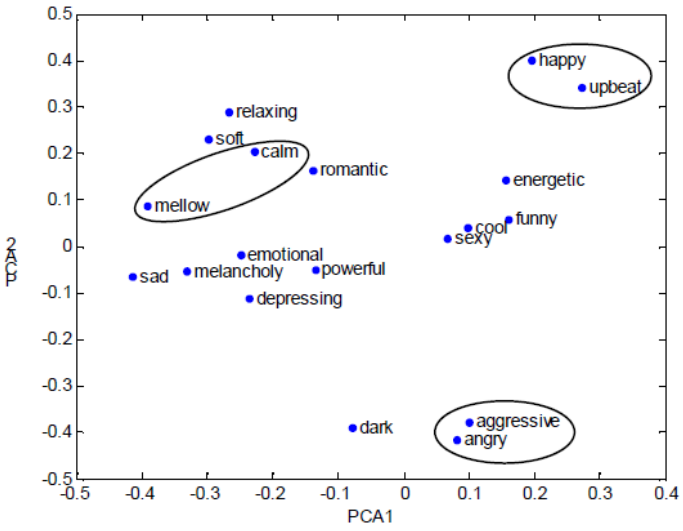

Figure 2. 5: K-means emotion clusters [38].

Despite this being a music specific emotion model, it has not often been applied in other related researches since forming merely three clusters has been viewed as domain oversimplification [39]. Furthermore, it's worthy to note that forming three clusters such that each emotion belongs to some cluster is not successful. This is well depicted in Figure 2.5 where only two tags from each cluster are close to one another when distance is considered. As discussed in 2.2.3, hierarchical approach better reflects repeated combination of tags based on distance unlike K-means.

2.2.5. Grouping Synonymous Emotion Tags

This approach has introduced a unique 18 emotion class model based on music tags retrieved from last.fm [39]. Tags irrelevant to music have been removed and synonymous words have been combined to form emotion clusters. Clusters which comprised of at least 20 songs have been retained, which has resulted in the 18 clusters [39] of emotion related tags depicted in Figure 2.6.

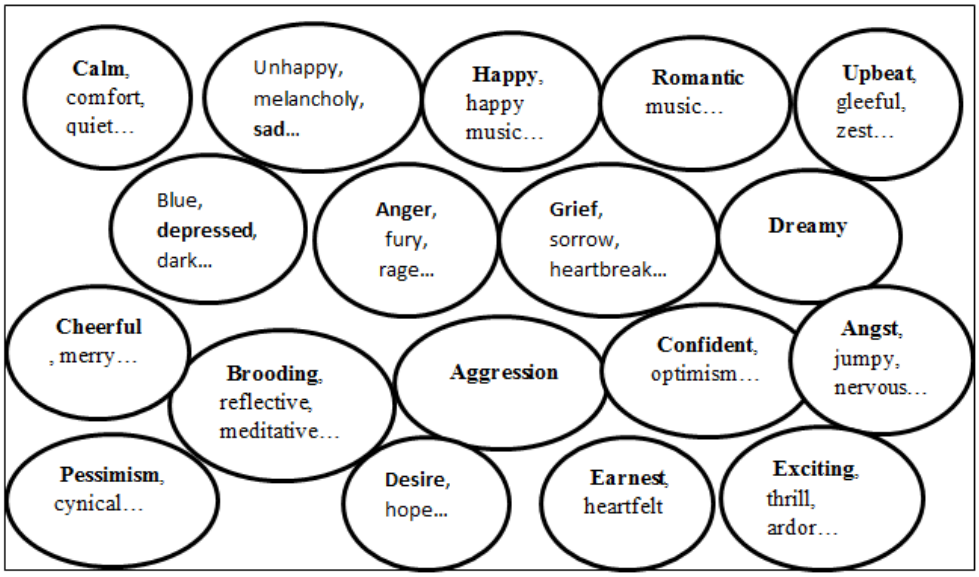

Figure 2. 6: Eighteen emotion categories [39].

Determining emotions expressed by a music piece by analyzing the tags associated with them helps build a music specific emotion model. However, merely the linguistic meaning of words has been considered to form emotion clusters in this scenario. It has resulted in overlooking the possibility of combining some clusters, based on their frequency of occurring together (i.e. set of emotions frequently occurring together in many songs could be combined to form a single emotion cluster). This latter phenomenon has been addressed by the researches described under sections 2.2.3 and 2.2.4.

2.2.6. Geneva Emotional Music Scale (GEMS)

GEMS [12] is considered to be the most music specific emotion scale to be constructed to date [24]. Unlike in former approaches this scale makes a clear distinction between perceived and felt emotion [12]. GEMS has been constructed as a result of 4 related studies, where the first two have been conducted for the purpose of compiling a list of music specific emotion terms and to determine how often individuals have felt and perceived those emotions. The 3rd study has been executed at a music festival for determining whether identified emotions could be categorized, utilizing confirmatory factor analytic procedures. This has resulted in the 9-factorial model of music induced emotion which has been validated as a reliable music emotion model in study four. The 9 dimensions are as follows, whereas the higher level and lower level dimensions are further described in [12].

1) wonder
2) transcendence
3) tenderness
4) nostalgia
5) peacefulness
6) power
7) joyful activation
8) tension
9) sadness

While the GEMS emotion scale is highly recommended for MER tasks [24], the approach of utilizing tags for music emotion detection is more practical for a research to be conducted within a time frame of one year. Despite the potential of distributing questionnaires to determine the emotions felt/ perceived by individuals when listening to songs, a dataset of considerable magnitude would have been difficult to be composed if that measure was taken. The number of data instances used to form the final model has a considerable impact on its accuracy and acceptability. Thus the best approach was decided to be utilizing tags for determining emotion classes. A limitation of this approach is the inability to determine whether tags express felt or perceived emotion. Since emotions are more often perceived than felt [12], the assumption (as specified under section 1.4) that tags express perceived emotion is acceptable.

2.3. Classification of Music into Identified Emotion Categories

Subsequent to deciding on an apt emotion model, it's possible to incorporate these emotion classes in the dataset. Thus, each song in the dataset would have one or several emotion labels associated with it (label is an emotion class). This dataset is used to determine the final model that could predict emotions elicited by a song. Methods used in former researches to realize the optimal model for prediction of emotion induced by music, have been presented under this chapter.

Existent research work could be categorized with respect to the features of music (audio and/or lyrics) utilized for classification or the classification approach followed. Classification approaches could be primarily categorized as multi-label and single-label approach. In the former instance, the capability for one song to evoke several emotions is reflected, whereas only the most representative emotion is considered in the latter. Thus, if the first approach is adopted, a music piece would get classified into several emotion categories. The second approach classifies a song into one emotion class alone. This section has been organized based on the music features considered in each research.

2.3.1. Classification Based on Lyrics

Words appearing in lyrics of music have been considered in this approach to classify music according to identified emotion categories.

Words appearing in lyrics have initially been retrieved and represented in a manner, such that feeding them to a model is simplified. Following representations of words have been extensively adopted in MER domain [39].

- Bag-of-words (BOW): A collection of unordered words from lyrics. Stemming in BOW refers to merging words of same morphology.
- Part-of-speech (POS): Lexical categories such as nouns and verbs
- Function words: Words like 'the', 'a'. Effective in text style analysis

After determining lyric features to consider, these have been presented as a feature vector. This allows for a song to be replaced by a vector of lyric features together with an emotion label(s). A song is required to have some value corresponding to each lyric feature. The feature value of a song has been assigned using one of the following methods in MER research work [39].

- Frequency or normalized frequency of words with respect to a music excerpt
- Term Frequency – Inverse Document Frequency (TF-IDF) weighting
- Boolean value indicating presence or absence of a term in a music piece.

To clarify this procedure further, assume the BOW model was used and the feature vector comprised of terms <a1, a2, a3>. In such a scenario if Boolean values were used as the scores, the vector of song s1 would be <1, 0, 1> assuming words a1 and a3 appeared in s1 but not a2.

A comparison of the accuracy of each method is depicted in Table 2.2, which conveys that BOW-Stemming method with TF-IDF weighting has produced best results in research [39]. Classification has been performed using Support Vector Machine (SVM) with linear kernel [39].

Representation	**Boolean**	**Term Frequency**	**Normalized TF**	**TF-IDF weighting**
BOW-Stemming	0.5748	0.5819	0.5796	0.6043
BOW	0.5817	0.5829	0.5840	0.5923
POS	0.5277	0.5763	0.5691	0.5571
Function words	0.5653	0.5733	0.5692	0.5723

Table 2. 2: Comparison of Lyric Based Methods [39]

Research [14] has attempted three different lyric based classification approaches. In the first experiment lyrics have been represented in BOW model with TF-IDF weighting, to retain words most significant for a specific song. K Nearest Neighbor (KNN) classification has been performed on this song set to realize the k groups of most similar songs. The emotion class of each group is then considered to be the emotion associated with most number of songs in that group. This experiment has relied on the assumption that 'similar' songs are most likely to belong to a single emotion category. The optimal average classification accuracy of 62.5% has been obtained when k = 7.

The second experiment has dealt with the dimensionality issue of lyrics via adopting Latent Semantic Analysis (LSA) method in unison with TF-IDF weighting. LSA is a commonly applied method for identification of documents which are semantically related, or depict a high rate of common words. When applied in the music context, the significant terms contributing to a specific dimension would depict the common lyric-words, associated with the songs related to that dimension. The number of dimension selected has significant impact on classification accuracy. Highest classification accuracy of 61.3% has been obtained when SVM classification is applied for 30 dimensions.

The final experiment adopts Language Model Difference (LMD) where lyric words associated with contradicting emotion categories have been compared. For instance, the 200 lyric-words most frequently associated with 'angry' emotion category are compared with that of the 'not-angry' emotion category. The most discriminative terms are then selected to represent each emotion class. Difference of document frequency associated with each term is used to evaluate how discriminative two terms are. This difference has been calculated as a combination of both absolute and relative differences since both reflect advantageous and disadvantageous aspects. Best accuracy of 80.7% has been obtained in this experiment, when SVM classification is performed on 100 most discriminant terms retrieved using LMD.

A significant disadvantage of lyrics only classification methods is their incapability of classifying instrumental music. Such music pieces have audio features alone, thus making their classification using this method impractical

2.3.2. Classification Based on Audio Features

Audio features retrieved from sound waves generated when performing music pieces have been considered in this approach to classify music according to identified emotion categories.

Classifying music into emotion categories has often been performed based on audio features such as Rhythmic features [29] [34], Timbre features [29] [34], Pitch content features [34] and Spectral features [39] [40] (e.g. Mean and variances of Spectral Centroid, Roll off, Flux and MFCC). Marsyas tool [40] has often been utilized for these feature retrieval tasks [29] [34]. A numerical value would correspond to each audio feature, with respect to a song, thus allowing a song to be represented as a pattern of audio features.

Research [29] has classified music based on audio features, into emotion categories defined in reduced Tellegen-Watson-Clark model (shown in Figure 2.3). Classification has been performed using the multi-label classifiers; Binary Relevance (BR), Label Power-set (LP), Multi Label K Nearest Neighbor (MLKNN) and Random K Label-set (RAKEL) algorithms; using SVM as base classifier. Comparison of these methods has proven RAKEL to be most accurate as shown in Table 2.3 [29].

	BR	**LP**	**RAKEL**	**MLKNN**
Average precision	0.7378	0.7669	0.7954	0.7104

Table 2. 3: Audio based classification: Multi-label approach [29]

Research [34] depicts another attempt at classifying music into emotion classes using audio features. The attempted method is a single-label approach utilizing SVM classifier. Music has been classified into emotion categories identified in extended version of Farnsworth's model (shown in Figure 2.4). However they have failed to produce satisfactory results, which they have accounted to emotion labels not being equally distributed across music types.

Even though classification of music into emotion categories using either lyric or audio features is a common approach, hybrid methods where features are combined are found to improve accuracy [39] [14]. Comparison of single feature and hybrid methods depicted in Table 2.4 is clear evidence for this matter.

Features / **Emotion class**	**Audio**	**Lyrics**	**Hybrid (Lyrics + Audio)**
Angry	98.1%	77.9%	98.3%
Happy	81.5%	80.8%	86.8%
Sad	87.7%	84.4%	92.8%
Relaxed	91.4%	79.7%	91.7%

Table 2. 4: Comparison among lyric, audio and hybrid approaches [14]

2.3.3. Hybrid (Lyrics + Audio) Classification

Combination of lyrics and audio has been considered in this approach to classify music according to emotion categories. Research [25] has adopted multi modal fusion methods to conduct hybrid classification as shown in Figure 2.7. Early Fusion by Feature Concatenation (EFFC), Late Fusion by Linear Combination (LFLC) and Late Fusion by Subtask Merging (LFSM) are these multi modal fusion methods utilized. Modals have been trained using SVM classifier either prior to or after merging audio and lyric features, as specified by the used Multi-modal fusion method. Table 2.5 depicts how audio and lyric features have been retrieved to execute this research experiment.

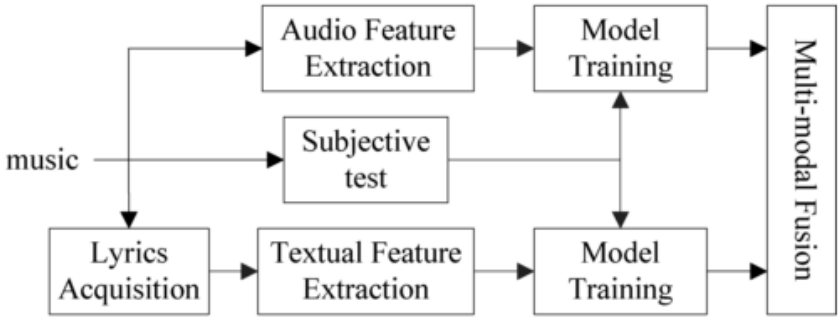

Figure 2. 7: System diagram of hybrid method [25].

	Method/ Tool	**Features**
Audio	Marsyas	MFCC
	PsySound	Spectral centroid, moment, roughness
Text from lyrics	Uni-gram: Counts unique terms. TF-IDF weighing applied	BOW
	PLSA: Probabilistic LSA which improves classification with increased overlapping of semantic words.	Latent vectors
	Bi-gram: Counts unique pairs of words. TF-IDF weighing applied	BOW

Table 2. 5: Feature extraction methods (audio, lyric) [25]

SVM classifier has been adopted for training models with relation to Thayer's emotion classes in this research. Multi-modal fusion methods and their accuracies are displayed in Table 2.6. 'Audio only' and 'Text only' methods serve as baseline for conveying superiority of hybrid models.

Method	**Accuracy**
Single feature methods	
Audio only	46.63%
Text only	40.01%
Multi-modal fusion methods	
EFFC: Combines audio and lyric features prior to training a single model	52.48%
LFLC: Separately trained models for audio and lyrics are combined linearly	55.34%
LFSM: Classification results for arousal and valance planes using lyric and audio models are eventually combined	57.06%

Table 2. 6: Comparison of Lyric+ Audio Models [25]

Research [14] depicts another attempt at combining audio and lyrics for classification of music into emotion categories. Audio features have been represented in the same vector space with lyric-words derived using LMD. SVM classification has then been performed on this feature space. The averages obtained via 10-fold cross validation have been considered as final results, as depicted in Table 2.4.

According to Tables 2.4 and 2.6, the performance of classification using different sets of features, in descending order would be: Hybrid approach, Audio based approach and Lyric based approach.

2.4. Summary

This chapter discussed how the two primary research questions have been addressed in existent literature. The first research question of determining an appropriate emotion model has been addressed by using either generic emotion models or by forming music specific emotion models. While adopting generic emotion models has been realized to lack emotion specificity, forming an emotion model using the music dataset itself is considered more appropriate. Tags associated with music pieces have often been used to determine emotions elicited by a music piece, when forming these music specific emotion models. These emotion related tags have been further combined based on their synonymy or frequency of occurring together in number of songs. Unsupervised learning methods (i.e. since actual emotion cluster of a tag is unknown, learning is unsupervised) as hierarchical and k-means clustering have been executed to perform the latter. Despite it not being attempted, performing reduction of dimensionality of emotion space considering both synonymy and frequency would potentially be a better approach. The second research question of identifying a model which is capable of predicting emotions expressed by a given song has been addressed in numerous ways in existent literature. Certain studies have used a single feature type (either lyrics or audio) for predicting the emotions of a song, whereas others have adopted hybrid techniques (combination of audio and lyrics). The latter method has depicted higher accuracy according to literature, which is followed by the audio based approach. The mannerism in which second question has been addressed could be further categorized based on the number of emotions associated with each song in dataset. Multi-label classification has been executed in certain researches which reflect the potential for one song to evoke several emotions. Others have utilized single-label approach which has the potential of predicting the emotion most representative of a specific song. Thus the second question is addressable in six distinct ways as; hybrid + multi-label, audio + multi-label, lyrics + multi-label, hybrid + single-label, audio + single-label and lyrics + single-label.

Chapter 3 elaborates on the methodology proposed to address the research problem, which has been devised based on knowledge acquired from studying the literature.

Chapter 3 - Methodology

3.1. Introduction

The former chapter provided with insight regarding how the two research questions specified under section 1.3 have been addressed in existent literature. The survey helped determine which approaches possess better potential for addressing the research problem effectively.

The objective of this chapter is outlining the overall methodology proposed for addressing the research questions, which would ultimately lead to a solution for the research problem. It is inclusive of the non-trivial assumptions and key concerns which would have considerable impact on the eventual outcome. The technical aspects of the approach are elaborated under latter chapters whereas the reader is provided with a high level view of research procedure in this chapter.

3.2. Fundamental Approach

Since the ultimate objective of this methodology is addressing the research problem specified in section 1.3, for which we are required to address the research questions, the methodology devises a music specific emotion model and evaluates the capability of different classification models to predict the emotions expressed by a song. On par with the scope of research specified under section 1.4, this methodology would be utilizing merely structural features of music pieces for prediction of musically elicited emotions. Furthermore, the study adopts a data-driven approach similar to those encountered in former research. This necessitates the collection of structural data of music pieces and the emotions people have expressed to have perceived when listening to those music pieces. Using this information, a model could be formed for predicting emotions expressed by any song (subject to constraints specified in section 1.4).

Several limitations and inadequacies with relation to former approaches were encountered when studying the literature. Two noteworthy limitations were inadequacy of the magnitudes of the datasets utilized and basing majority of researches on generic emotion models. Despite the availability of millions of music compositions, many researchers [38] [14] [29] have resorted to using datasets comprising of less than 3000 songs, which limits the generalizability of their findings. In general larger samples are recommended based on which to make predictions [41]. Thus, this research would be based on a dataset which is comparatively of much larger magnitude (larger than twice the size keeping 10,000 as baseline) than datasets utilized in former research. However the generalizability of the conclusions of this research is constrained by assumptions made under section 1.4. Associating generic emotion models with music pieces is considered less appropriate [12] due to which this research focuses on forming a music specific emotion model. Music specificity is defined by deriving emotions from the music dataset itself. This objective is achieved by using tags associated with songs to form an emotion model for the study. Hence, this methodology focuses on resolving these and other limitations (specified in section 1.5) encountered in former attempts.

Proposed methodology is grounded on the knowledge acquired from studying the literature while resolving their inadequacies. Therefore, the overall methodology has been devised via incorporating the positive aspects of different research work studied in the literature, with additional preprocessing, feature engineering tasks included.

3.3. Key Concerns

Several key concerns are associated with attempting to devise a methodology for addressing research problem while resolving shortcomings formerly specified. The primary concerns required to be addressed are as follows.

1) Since magnitudes of formerly utilized datasets were found to be inadequate, a comparatively larger music dataset comprising of necessary information (structural features and tags) requires to be obtained. Evaluating the aptitude of dataset for realizing research objectives would be immensely beneficial in ensuring the level of acceptability of final outcome.
2) A comprehensive method of devising a music specific emotion model must be determined.
3) The structural features which contribute most to the final prediction model require to be retained for enhancing efficiency and accuracy.
4) The level of accuracy with which the final model can predict emotions must be discernible.
5) Model must be highly scalable and adaptable to new data, since it concerns a field which grows with immense rapidity. Thus, it must be possible to easily incorporate novel music pieces in the model.

3.4. Overall Architecture

This section presents the overall architecture diagram of proposed methodology, which has been modeled to address the research questions effectively. Key concerns have been focused upon when developing the architecture. Figure 3.1 depicts the architectural view which comprises of five primary components.

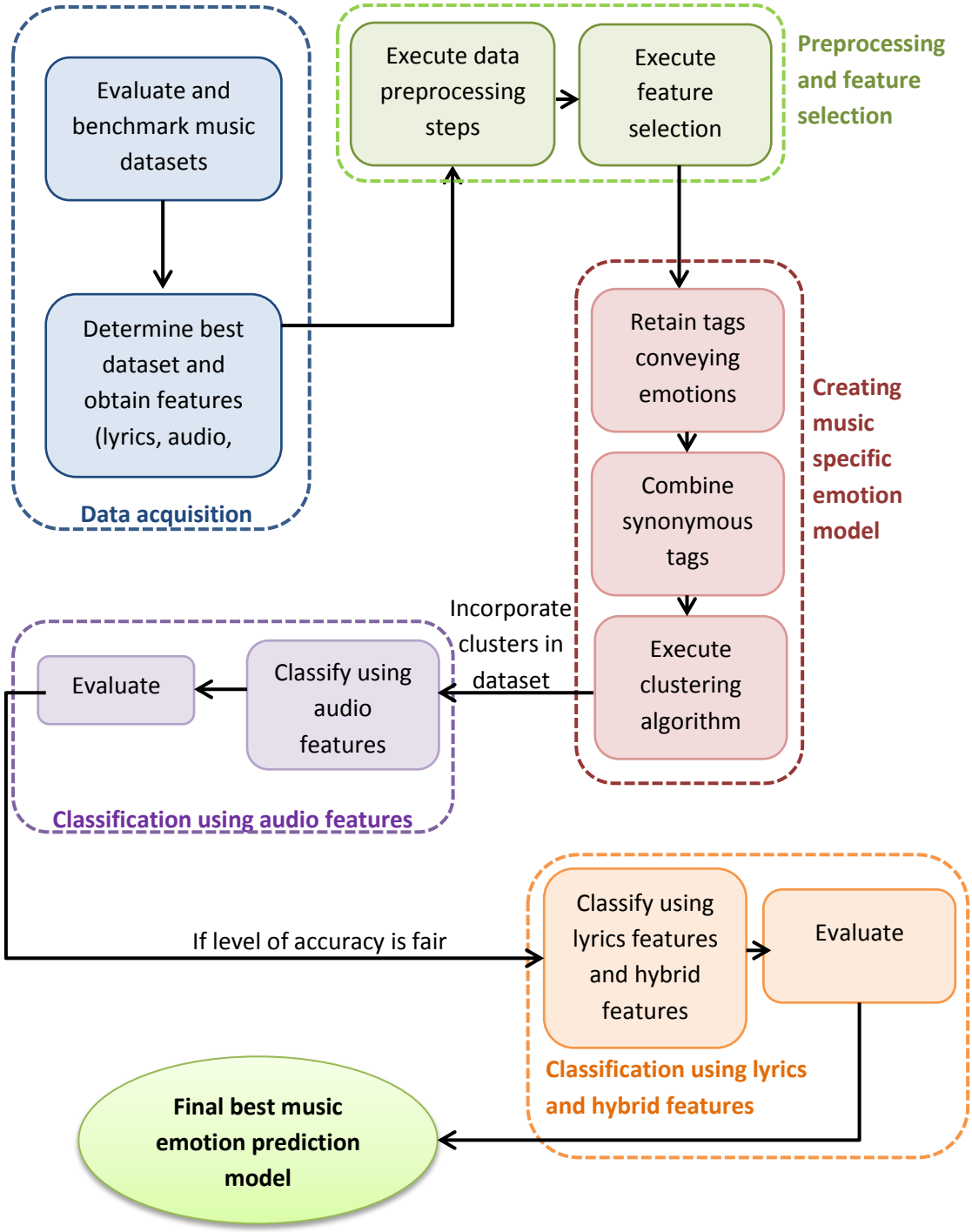

Figure 3. 1: Overall architecture of music emotion prediction model

As depicted in Figure 3.1, the initial phase of the research is dedicated to benchmarking and evaluating several music datasets, to evaluate their aptitude for achieving the research objectives. Each dataset is benchmarked with relation to relevance, quantity and quality subsequent to which the most suitable dataset to utilize for research is determined.

Preprocessing plays a vital role in refining and cleaning the dataset, such that the eventual outcome is undeterred by outliers/noise, etc. Furthermore, this stage is inclusive of steps which helps organize the dataset in a manner exploitable by automated methods. Feature selection algorithms allow retention of the most valuable features for modelling eventual prediction system. The efficiency of the classifier is highly reliant on application of proper feature selection.

The third phase addresses a primary research question of this study. Since importance of music specificity of the emotion model was emphasized, this component plays a significant role in the overall architecture. Most Music APIs allow individuals to tag the songs they listen to, thus enabling us to retrieve collection of tags associated with songs in our dataset. Some of these tags may semantically express emotions. Such tags would be retained, subsequent to which the synonymous tags would be combined. Eventually a clustering algorithm would be executed to further reduce the dimensionality of this emotion space. Since we possess no prior knowledge as to what a true emotion specific music model is, it's necessary to adopt an unsupervised learning technique in this stage. The outcome of this phase would be a set of emotion clusters, where each cluster comprises of set of emotion related tags. Since the original tags associated with songs in dataset are known, we can incorporate emotions clusters in the dataset prior to proceeding with research. The next two phases concern determining the final music emotion prediction model. Classification is initially attempted only on audio data since audio based classification generally outperforms lyric based classification (refer section 2.3.3). Subsequent chapters further elaborate execution of methodology.

3.5. Summary

This chapter was dedicated to providing an analysis of the proposed methodology and overall architecture (depicted by Figure 3.1) of the research. Justification was provided for assuming certain measures in the methodology whereas key concerns were highlighted. The setup for deploying this architecture would be extensively elaborated under next chapter.

Chapter 4 - Experimental Setup

4.1. Introduction

The former chapter extensively elaborated on the proposed methodology, according to which the research would be guided. The five components identified in the architecture comprise of the primary phases of the research, which would ultimately lead to a solution for the research problem. Each of these phases would be covered in subsequent chapters of the thesis.

This chapter is dedicated to specifying the experimental setup utilized to proceed with the research according to specified research methodology. The initial phase of the architecture is covered under section 4.2, whereas methods, tools, evaluation metrics and other resources requisite for completion of subsequent phases are introduced. Experiments would be conducted using freely available tools specified herewith, whereas self-implemented methods too would be incorporated.

4.2. Dataset

Being a data oriented research, the accuracy, impartiality and volume of the dataset on which the research is based, contributes greatly towards the success of the overall outcome. Thus, selecting a dataset satisfying these requirements and appropriate for the research objective is vital for ensuring acceptability of eventual conclusion. The initial option available was to create an in-house dataset of music. Subsequent to selection of a group of people, emotions associated with each music piece could have been derived via requesting listeners to convey what emotions they perceived, after listening to a song. However, retrieving this information from a significant number (thousands) of people would not have been feasible within provided timeframe. Determining emotions expressed by a song via questioning selected number of people may bias the eventual result to their perception. Thus the option of forming an in-house dataset was ruled out.

Therefore, basing the research on an existent, freely available music dataset was opted for. Prior to randomly selecting a specific dataset, several popular music datasets freely available for research work were benchmarked. These datasets were evaluated with relation to relevance, quantity and quality as specified in [42].

4.2.1. Dataset Evaluation Criteria

Relevance of data concerns the availability of necessary information required to conduct the research [42]. For this scenario, the necessary information comprises of structural features (lyrics, audio) and emotions associated with each music piece in dataset. Since emotions are to be derived using tags as specified under section 3.4, availability of tags for songs in dataset would be adequate to determine the emotions.

Quantity of data is primarily evaluated with respect to number of instances available for the research. Instances relate to the data points or the number of records available for the study. While having more than 5000 records is desired according to rule of thumb [42], conference paper [43] has proven empirically that classifiers achieve stability when training dataset exceeds 16 000 data points.

Quality of data is evaluated based on accuracy, consistency, believability and accessibility [42] in this research.

4.2.2. Dataset Benchmark

Six popular music datasets were benchmarked based on specified evaluation criteria.

RWC Database

Description:	*Evaluation:*
Comprises of four smaller databases; Popular Music Database (100 pieces), Royalty-Free Music Database (15 pieces), Classical Music Database (50 pieces), and Jazz Music Database (50 pieces). Each database comprises of audio, lyrics and MIDI files [44].	The magnitude of the dataset is not adequate for the research whereas Classical and Jazz DBs cannot be used due to being biased to a single genre. 80% of the Popular music database and 33.3% of Royalty-free music database comprise of Japanese songs [44]. Hence those too are much partial while being difficult to be explored and analyzed due to lack of familiarity with language.

GTZAN Genre Collection

Description:	*Evaluation:*
This dataset is created by George Tzanetakis; the designer of the infamous music analysis tool Marsyas. It has been utilized in researches [45] and [46] and comprises of 10 genre classes, with 100 thirty second long song excerpts per each class [46].	Magnitude of the dataset is not satisfactory. Furthermore, lyrics for all songs in the dataset are not acquirable due to copyright laws, whereas tagging of each song must be conducted with the assistance of fellow students. Thus utilizing this dataset is not very appropriate for the research.

MagnaTagATune

Description:	*Evaluation:*
This is a fairly large dataset utilized in a number of researches [47], [48], [49] comprising of 25 863 thirty second song excerpts. The music from Magnatune label have been associated with user tags acquired from TagATune game [50].	Being a considerably large dataset on which many former researches have been based, this dataset is somewhat appropriate for realizing objectives of this project. Furthermore, the association of songs with user tags acquired based on a game ensures that a considerable number of persons' emotion opinion is conveyed by this dataset. Retrieval of audio features utilizing an open source tool too is possible since audio files are made available. However lyrics features are not made available in the dataset itself. Thus, if this dataset is to be utilized, lyrics associated with dataset would have to be acquired separately.

Uspop2002

Description:	*Evaluation:*
This dataset comprises of tags, genre and audio features for 8752 music pieces [51]obtained from around 700 albums [38]. The dataset has been used for several researches [38], [52].	Despite the dataset comprising of tags and audio features, its magnitude is much smaller in comparison with other existent music datasets

Musicbrainz

Description:	*Evaluation:*
This is an invaluable dataset for researches which are based on metadata of music. Valuable data related to albums, artists, CDs could be derived from this dataset [53] whereas tags and ratings too are provided as supplementary information.	This dataset does not provide with audio files or lyric features required for research. Obtaining audio files for thousands of songs is not feasible considering the duration of research.

Million Song Dataset (MSD)

Description:	*Evaluation:*
MSD is a freely available dataset which provides with audio features and metadata for one million contemporary music tracks (i.e. songs) [51]. The primary source of this data is The Echo Nest[8] which is a music intelligence and data platform for developers and media companies. The API provided by Echo Nest has been utilized for acquisition of data associated with a million songs. The Laboratory for the Recognition and Organization of Speech and Audio (LabROSA[9]) has collaboratively formed this dataset with Thee Echo Nest, whereas it has been partially supported by the National Science Foundation [54]. Each music composition in the dataset is identified by the Track ID, which is a unique ID comprising of 18 characters. Due to the extensive application of MSD in music research domain [38] [39], several communities have contributed to providing features associated with the music pieces in this dataset. This is inclusive of lyrics, audio and tags, which are the features necessary to conduct this research.	MSD is presently the largest music dataset available for researches [51]. Appropriateness of MSD for this research is superior to the rest due to its magnitude and availability of all attributes required to conduct the study.

[8] http://the.echonest.com/
[9] http://labrosa.ee.columbia.edu/

4.2.3. Data Sources

"There is no data like more data" - Bob Mercer of IBM in 1985 [51]

Considering the properties of the freely available, popular datasets for music information retrieval tasks, MSD was chosen as the best dataset on which to base the research. The features necessary to conduct the research were obtained from three different sources as presented graphically in Figure 4.1.

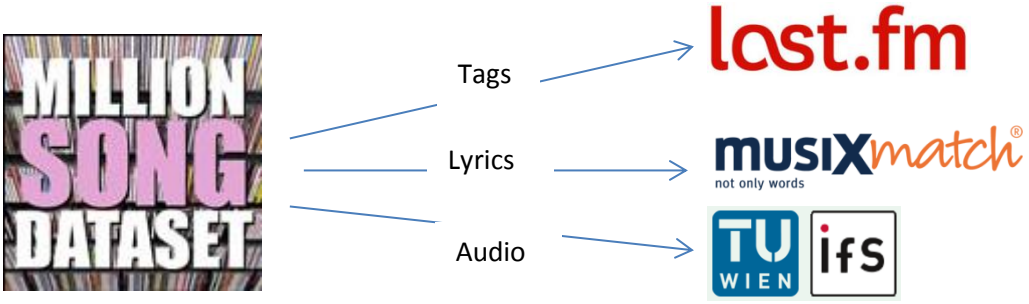

Figure 4. 1: Data sources for obtaining tags, lyrics and audio features

Last.fm Dataset[10]

This dataset provides with song level tags, which were utilized in the research for determining emotions associated with each song. Tags are 'terms' which song listeners have associated with the music pieces in the MSD, via the API provided by Last.fm[11]. Due to this being a site used by millions of people around the world to satisfy their musical requirements, tags appearing in this dataset reflect the opinions of a global community, thus making it a suitable resource based on which to determine emotions perceived by music. Despite the MSD dataset providing with its own tag collection, those are not apt to achieve the research objective since they reflect artist level tagging (terms used to describe an artist)[51]. Thus the tag features utilized for research were those provided in the Last.fm dataset.

[10] http://labrosa.ee.columbia.edu/millionsong/lastfm
[11] http://www.last.fm/

The tag dataset has been provided as a SQLite database which could be downloaded via their site. Track IDs from MSD have been associated with list of tags in this dataset, thus allowing us to retrieve the tags associated with a specific track via SQLite queries. There exist 505,216 tracks with at least one tag, whereas 522,366 unique tags are existent. The total number of track-tag pairs is 8,598,630.

musiXmatch Dataset[12]

This dataset provides with lyrics for 237,662 MSD tracks in bag-of-words format. Lyrics have been obtained from the musiXmatch site[13] which is presently the largest lyrics catalogue in the world. Certain tracks from MSD have been omitted due to numerous copyright restrictions and some being instrumental music. The musiXmatch dataset is the largest clean lyrics collection available for research. The bag-of-word representation adopts stemming to map several similar forms of a word to one stemmed-term. Mapping of stemmed terms to un-stemmed terms has been provided in the site to enhance comprehension.

The musiXmatch dataset too has been provided as a SQLite database, which reflects whether a certain word appears in lyrics of a song. Number of occurrences of that word in the song lyrics too is provided.

[12] http://labrosa.ee.columbia.edu/millionsong/musixmatch
[13] https://www.musixmatch.com/

The Information Management and Preservation Lab at the Department of Software Technology and Interactive Systems at Vienna University of Technology[14]

The Vienna University of Technology has provided with a multitude of audio features for 994 960 tracks in the MSD. Despite the MSD dataset providing with audio analysis of tracks as depicted in Figure 4.2, it was noted that this was not inclusive of significant low-level audio representations as spectral centroid, spectral flux, etc. Thus, it was deduced that utilizing audio features provided by Vienna University for research would be more appropriate.

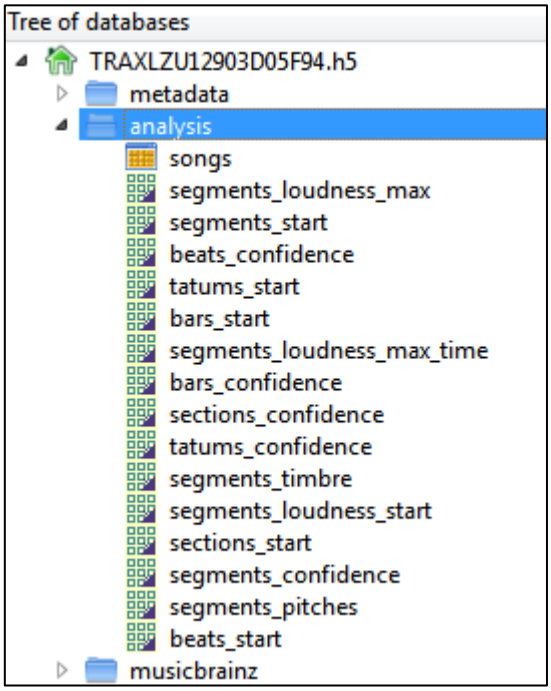

Figure 4. 2: Audio features provided in MSD

Of the many audio features provided in the dataset, the ones derived using the infamous audio analysis tools; jMIR[15] and Marsyas[16] were retained for the study due to their extensive usage for research work [15] [29] [54]. 16 spectral features and 24 timbre features retrieved from jMIR (jAudio component) and Marsyas respectively were retained for study. Table 4.1 provides a concise description regarding audio features utilized. Description of jMIR features were obtained from [55].

[14] http://www.ifs.tuwien.ac.at/mir/msd/download.html
[15] http://jmir.sourceforge.net/
[16] http://marsyas.info/

jMIR	
Feature	**Description**
Spectral centroid	The center of mass of the power spectrum.
Spectral roll-off point	The fraction of bins in the power spectrum at which 85% of the power is at lower frequencies. This is a measure the right-skewedness of the power spectrum.
Spectral flux	A measure of the amount of spectral change in a signal. Found by calculating the change in the magnitude spectrum from frame to frame.
Compactness	A measure of the noisiness of a recording. Found by comparing the components of a window's magnitude spectrum with the magnitude spectrum of its neighboring windows.
Spectral variability	The standard deviation of the magnitude spectrum. A measure of how varied the magnitude spectrum of a signal is.
Root mean square	A measure of the power of a signal over a window.
Zero crossings	The number of times the waveform changed sign in a window. An indication of frequency as well as noisiness.
Fraction of low energy windows	The fraction of the last 100 windows that has an RMS less than the mean RMS of the last 100 windows.
Marsyas	
Measurement on the Fast Fourier Transformation (FFT) of sounds generated by octave notes. A keyboard octave has notes: A, A#, B, C, C#, D, D#, E, F, F#, G, G# (12 separate features)	

Table 4. 1: Description of audio features utilized for research

Table 4.1 provides with description of audio features retained for research task. MFCC features have been disregarded since it's primarily of importance for automatic speech and speaker recognition [56], thus being of less relevance for this study. Both mean and standard deviation values of each feature specified in Table 4.1 have been used for the research. Hence, 40 audio features have been retained, where 16 (8*2) have been extracted from jMIR tool and 24 (12*2) from Marsyas. Audio features were provided in ARFF file format.

4.3. Tools and Programming Languages

This section provides a brief description regarding tools and programming languages used to conduct research experiments. *Python* language was extensively used for self-implementation of certain methods to simplify tasks and to execute data-preprocessing. Several data analysis tools were used to perform numerous research tasks effectively. These comprise of *R, Weka, Meka* and *RapidMiner.* Furthermore, *WordNet* tool was used for determining emotion related tags and for identification of synonyms.

4.3.1. Python

This is an extensively used high level programming language with a powerful syntax that allows users to express concepts in few lines of code. This is a creation of Guido van Rossum from Netherlands, who commenced implementation of the language in 1989. Python 2.7 version was used in this research, which was installed on both the Windows and Ubuntu platforms.

4.3.2. R Tool

R [57] is a free software environment which is widely used for statistical computing and graphics. This valuable tool for data analysis is a creation of Ross Ihaka and Robert Gentleman from University of Auckland.

4.3.3. Weka

Waikato Environment for Knowledge Analysis (Weka) [58] is an invaluable tool for data related research. This software has been developed at University of Waikato, New Zealand, whereas the language on which it's based is java. Weka supports classification of single-label (multi-class and binary-class) data only.

4.3.4. Meka

Meka is a multi-label extension to the Weka tool which allows classification of multi-label datasets. This tool provides with a wrapper to Mulan framework which has been built to enable classification of multi-label data. Thus, Meka combines the power of Mulan and Weka to provide with a comprehensive environment for multi-label classification.

4.3.5. RapidMiner

RapidMiner is a software platform developed by the company of the same name. It is extensively used for data mining tasks and research work. The starter edition which is available for free download has been used in this research.

4.3.6. WordNet Tool

WordNet [13] is a large lexical database of English. This tool groups words according to their semantic meaning, thus making it a useful instrument for computational linguistics and natural language processing. This tool is important for the research since emotional significance of tags is necessary to be determined. It further assists in combining synonymous terms.

4.4. Evaluation Procedure

Stratified k-fold cross validation [59] is a common and efficient evaluation procedure for classification tasks. This algorithm is applicable for both single-label and multi-label classification, thus being implemented in the tools Weka and Meka. A concise overview of the algorithm is as follows.

Prior to performing k-fold cross validation, the dataset is randomly split into k number of partitions, such that each partition is approximately of equal size. Subsequent to partitioning, the classification algorithm is executed on this space k number of times, using (k-1) subsets for training the classifier, and the remaining subset for validation. A different subset is used for the validation process in each round thus ensuring utilization of a partition for validation exactly once. Eventually, the average performance measure would be output as final result. Stratified k-fold cross validation is a variant of the original algorithm, where the algorithm strives to ensure that each class is represented equally (approximately) across each fold. Despite this being somewhat computationally expensive, it helps minimize the overfitting[17] problem which many classifiers suffer from. Stratified 10-fold cross validation was applied in each classification experiment conducted in this research. Figure 4.3[18] depicts 5-fold cross validation where 'red' and 'green' depict target class labels. In stratified 5-fold cross validation, each of these folds would contain approximately the same number of reds and greens.

[17] *Overfitting:* This is a modelling error which reflects the poor performance of classifier model with relation to predicting class-label of unseen data, due to adapting 'too well' to training dataset. Presence of error and noise is often the cause of this issue, which could be minimized by application of validation techniques such as k-fold algorithm. Apart from this, pre-pruning and post-pruning are renowned techniques applied in decision tree classifiers for reducing effect of overfitting.

[18] https://www.kaggle.com/c/chess/forums/t/112/alternatives-to-month-aggregated-rmse

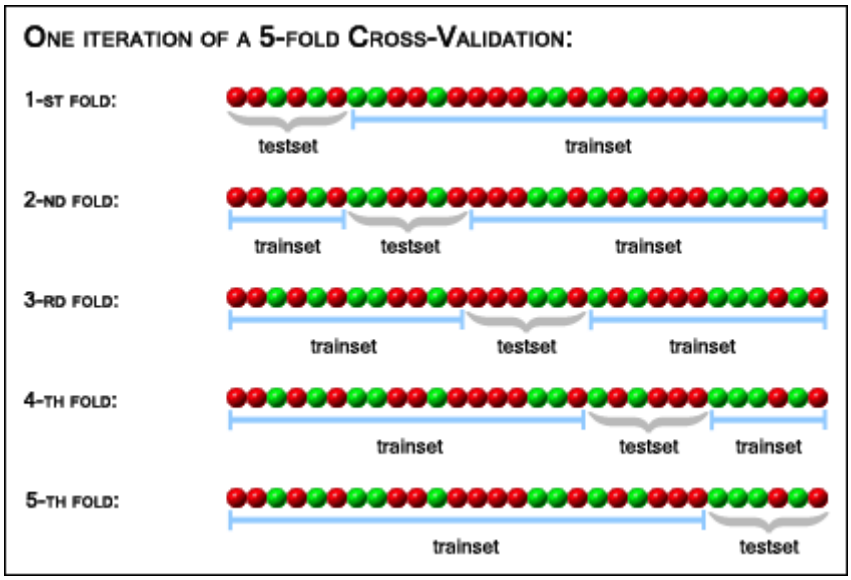

Figure 4. 3: 5-fold cross validation

4.5. Evaluation Metrics

Evaluation metrics help evaluate performance of classifiers assisting researchers to determine what the best classification model for prediction task is. Different metrics present different viewpoints of classifier aptitude, thus necessitating usage of several metrics to acquire a holistic view of overall performance of the model. Certain metrics utilized in the research are applicable for both single-label and multi-label classifier evaluation as specified under this section.

4.5.1. Individual Evaluation Measures

Individual evaluation measures are direct interpretations derived from the confusion matrix. In a binary-classification scenario, the confusion matrix would reflect how positive and negative labels (e.g. assuming only two emotions were considered, positive label could be 'happy' whereas negative label could be 'sad') have been classified. The confusion matrix is easily extensible to handle multi-class and multi-label scenarios, where instead of positive and negative labels a series of labels would be introduced. Figure 4.4[19] depicts confusion matrix for binary classification.

[19] https://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/

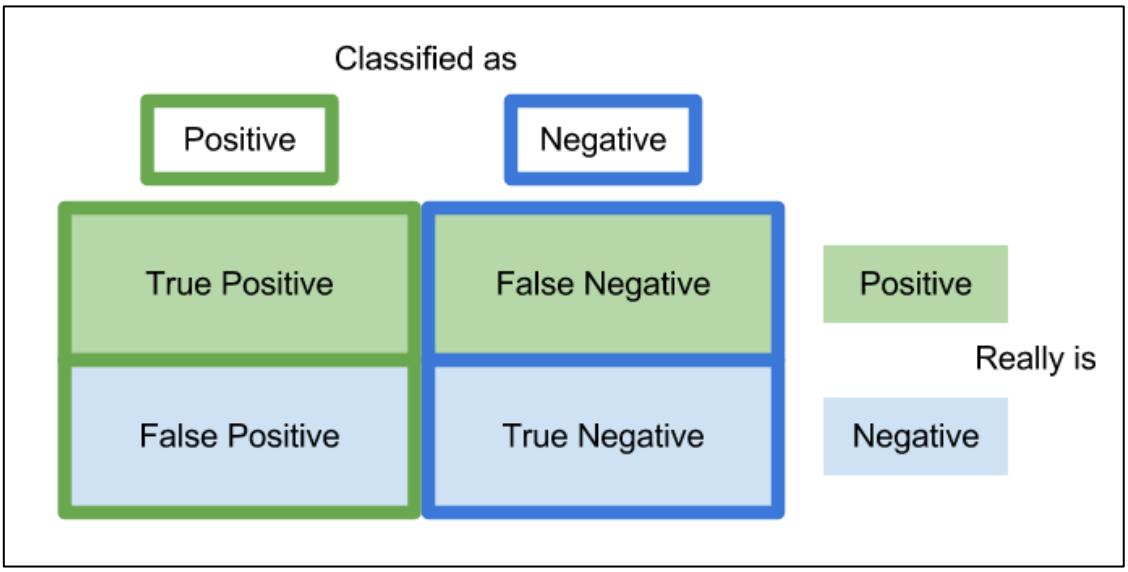

Figure 4. 4: Confusion matrix

Individual evaluation measures depicted are based on Figure 4.4, using following abbreviations.

TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative

All these measures are applicable in single-label classification scenario, whereas those applicable for multi-label evaluation have been specified.

Accuracy

This measure evaluates the proportion of correct predictions by a classifier [60] [61]. Accuracy is a measure applicable in multi-label scenario as well.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision

Also known as positive predictive value, this helps evaluate the probability that a positive prediction is correct [61]. This metric could be used to evaluate multi-label classifiers as well.

$$\text{Precision} = TP/ (TP + FP)$$

Recall

Also known as true positive rate, sensitivity and hit rate, this helps evaluate the probability of correctly labeling members of the target class [61].

$$\text{Recall} = TP/ (TP + FN)$$

False Positive Rate

Otherwise known as false alarm rate, this reflects the probability of falsely rejecting the null hypothesis for a particular test (i.e. classifier inaccurately states that an instance belonging to negative class is positive) [61].

$$\text{False Positive Rate} = FP/ (FP + TN)$$

4.5.2. Combined Evaluation Measures

F-measure

This is the harmonic mean of precision and recall [61]. F-measure is a sound evaluation metric for both single-label and multi-label classifiers.

$$\text{F-measure} = 2 * \text{Precision} * \text{Recall}/ (\text{Precision} + \text{Recall})$$

Hamming Loss

Hamming loss [60] is a measure which reflects number of times an incorrect label is predicted and number of times a relevant label is not predicted. XOR operation is applied to evaluate symmetric difference of two sets. This measure is a valuable metric to evaluate multi-label classifiers.

$$\text{Hamming loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i \,\Delta\, y_i)}{l}$$

n= number of instances, l = number of labels, x_i = prediction, y_i = ground truth (actual label).

4.5.3. Graphical Evaluation Measures

Area under the Curve (AUC)

AUC [61] is a graphical measure which depicts the area under ROC curve. ROC curve represents true positive rate against false positive rate, thus providing with an indication regarding overall performance of classifier. While value of AUC lies between 0.5 and 1, Table 4.2 depicts the inference one can make of classifier performance based on AUC value.

Area under the Curve	**Reflection on performance**
0.9 - 1	Excellent
0.8 – 0.9	Good
0.7 – 0.8	Fair
0.6 – 0.7	Poor
0.5 – 0.6	Fail

Table 4. 2: Classifier evaluation based on AUC value

4.6. Algorithms Explored

Algorithms which were explored in the research are described under this section. The unsupervised learning technique attempted in research for initial clustering is the agglomerative hierarchical clustering approach. Different algorithms are applicable in the multi-label and single-label scenarios, thus allowing us to categorize them under two sections. Utilized multi–label classifiers are those which were encountered in the literature. SVM classifier has been extensively used for single-label classification in former research attempts. Apart from SVM several other classification approaches too have been attempted in this study for the single-label classification scenario.

4.6.1. Hierarchical Clustering

In agglomerative hierarchical clustering approach, each data point (tags in this instance) would be in separate clusters at the beginning, which would be successively merged to form larger clusters. The merging could be halted either when all data points are in one cluster or when a certain stopping criterion is met [42].

Merging of two clusters in agglomerative hierarchical clustering is based on the linkage type opted for. Of the two unsupervised learning techniques identified in literature, hierarchical clustering was chosen, since it allows the flexibility of choosing which tags to merge via comparison of clusters formed for different linkage types. This is disallowed in k-means clustering, since it specifically assigns each tag to one of the clusters, thus prohibiting a tag to be a cluster of its own (cluster with one tag).

A distance measure is concerned with hierarchical clustering regardless of the linkage method used. *Jaccard Distance* was chosen since it could be used to reflect the proportion of songs a tag has in common.

Jaccard Distance

This is a distance measure adopted in clustering, when objects concerned have no 'location' which is numerically presentable. Jaccard distance [62] is applicable when objects could be represented as sets. For two sets A and B,

$$jaccard\ distance = 1 - \frac{|A\ \cap B|}{|A\ \cup B|}$$

Objects with minimal Jaccard distance are likely to be in same cluster. When songs associated with tags are represented as sets, this distance measure could be effectively applied.

As mentioned afore, different linkage types could be applied to determine how clusters must be formed using the distance measure. Five distinct linkage types described herewith were applied in this research.

Single Linkage (Nearest Neighbor)

Definition of the distance between two clusters in this methodology is the shortest distance between them; that is the distance between the points closest to each other in two clusters. The clusters which possess the minimal among these shortest distances are chosen as the ones to be merged [42] [63]. In Figure 4.5, for shortest distances between each pair of clusters, if q < r < s, clusters A and B would be combined.

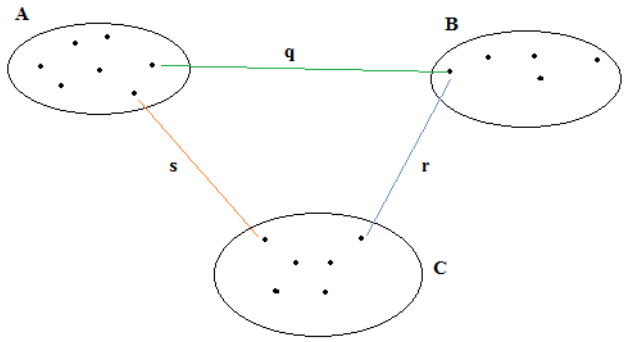

Figure 4. 5: Single linkage

Complete Linkage (Furthest Neighbor)

Distance between the two data points furthest from each other is taken into consideration in defining the distance between two clusters. The clusters whose such defined distance is minimal are then merged to form a larger cluster [42] [63]. In Figure 4.6, for longest distances between each pair of clusters, if q < r < s, clusters A and B would be combined.

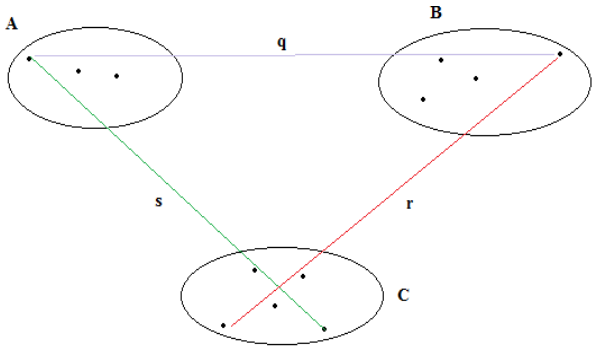

Figure 4. 6: Complete linkage

Group Average (Average Link Clustering)

This method takes the average of distance between each data point in one cluster against each data point in the other. Hence it considers all dissimilarities, helping us avoid the pit falls of single link and complete link criteria. After acquiring average of dissimilarities between each two clusters, the ones with the minimum average dissimilarity are merged [42] [63]. Figure 4.7 depicts group average clustering where

$$\text{Average of dissimilarities} = \frac{\sum_i \sum_j d(i,j)}{i * j}$$

{dissimilarity d (i,j), iϵA, jϵB}

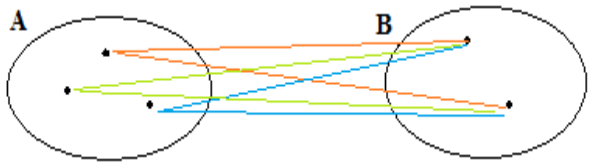

Figure 4. 7: Average link clustering

Ward's Method [64]

This linkage method is adopted in research [35] due to it being considered highly efficient [63]. In each phase of clustering, this mechanism attempts to minimize the error sum of squares of formed clusters [63]. Thus the intra cluster variance is minimized in this method, which has led to this algorithm being referred to as Ward's minimum variance method [64]. **Ward1** and **Ward2** are methods based on this concept which slightly vary according to implementation. Both methods were attempted in this research.

4.6.2. Single-label Classification

Support Vector Machines (SVM)

SVM[65] is an extensively used supervised learning technique which attempts to construct optimum hyperplanes which best separate a dataset into classes. A good separation is said to be obtained by finding the hyperplane which has largest distance to nearest training record from any class. Figure 4.8 depicts optimal hyperplane as a solid line, where training data belongs to two distinct classes depicted by crosses and circles.

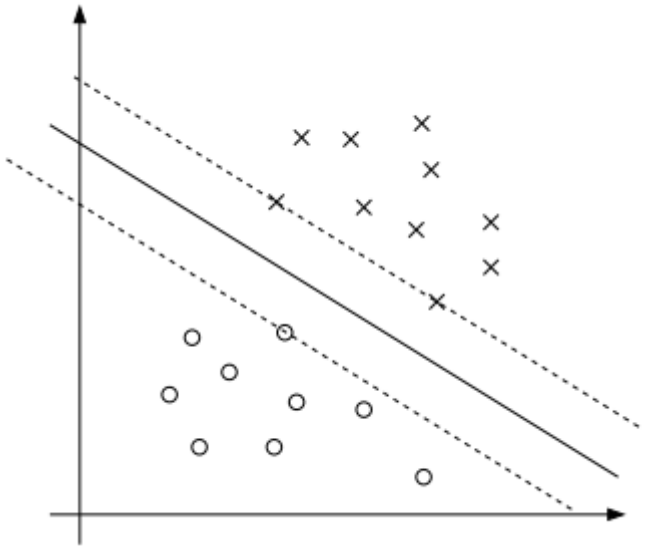

Figure 4. 8: Optimal hyperplane in SVM

Unlike the example depicted in Figure 4.8, training data is usually not linearly separable. Thus, prior to conducting separation, the data is mapped to a higher dimensional space. Kernel functions are utilized to simplify this mapping process, whereas the best kernel for a specific dataset must be learnt empirically. The Sequential Minimal Optimization (SMO) [65] implementation of SVM is made available in Weka and Meka tools, which would be utilized in this research.

Naïve Bayes

Naïve Bayes [66] is based on Bayes' Theorem which describes the probability of an event, based on conditions that might be related to the event. Bayes' theorem is described as follows.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

P(A) and P(B) are the probabilities of A and B without regard to each other. P(A|B), a conditional probability, is the probability of observing event A given that B is true. P(B|A), is the probability of observing event B given that A is true.

Assume there's a problem instance x(a1, a2, …, an)to be classified where the class attribute C has k outcomes. Attribute values are represented by a1, a2, …, an where n unique attributes exist. Naïve Bayes finds probability of x belonging to each class category C_k given the attribute values (a1, a2, …, an). That is, P(C_k | a1, a2, …, an)is calculated for all k. Class value C_k for which we get highest probability is chosen as class label of x.

P(C_k|a1, a2, …, an) = P(C_k|x) could be depicted as follows using Bayes' theorem.

$$P(C_K|x) = \frac{P(C_K)P(x|C_K)}{P(x)}$$

Using assumption P(C_k|x) = P(C_k)P(x|C_k) and the conditional independence assumption P(x|C_k) = P(a1, a2, …, an |C_k) = P(a1|C_k). P(a2 |C_k)… P(x|C_k), the final output of Naïve Bayes algorithm could be depicted as,

$$predicted\ class = \max_{k\in\{1,\ldots,k\}} P(C_K)\prod_{i=1}^{n} P(x_i|C_K)$$

Naïve Bayes implementation in Weka utilizes a kernel function for handling continuous attribute values.

C4.5

C4.5 [67] is a decision tree algorithm which is a variant of the original ID3 algorithm devised by Ross Quinlan. Information gain is the function used in this method to determine the attribute to be chosen at each level of the tree. Attribute with highest gain is chosen as root whereas the next attributes are chosen in descending order of their gains. Higher information gain reflects the capability of an attribute to split the data such that each partition comprises of higher proportion of data from a single target class only. This algorithm is implemented as J48 in Weka tool.

Random Forests

Random forests algorithm [68] forms a combination of prediction trees and produces the eventual result using ensemble methods. Thus the final class label is predicted via studying the class label prediction of each tree. Each tree is constructed using a sample of the training dataset, where each sample is created with replacement, adhering to bootstrapping mechanism. The overfitting problem of general decision trees is minimized by this classifier. Among many desirable characteristics of this algorithm are its robustness to noise and outliers. Furthermore the accuracy of Random forests tends to be either as good as Adaboost (a boosting mechanism) or in certain instances better. Figure 4.9[20] provides an overview of this classifier.

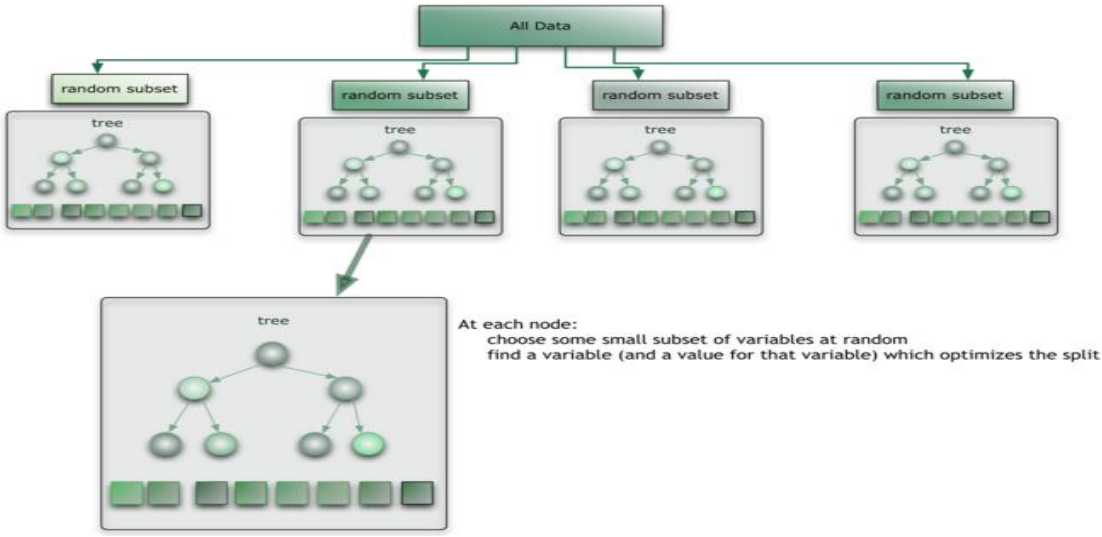

Figure 4. 9: Random forest

[20] https://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics/

4.6.3. Multi-label Classification

Binary Relevance (BR)

BR[60] is a popular problem transformation method used to solve multi-label classification problem, where the dataset is transformed to handle the problem as a series of binary classifiers. Union of labels predicted as positive is output for new element classified. For instance, in this research, each song would be associated with a number of emotions when it's depicted as a multi-label problem. BR creates several datasets, associating a song with a single emotion class in each case. This allows each new dataset to be dealt with as a binary classification scenario. Eventually the classifier outcomes would be combined to decide the emotions of unseen songs.

Label Power-set (LP)

Despite BR classifiers being widely used, it doesn't take label dependency into account, which has been noted as a major limitation of BR. This inadequacy is resolved in LP [60] where set of labels associated with each data point is considered as a unique class. For instance, if emotions e1, e3, and e7 are associated with a particular song, LP would define a new class label as e137. The problem is subsequently addressed as a single-label (multi-class) classification instance, where the most probable class of a new data point is output.

Multi Label K Nearest Neighbor (MLKNN)

MLKNN [60] is an extended version of the k nearest neighbor algorithm adopted for single-label classification. Initially the k nearest neighbors of a data point in training set are identified, subsequent to which maximum a posteriori probability is used for classifying new data point.

Random K Label-sets (RAKEL)

Class imbalance and incapability of predicting unseen labels is a limitation of the LP method. This problem is resolved in RAKEL [60] by partitioning collection of all labels to k subsets. LP algorithm is then executed on each of these subsets separately and averages the decisions per label, when novel data requires to be classified.

Since each of these classifiers eventually address the problem as a multitude of single-label problems, it's necessary to incorporate a single-label classifier as well in each instance. SVM was chosen as the base classifier for this research task as specified in the literature.

4.6.4. Parameter Tuning

Among the algorithms explored in this study, SVM, Random Forest and C4.5 have parameters associated with them, which need to be tuned for achieving optimal performance of each classifier. *Two-step-evaluation* was assumed in our research for deducing these optimal parameters. Figure 4.10 provides with an overview of this evaluation procedure.

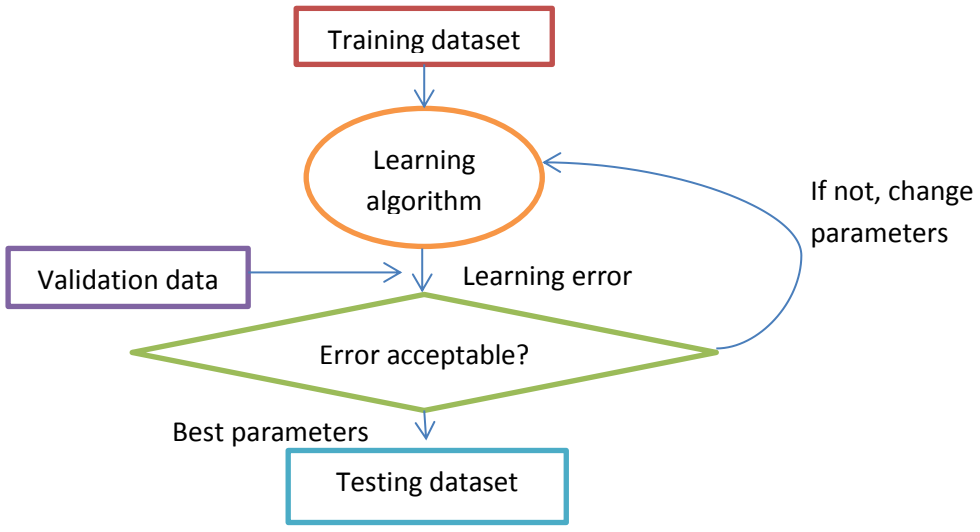

Figure 4. 10: Two step evaluation for parameter configuration

When conducting two-step-evaluation, 50% of the dataset was utilized for training whereas 25% each were used for validation and testing of the classifiers. Different metrics could be used as learning error depending on the classification approach. While Root Mean Squared Error (RMSE) was used as learning error of all three classifiers; SVM, Random Forest and C4.5; the additional learning error measure known as Out of bag error (OOBE) too was considered for Random Forest algorithm.

4.7. Handling Class Imbalance

Class imbalance phenomenon is encountered when the training dataset is not equally distributed among the target class values. In this context, unequal distribution is resulted due to certain emotions being frequently expressed by songs in training set and other emotions being infrequent. When imbalance is present in the training set, classifiers are prone to execute learning on majority label only, thus resulting in classification of unseen data into this majority class often [69]. As a result, the minority class(es) is entirely disregarded, yielding poor accuracy for that class. Generic methods [69] (not specific to one classifier) of minimizing the effect of class imbalance are depicted in Figure 4.11.

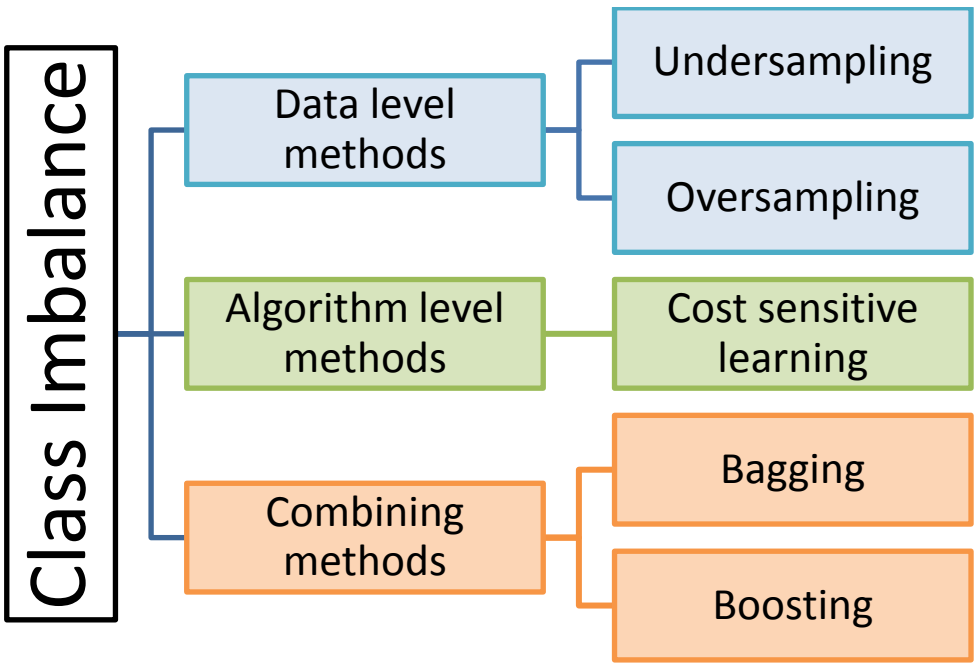

Figure 4. 11: Methods of handling class imbalance

4.7.1. Data Level Methods

Undersampling

This technique strives to resolve class imbalance problem via randomly eliminating records from majority class [69]. Elimination is often conducted to achieve 1:1 ratio (approximately) among class labels. A drawback of this mechanism is loss of important information from the test dataset. Furthermore, in an instance where the minority class comprises of very few data points, adopting this method would be highly detrimental to the eventual prediction, since classifier model would be formed using very few records. *SpreadSubsample* implementation of Weka was used as undersampling algorithm in this study.

Oversampling

Class imbalance is resolved via random replication of records from minority class [69]. Increased tendency for classifiers to over fit training data is viewed as a limitation of this method. Weka offers the *SMOTE* implementation of this algorithm, which replicates data via generating new synthetic records for minority class.

Figure 4.12[21] provides a graphical interpretation of undersampling and oversampling techniques.

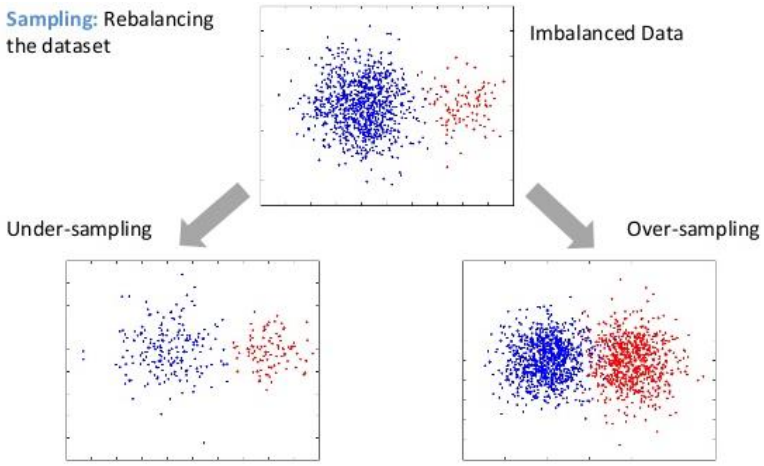

Figure 4. 12: Undersampling and oversampling techniques

[21] http://www.slideshare.net/barnandas/barnan-das-icost2011talk

4.7.2. Algorithm Level Methods

Cost Sensitive Learning

Cost sensitive learning [69] [70] helps minimize the effect of class imbalance via specifying that one misclassification cost is greater than another misclassification. To achieve this objective, a misclassification cost of minority class is assigned a higher value than misclassification cost of majority class. With reference to Figure 4.4, if the positive class comprises of fewer records than negative class, misclassifying positives as negatives (False Negative) could be specified to be costlier than misclassifying negatives as positives (False Positive). Figure 4.13 depicts this graphically, where False Negative misclassification is indicated cost five times as much as False Positive misclassification. Accurate classification assumes cost value 0.

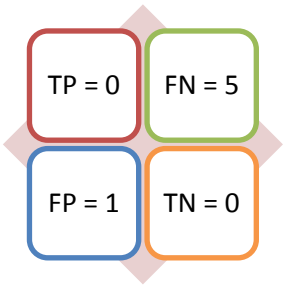

Figure 4. 13: Example cost matrix

Despite this being an effective method, assigning cost requires one to have prior knowledge regarding which misclassification error is worse. Such prior knowledge is lacking in the music emotion recognition domain, thus making it difficult to utilize this mechanism in our research. However, empirically evaluating the impact of different costs on accuracy of classifier is a feasible method of fine tuning the cost values.

4.7.3. Combining Methods

Combining methods help combine the classification of different subsamples of the dataset, thus minimizing the effect of class imbalance phenomena. *Bagging* and *Boosting* are frequently used classifier combination methods, which were attempted in this research as well.

Bagging

Bagging, otherwise known as Bootstrap Aggregating [69] creates a number of new training sets using the original dataset. These new training sets are created using sampling with replacement, subsequent to which an array of classifiers would be modeled on each set. Eventual prediction of a record is produced by averaging the output of each model or by plurality voting (i.e. outputs prediction of majority of models).

Boosting

Similar to Bagging [69], sampling with replacement is applied in Boosting as well, where a weighting mechanism enhances the possibility of formerly misclassified data to be selected in second round. Thus Boosting follows an iterative process which has to be performed sequentially. *AdaBoostM1* implementation of Boosting rendered by Weka was applied in this research.

4.8. Experimental Environments

Experiments were executed on three different environments depending on varying resource requirements and computational intensity of the algorithms. Algorithms requiring significant processing power were executed on the remote server environment depicted by Table 4.3 whereas the bulk of data preprocessing and feature selection tasks were executed on environments depicted by Table 4.4 and Table 4.5.

Measure	**Description**
Processor	Intel(R) Xeon(R) - 2.5 GHz, 2500Mhz, 16 Cores, 32 Logical processors
RAM	120 GB
Operating System	Microsoft Windows Server 2012 R2 Datacenter

Table 4. 3: Remote server environment

Measure	**Description**
Processor	Intel(R) Core(TM) i3 – 2.4GHz
RAM	4 GB
Operating System	Windows 7 Home Premium 64-bit

Table 4. 4: PC environment: Windows

Measure	**Description**
Processor	Intel(R) Core(TM) i3 – 2.4GHz
RAM	4 GB
Operating System	Ubuntu 11.04

Table 4. 5: PC environment: Ubuntu

4.9. Experimental Flow

This subchapter provides with a concise description regarding steps followed in conducting experiments. All experiments are based on data acquired from Million Song Dataset (MSD) and associated data sources described under section 4.2.3.

4.9.1. Data Acquisition Based on Music Dataset Evaluation

Prior to commencing the experiments it was necessary to obtain the dataset most suited for achieving research objectives. Thus evaluation of six freely available music datasets was performed as extensively described under section 4.2. Considering quantity, quality and relevance of the datasets, the MSD was opted for as the most appropriate dataset on which to base the research.

4.9.2. Preprocessing and Feature Selection

Data preprocessing tasks which are vital for derivation of reliable results were conducted under this phase. Noise/outlier removal, deduction of a mechanism of handling missing data, data integration, conversion of data files to standard format, elimination of redundant features, etc. were executed in the preprocessing stage. Several data files were constructed for the same set of songs as necessitated for varying clustering and classification tasks. Subsequently a subset of features was selected based on features redundancy analysis.

4.9.3. Constructing Music Specific Emotion Model

A music specific emotion model was constructed by retrieving emotion-related tags associated with music and combining them to form emotion clusters. Combination was performed based on synonymy of tags and using hierarchical clustering approach for five different linkage types, as described under section 4.6.1.

4.9.4. Training Models

Subsequent to determining final emotion clusters, those were incorporated in the dataset, prior to conducting classification. Classification was initially attempted on audio feature set only due to time constraint. The models which perform commendably for audio features were then attempted for lyrics features and hybrid features separately. Algorithms attempted in this phase were elaborated under sections 4.6.2 and 4.6.3.

4.9.5. Testing Models

The classification models were evaluated utilizing evaluation metrics specified under section 4.5. Research conclusion was made via critical analysis of the evaluation results of these classifiers.

4.10. Summary

This chapter was dedicated to providing insight to reader regarding how and why the Million Song Dataset was selected for conducting research. The rest of the experimental setup was extensively elaborated, specifying tools and programming languages used to execute experiments and the algorithms explored. Evaluation metrics described under this chapter were used for attesting aptitude of attempted classification algorithms. An introduction to the class imbalance problem was provided, specifying techniques adoptable for dealing with it.

The next chapter is dedicated to discussing the preprocessing and feature selection steps followed, prior to conducting research experiments.

Chapter 5 - Feature Engineering

5.1. Introduction

The former chapter provided detailed information regarding the experimental setup followed in the research project. This setup was applicable throughout the research procedure, whereas reader was enlightened on the experimental flow adhered to in attempting to achieve research objective.

This chapter focuses on the feature engineering tasks executed, which is inclusive of data preprocessing and feature selection phases. The accuracy of the eventual outcome significantly relies on the quality of the feature engineering process, thus placing considerable importance on this phase. Second component of the overall architecture, depicted by Figure 3.1, is entirely addressed under this chapter, whereas retrieval of emotion related tags associated with third phase too is elaborated herewith.

5.2. Data Preprocessing

Data preprocessing comprises of mandatory stages which must be executed to ensure that the dataset is apt for classification. Apart from the customary data preprocessing stages as data cleaning and data integration, this study required the retrieval of emotion related tags from acquired song level tags. As described in Figure 3.1, initial stage of emotion class deduction is retrieving tags conveying emotions. This was addressed under data preprocessing. The original dataset comprised of songs of various languages, from which merely English songs required to be retained. Subsequent to execution of all these steps, feature selection was conducted.

5.2.1. Creation of Initial Data Files

As mentioned under section 4.2.3, the features associated with songs from Million Song Dataset were obtained from three different sources. Thus the presentation of data was required to be converted to a common format, to proceed with experiments. Audio files obtained were originally in the ARFF (.arff extension) format whereas the lyrics and tag features were provided as database files (.db extension). Two file formats opted for were CSV (.csv) and ARFF (.arff). Tools RapidMiner and R support the former whereas Meka and Weka support the latter.

Python based self-implemented programs were utilized for file creation purposes. Subsequent to completion of initial file creation phase we were in possession of four files which are as follows.

I. An ARFF file comprising of audio features of music tracks. The original ARFF files were merged to represent the 40 audio features described in Table 4.1 as a single vector. Thus, a series of audio patterns for music pieces in the dataset was the content of this file.

II. The audio ARFF file was converted to an equivalent CSV file. Both file formats were retained to deal with different requirements of tools utilized. Figure 5.1 provides with a partial view of the CSV file created, when loaded on RapidMiner tool. Each row in this file corresponds to a song in dataset.

III. The initial lyric file was in ARFF format alone, where the attributes were 5000 most commonly appearing words in songs. This initial file merely reflected the presence or absence of a word with respect to a song. 1 was used to mark presence of a word in song lyrics whereas 0 represented absence of it. This file was subject to considerable further refinement in latter stages, which has been described later in this section.

IV. A third CSV file was created to represent tags, as a collection of song vectors. Each tag was replaced by a binary vector reflecting whether a particular song in the dataset was associated with the tag or not. 1 and 0 were used to mark presence or absence of a song. Figure 5.2 depicts partial view of this file when loaded on Microsoft Excel. A row corresponds to a tag and a column to a song.

Row No.	Spectral_C...	Spectral_R...	Spectral_Fl...	Compactne...	Spectral_Va...	Root_Mean...	Fraction_Of...	Zero_Cross...
1	6.229	0.053	0.001519	172.500	0.002	0.057	0.056	16.230
2	3.808	0.036	0.005712	244	0.003	0.087	0.074	11.810
3	10.600	0.081	0.01223	237.200	0.005	0.173	0.054	24.360
4	12.240	0.091	0.001293	215	0.001	0.054	0.053	27.790
5	12.040	0.108	0.005138	212.700	0.003	0.080	0.041	26.610
6	5.825	0.048	0.007174	204.100	0.003	0.094	0.056	14.330
7	9.872	0.085	0.00033	226.900	0.001	0.045	0.077	23.910
8	6.247	0.061	0.004911	179.200	0.003	0.088	0.054	16.940
9	6.603	0.051	0.004536	190.600	0.003	0.091	0.051	15.490
10	5.465	0.048	0.009556	195	0.003	0.094	0.059	16.410
11	5.738	0.053	0.002624	199.600	0.002	0.048	0.068	14.890

Figure 5. 1: CSV file for audio features

	A	B	C	D	E	F	G	H	I	J
1	0	0	0	1	0	1	0	0	0	0
2	0	0	0	0	1	1	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	1	0	0	0	1	1	1	0	0
10	0	1	0	0	0	0	0	0	0	0
11	1	0	0	0	1	1	1	0	0	0
12	1	0	0	1	0	0	0	1	0	0
13	0	1	0	1	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0

Figure 5. 2: Tag-track file

5.2.2. TF-IDF Based Lyric Attribute Values

The ARFF file initially created for lyrics features as described under section 5.2. 1 was extremely inefficient to work with, since it comprised of 5000 attributes. Furthermore, as specified under section 1.4, the research utilized English songs alone to conduct experiments. Thus, prior to refining the lyrics dataset via assignment of TF-IDF values, the non-English words in dataset were required to be identified for removing non-English songs.

The python module PyEnchant[22] was used to achieve the objective of identifying non-English words. Subsequent to removing attributes which corresponded to such non-English words, it was possible to discard the non-English music pieces as well, via analyzing the binary vector values of those songs (i.e. once non-English attributes are removed, those songs would correspond to a vector comprising only of 0 values). Subsequently the TF-IDF value assignment was executed.

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF weighting helps determine the importance of a term in a document, with relation to a collection of documents considered [33]. This extensively applied score in document classification context could be utilized in our research by considering songs as 'documents' and words from song lyrics as 'terms'. Rather than merely depicting the presence or absence of a word from song lyrics, using TF-IDF value allows reflecting how important a word is for a song. Calculation of TF-IDF score is as follows.

Term Frequency (TF) calculates the frequency with which a term occurs in a document [33].

$$TF = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{count\ of\ all\ words\ occuring\ in\ document\ d}$$

Inverse Document Frequency (IDF) of a term is defined as follows. It assigns a high value to those terms which rarely occur, whereas words commonly appearing in a number of documents get a low value.

$$IDF = \frac{total\ number\ of\ docuents\ in\ the\ corpus}{total\ number\ of\ documents\ with\ term\ t}$$

[22] http://pythonhosted.org/pyenchant/

Final TF-IDF value of term *t* in document *d* is calculated by multiplying these two values. It reflects importance of *t* with respect to *d*.

Using IDF together with TF helps give more importance to words that often occur in a given song, but less in the collection of music. Furthermore, it helps minimize the importance given to frequently occurring terms such as 'the' and 'a'.

Using this measure, the importance of each lyric word with relation to a specific track was calculated. If TF-IDF of a word was less than the average TF-IDF value of all words for corresponding song, we considered the word to be of no significance for that song. Commonly occurring words were thus removed from the lyrics dataset. This reduced the lyric feature dimensionality to 1536 English words, which was more convenient to be handled by a classifier. A CSV file corresponding to this lyric ARFF file was created, a portion of which is depicted in Figure 5.3. A row corresponds to a song whereas the columns depict lyric attributes (words in lyrics). Value 0 indicates that word doesn't appear in the song. Otherwise TF-IDF value is indicated.

Row No.	hatred	yellow	four	upside	captain	whose	spinning	sorry	sweetest
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0.008	0.018	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0.053	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0

Figure 5. 3: Lyrics data file

5.2.3. Handling Missing Values

As specified in section 4.2.3, of the one million songs provided in Million Song Dataset, lyrics, audio and tag features have not been provided for all. Tags have been provided for merely 505 216 of them, audio features for 994 960 and lyrics features for 237 662. For a certain track there may be tags associated with it but not audio features. This has resulted in occurrence of missing values in the dataset. This missing value problem primarily occurs when we later attempt to merge lyrics and audio feature for hybrid classification. Four primary mechanisms of handling missing data has been specified in [42].

I. Removing the entire tuple

II. Manually entering missing values

III. Assuming a constant value to fill missing value (e.g. 'null')

IV. Imputation - Mean or most common attribute value is used to replace missing values.

Mechanism of manually entering missing values was ruled out since it necessitates acquiring features for more than 100 000 songs. Assuming imputation method significantly corrupts the dataset via associating words with songs despite its lyrics not containing it. Using a constant to fill missing values is disregarded, since those could be erroneously considered either as a unique emotion or lyric word by the classifier. Hence the option of removing entire tuple when a value is missing was applied, which results in the retention of 167 023 data instances. Applying this solution is not detrimental to the research outcome, since magnitude of retained dataset is considered highly adequate [42] [43] for data related tasks. Rows corresponding to these 167 023 data points alone were thus retained in created ARFF and CSV files. It resulted in removal of certain columns with relation to the tag-track file depicted in Figure 5.2. Subsequent to this stage another ARFF file was created by combining audio and lyric features. A row in this file corresponds to a song whereas columns depict both lyric and audio features.

5.2.4. Identifying Emotion Related Tags

Since merely emotion related tags are important for identification of emotion clusters, a mechanism had to be adopted for removing non-emotion related tags from the dataset. WordNet [13] tool and GEMS [12] were primarily utilized for derivation of emotion related terms.

WordNet is a tool applied in a number of former researches [26] [14] [39] [36] which helps find whether a given tag has term 'emotion' as its hypernym. A hypernym is a word more generic than a given term [13] which allows one to infer which category a term comes under. GEMS is identified as the most music specific emotion scale existent [24]. Hence, these were utilized as the primary resources for identification of emotion related tags, whereas former research work too was referenced [38].

As mentioned in section 4.2.3, there exist 522,366 unique tags associated with 505,216 tracks in the MSD. Since assessing all these tags to determine whether they were related to emotion was impractical, merely those which were associated with at least 100 music pieces from 167 023 data instances were retained for further study.

The 1446 tags thus derived were further filtered using WordNet and GEMS to retain terms related to emotion. Subsequently inconsistencies among tags were identified and corrected. Certain emotion related tags depicted inconsistencies with relation to spellings and how they were written (e.g. Humor, humour; Feel good, feelgood; Chill-out, chill out, chillout; Love, love love love; Dance, dance dance dance). Such inconsistencies were resolved by selecting a single term from a set of inconsistent terms for research purpose. Rows corresponding to emotion tags alone were retained in tag-track file depicted in Figure 5.2. The preprocessing conducted under this section is of considerable relevance to Chapter 6, which discusses creation of music specific emotion model. Emotion specific tags retained are depicted in Table 6.1.

Note: Data preprocessing tasks described in this chapter from here onwards were executed after creating music specific emotion model, and incorporating emotion clusters in data files. Creation of emotion model is described under Chapter 6.

5.2.5. Removal of Outliers and Extreme Values

Outlier could be defined as “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” [71]. Presence of outliers is detrimental to the performance of a number of classifiers, thus making removal of outliers an important preprocessing task. Among the three features considered in this research experiment, we were concerned of the presence of outliers specifically in audio features set. The lyrics dataset and tags dataset are not affected by the presence of outliers, since they merely reflect presence of words in lyrics and tag-track associations respectively. On the contrary, the audio features are valuable in detecting whether a song is very different from other songs in dataset. Figure 5.4 graphically depicts data points identified as outliers using two audio features, spectral flux and spectral variability. Both the standard deviation and average value of each feature has been considered. Circles correspond to outliers.

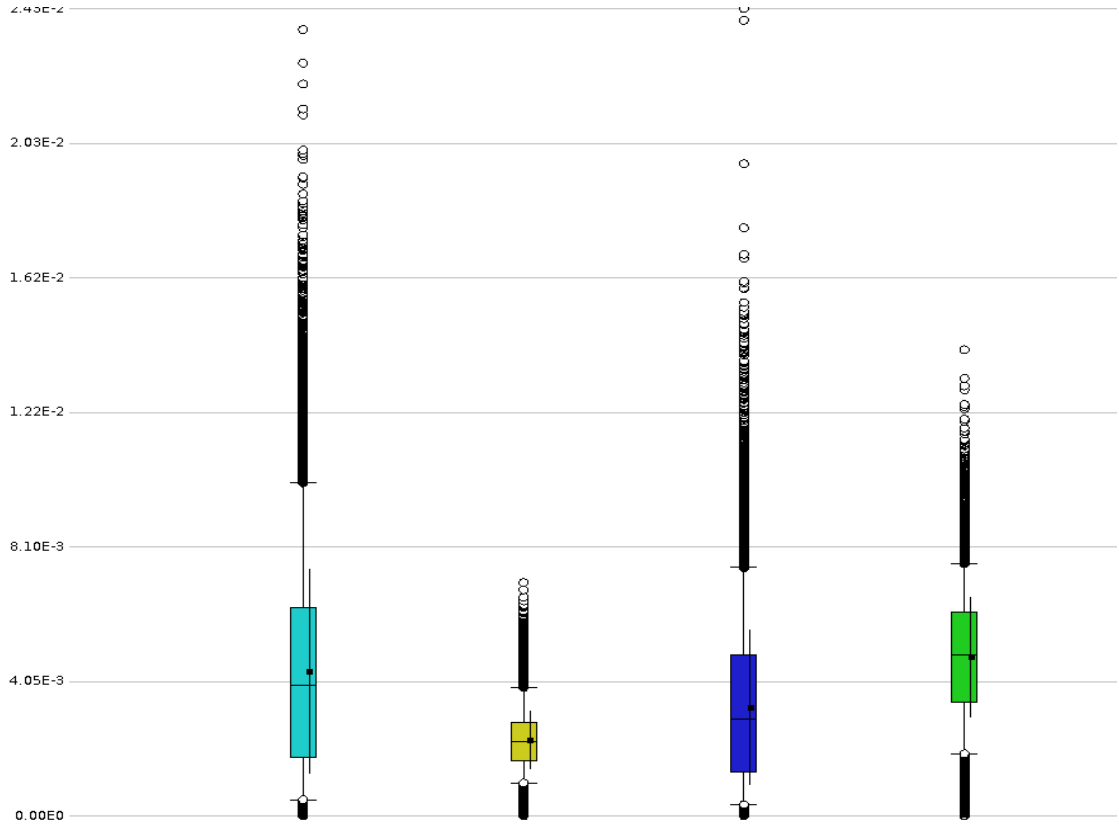

Figure 5. 4: Outlier detection using audio features. A boxplot representation

Inter Quartile Range (IQR)[23] is a statistical measure widely used for outlier detection, which is implemented in Weka. It evaluates presence of outliers and extreme values (extreme outliers) as follows[24] where x denotes a data point.

Outliers	Extreme values
Q3 + OF*IQR < x <= Q3 + EVF*IQR or	x > Q3 + EVF*IQR or
Q1 - EVF*IQR <= x < Q1 - OF*IQR	x < Q1 - EVF*IQR

Where Q1 = 25% quartile, Q3 = 75% quartile, IQR = Difference between Q1 and Q3, OF = Outlier Factor and EVF = Extreme Value Factor

Figure 5.5 depicts distribution of data among 25 emotion classes prior to removal of outliers and extreme values.

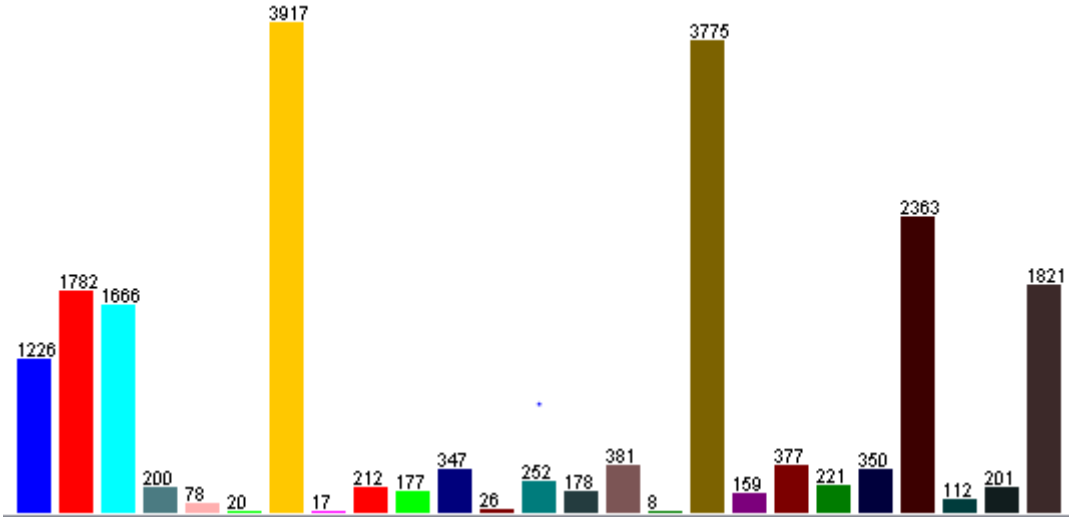

Figure 5. 5: Distribution of data prior to outlier removal

When outlier detection algorithm of Weka was executed on this space, 338 outliers and 34 extreme values were detected. Subsequent to removal of these records, the distribution was as depicted in Figure 5.6. Classification was executed on this refined dataset to minimize the effect of outliers.

[23] Prior to calculating IQR for a particular attribute, attribute values are sorted in ascending order.
[24] http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/InterquartileRange.html

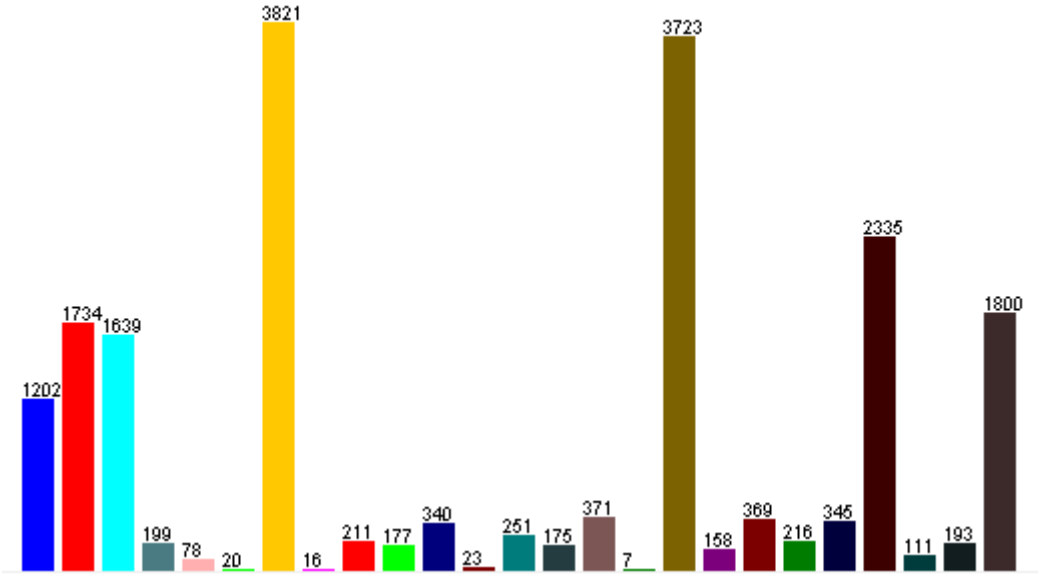

Figure 5. 6: Distribution after outlier removal

5.2.6. Feature Discretization

Feature discretization is the process of converting continuous attribute values into discrete values. As depicted in Figure 5.1, audio features values are continuous which enables application of feature discretization techniques. Attempting classification on discretized data was considered to be important since studies [72] have shown improvement in performance of Naïve Bayes and C4.5 algorithms when data is discretized. Discretization algorithms are primarily of two types, namely, supervised discretization and unsupervised discretization [72]. Algorithms under each category are depicted in Figure 5.7.

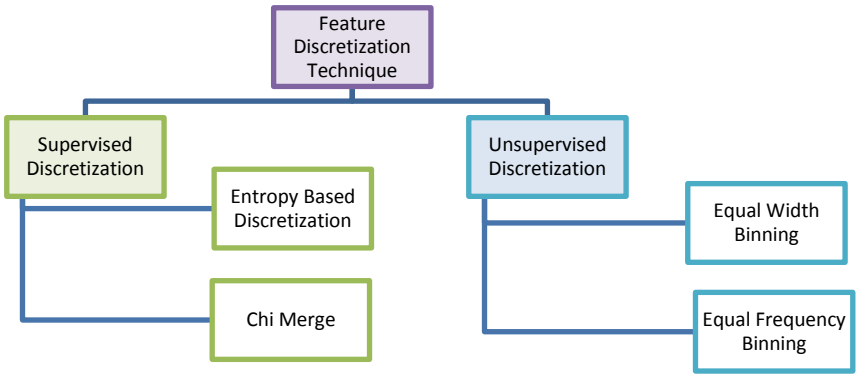

Figure 5. 7: Feature discretization methods

While supervised discretization utilizes class information for deciding how best to perform discretization, unsupervised discretization does not rely on class values.

Supervised Discretization

Entropy discretization method utilizes minimal entropy heuristic, which reflects how well the discretization partitions dataset to distinct classes. For a dataset S, feature A and partition boundary T, entropy of partition T is given by following equation [72].

$$E(A,T;S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2)$$

Chi Merge is a statistical measure which executes supervised discretization method via iteratively merging data points based on a threshold. A high threshold has been found to be effective when discretizing random data, to avoid construction of too many intervals [72].

Unsupervised Discretization

Equal width binning [72] is a simple statistical measure allowing discretization of data into bins (categorical values) of equal width. When maximum value of an attribute is V_{max} and minimum value is V_{min}, bin width of that is specified as follows. Since max and min values are used, this algorithm is highly sensitive to outliers, thus necessitating outliers to be removed from dataset prior to calculating bin width.

$$Bin\ width = \frac{V_{max} - V_{min}}{number\ of\ bins}$$

Equal Frequency Binning [72] strives to obtain approximately the same number of data points for each bin. This algorithm is subject to the problem of repeating same data point in several bins.

Of these four algorithms, equal with binning and entropy based discretization was attempted on the dataset. Chi Merge was not applied due to practical difficulties of determining optimal threshold, whereas equal frequency binning was ruled out due to possible repetition of data in bins.

As shown in Figure 5.8, dispersion of data considerably varies from one another, for different audio attributes.

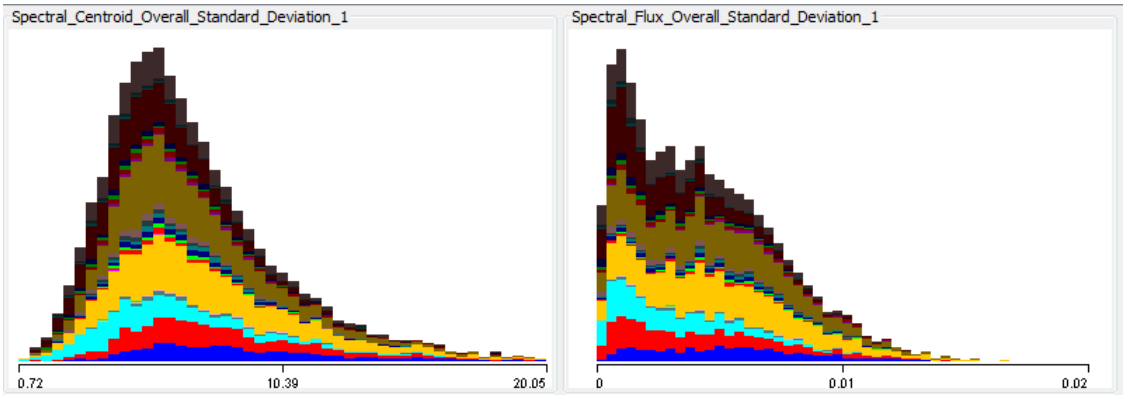

Figure 5. 8: Data dispersion for spectral centroid and spectral flux.

Thus, when applying the equal width binning algorithm, it would be inaccurate to use same number of bins for each distinct audio attribute. Freedman-Diaconis rule[73] is a robust statistical measure which is considered rule-of-thumb for determining optimal number of bins in creating histograms. This measure was thus recognized to be considerably appropriate for determining *number of bins* with respect to each attribute, prior to executing equal width binning.

Freedman-Diaconis rule is specified as follows, where IQR(x) is the interquartile range of the data and n is the number of observations in the sample x.

$$bin\ size = 2IQR(x)n^{-1/3}$$

This allows *number of bins* to be calculated as *(V_{max} – V_{min})/bin size.* Calculation of this value for spectral centroid (standard deviation) audio feature is as follows. IQR of spectral centroid was assessed to be 3.763. For n = 19 514, the bin size obtained is 202.62. The minimum and maximum values of spectral centroid are 0.723 and 20.05 respectively. Thus the *number of bins* obtained for spectral centroid is 0.095.

Getting *number of bins* < 1 disallows application of equal width binning method for research dataset. This is primarily resulted by the high number of records against the limited range within which values of audio features vary. Thus entropy based discretization was applied on dataset. Weka implements Fayyad & Irani's entropy method, which was applied on dataset to perform discretization. Figure 5.9 depicts representation of audio features (corresponds to features depicted in Figure 5.8) subsequent to discretization. Classification was attempted on both the discretized audio dataset and dataset comprising of continuous values.

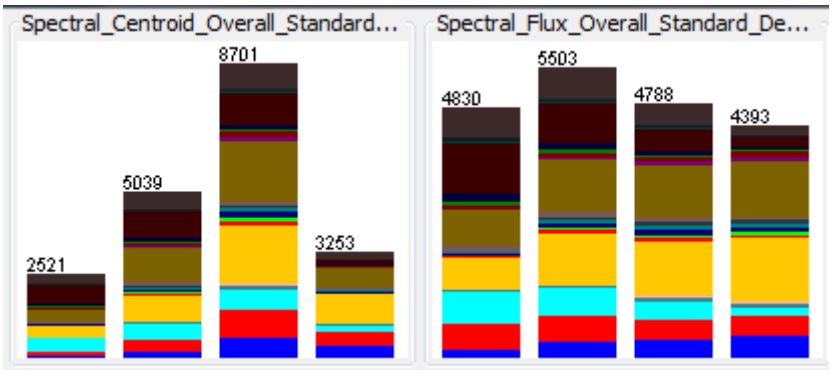

Figure 5. 9: Data dispersion for spectral centroid and spectral flux.

5.3. Feature Selection

Feature selection [74] is the mechanism of selecting a subset from feature space, which are most relevant to class attribute values. Selection of the ideal subset would thus assist in commendable discrimination of class labels. When a good subset is selected, data with similar features would have high probability of belonging to same class whereas records with different features would belong to separate classes. Feature selection was applied on the lyrics and audio feature files with the intent of reducing 40 audio features and 1536 lyric features to a lesser number. Various researchers have categorized feature selection algorithms in numerous ways. Langley's categorization was as *filter* and *wrapper* methods whereas Ben-Bassat extended this categorization as *uncertainty(information)*, *distance* and *dependence* measures [74]. Since dependence measures category from Bassat's definition includes Langley's groups, Ben-Bassat's categories were assumed in this research.

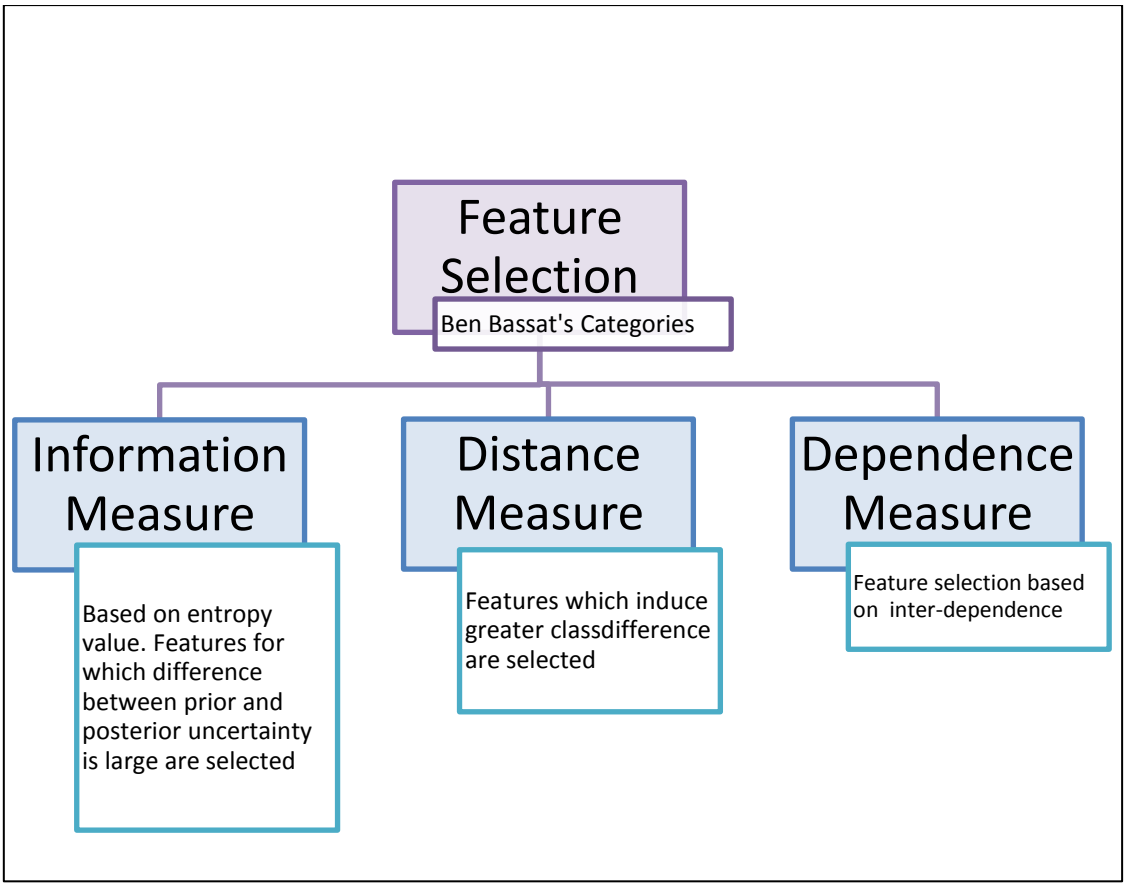

Figure 5. 10: Ben Bassat's feature selection algorithm categories.

Figure 5.10 pictorially represents Bassat's categories of feature selection algorithms. Dependence measure based feature selection methods, which could be further categorized as filter and wrapper methods, are the most commonly applied mechanisms for conducting selection of features [74]. Filter methods perform selection of features in an unsupervised manner. Thus, these methods are independent of the classification algorithm utilized, which improves generalizability of features selected using a filter technique [75]. As opposed to this concept, wrapper methods iteratively refine the subset of features selected, based on feedback from classifier. Wrapper methods are therefore classifier specific which reflects that a feature set selected for Naïve Bayes classifier would not be applicable for C4.5 based classification. This has led to utilization of wrapper methods being prohibitively expensive [75]. Since several classification techniques have been attempted in this study, executing feature selection using a filter method was considered to be more apt. The non-discretized version of audio CSV file was used for correlation analysis.

Feature redundancy analysis, otherwise known as correlation based feature selection is a renowned filter technique used for selecting subset of features [74] [75]. Linear correlation coefficient, which is the primary measure adopted in this technique, is defined as follows [76].

$$\rho = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\ \sqrt{\sum(Y_i - \bar{Y})^2}}$$

For two variables X and Y (X_i and Y_i indicates values of variables), where $\bar{X}$ and $\bar{Y}$ are their respective means, linear correlation coefficient is given by ρ as shown above. While $-1 \leq \rho \leq 1$, equality to -1 or 1 signifies complete correlation [76]. Features depicting high correlation are considered to be redundant, thus reflecting that retaining only one of such a feature pair is adequate for classification. This mechanism which is implemented in RapidMiner was utilized for selecting subsets of audio and lyric features respectively. Figure 5.11 depicts process of generating correlation matrix for audio feature set.

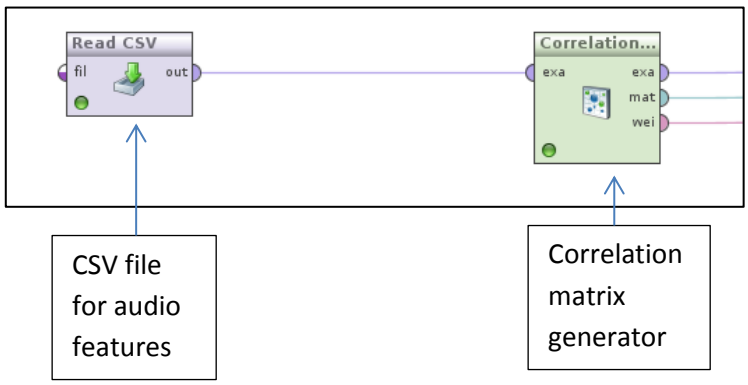

Figure 5. 11: Generating correlation matrix on RapidMiner.

Figure 5.12 depicts partial view of the correlation matrix generated for audio feature file, where higher correlation is highlighted in darker blue. Correlation analysis was performed separately for lyrics feature set, where the resultant matrix is presented in Figure 5.13. These feature types were considered separately since the study conducted audio based classification and lyric based classification individually, as described under a latter chapter. For pairs of features depicting correlation greater than 0.75 or less than -0.75, merely one was selected to be used for classification.

Attributes	Spectral_Centr...	Spectral_...	Spectral_...	Compactn...	Spectral_...	Root_Mea...	Fraction_...	Zero_Cro...	Spectral_...	Spectral_...	Spectral_...
Spectral_Centr	1	0.954	0.016	-0.047	0.303	0.336	0.117	0.976	0.637	0.588	0.003
Spectral_Rollof	0.954	1	0.002	-0.113	0.318	0.343	0.037	0.973	0.627	0.603	-0.014
Spectral_Flux_(	0.016	0.002	1	0.044	-0.067	-0.069	0.074	0.006	-0.004	-0.016	0.288
Compactness_	-0.047	-0.113	0.044	1	0.061	0.072	0.247	-0.080	-0.155	-0.176	0.054
Spectral_Varial	0.303	0.318	-0.067	0.061	1	0.983	-0.055	0.334	0.052	0.062	-0.103
Root_Mean_Sq	0.336	0.343	-0.069	0.072	0.983	1	0.002	0.363	0.096	0.095	-0.111
Fraction_Of_Lo	0.117	0.037	0.074	0.247	-0.055	0.002	1	0.103	-0.086	-0.141	0.065
Zero_Crossing	0.976	0.973	0.006	-0.080	0.334	0.363	0.103	1	0.560	0.514	-0.012
Spectral_Centr	0.637	0.627	-0.004	-0.155	0.052	0.096	-0.086	0.560	1	0.983	-0.007
Spectral_Rollof	0.588	0.603	-0.016	-0.176	0.062	0.095	-0.141	0.514	0.983	1	-0.016
Spectral_Flux_(	0.003	-0.014	0.288	0.054	-0.103	-0.111	0.065	-0.012	-0.007	-0.016	1
Compactness_	-0.130	-0.177	0.046	0.368	-0.224	-0.198	0.225	-0.141	-0.158	-0.203	0.022
Spectral_Varial	-0.032	0.038	-0.113	-0.037	0.720	0.676	-0.344	-0.005	0.060	0.114	-0.109
Root_Mean_Sq	-0.026	0.048	-0.115	-0.052	0.688	0.653	-0.353	-0.003	0.113	0.167	-0.110
Fraction_Of_Lo	0.169	0.189	-0.003	-0.058	0.241	0.230	0.030	0.176	0.156	0.179	-0.023
Zero_Crossing	0.622	0.665	-0.029	-0.221	0.123	0.149	-0.182	0.577	0.964	0.970	-0.033
Mean_Acc5_Me	0.115	0.182	-0.047	-0.051	0.782	0.733	-0.280	0.151	0.087	0.130	-0.019
Mean_Acc5_Me	0.114	0.183	-0.048	-0.054	0.778	0.731	-0.281	0.151	0.086	0.129	-0.021
Mean_Acc5_Me	0.104	0.172	-0.050	-0.052	0.771	0.725	-0.283	0.140	0.083	0.126	-0.024
Mean_Acc5_Me	0.106	0.174	-0.048	-0.051	0.780	0.730	-0.284	0.142	0.074	0.119	-0.022
Mean_Acc5_Me	0.111	0.179	-0.046	-0.052	0.786	0.734	-0.284	0.148	0.073	0.119	-0.019
Mean_Acc5_Me	0.110	0.177	-0.046	-0.050	0.788	0.735	-0.284	0.146	0.072	0.118	-0.017
Mean_Acc5_Me	0.111	0.177	-0.045	-0.048	0.792	0.738	-0.284	0.147	0.069	0.115	-0.016
Mean_Acc5_Me	0.108	0.174	-0.045	-0.046	0.791	0.737	-0.285	0.144	0.067	0.113	-0.017
Mean_Acc5_Me	0.113	0.179	-0.044	-0.047	0.794	0.739	-0.282	0.150	0.065	0.111	-0.016
Mean_Acc5_Me	0.121	0.187	-0.043	-0.049	0.795	0.741	-0.278	0.158	0.071	0.115	-0.015
Mean_Acc5_Me	0.122	0.189	-0.043	-0.050	0.793	0.740	-0.277	0.159	0.077	0.121	-0.014
Mean_Acc5_Me	0.121	0.188	-0.044	-0.052	0.789	0.737	-0.278	0.157	0.083	0.127	-0.016
Std_Acc5_Std_I	0.361	0.394	-0.015	-0.047	0.771	0.766	0.030	0.405	0.125	0.125	-0.058
Std_Acc5_Std_I	0.362	0.395	-0.014	-0.048	0.765	0.760	0.029	0.407	0.125	0.124	-0.057
Std_Acc5_Std_I	0.360	0.391	-0.015	-0.044	0.773	0.768	0.026	0.403	0.121	0.121	-0.059
Std_Acc5_Std_I	0.359	0.390	-0.016	-0.038	0.793	0.788	0.026	0.402	0.119	0.120	-0.062

Figure 5. 12: Correlation matrix for audio features.

Attributes	hatred	yellow	four	upside	captain	whose	spinning	sorry	sweetest	certain	pride	worth	lon
hatred	1	-0.002	-0.003	-0.002	-0.001	0.004	-0.002	-0.000	-0.002	-0.001	0.007	-0.000	-0.0
yellow	-0.002	1	0.003	0.001	-0.000	0.003	-0.000	-0.000	-0.000	-0.002	-0.002	-0.003	-0.0
four	-0.003	0.003	1	-0.000	0.007	0.008	-0.001	-0.002	-0.003	-0.004	-0.001	-0.003	-0.0
upside	-0.002	0.001	-0.000	1	-0.000	-0.001	0.001	0.008	0.000	-0.002	0.000	-0.001	-0.0
captain	-0.001	-0.000	0.007	-0.000	1	0.003	-0.002	0.002	-0.001	-0.001	-0.001	0.004	0.00
whose	0.004	0.003	0.008	-0.001	0.003	1	-0.001	-0.004	-0.001	0.003	0.001	0.002	0.00
spinning	-0.002	-0.000	-0.001	0.001	-0.002	-0.001	1	-0.003	-0.002	-0.002	0.000	-0.002	-0.0
sorry	-0.000	-0.000	-0.002	0.008	0.002	-0.004	-0.003	1	-0.001	-0.003	-0.000	0.007	0.00
sweetest	-0.002	-0.000	-0.003	0.000	-0.001	-0.001	-0.002	-0.001	1	-0.002	-0.002	0.010	0.00
certain	-0.001	-0.002	-0.004	-0.002	-0.001	0.003	-0.002	-0.003	-0.002	1	-0.002	0.008	0.00
pride	0.007	-0.002	-0.001	0.000	-0.001	0.001	0.000	-0.000	-0.002	-0.002	1	0.003	0.00
worth	-0.000	-0.003	-0.003	-0.001	0.004	0.002	-0.002	0.007	0.010	0.008	0.003	1	-0.0
lonesome	-0.002	-0.001	-0.001	-0.002	0.001	0.009	-0.002	0.000	0.008	0.004	0.002	-0.002	1
digital	-0.001	-0.000	-0.000	-0.000	-0.001	-0.001	-0.000	-0.001	0.009	-0.001	-0.002	-0.001	-0.0
void	0.004	0.000	-0.001	-0.001	-0.001	0.005	0.002	-0.004	0.003	-0.002	-0.001	0.000	-0.0
distorted	-0.001	0.001	-0.002	-0.001	-0.001	0.011	0.002	-0.003	-0.001	-0.001	0.003	-0.001	-0.0
government	-0.001	0.001	-0.000	-0.001	0.007	0.002	-0.001	-0.002	-0.001	-0.001	0.001	-0.002	0.00
alas	-0.001	-0.001	-0.001	-0.001	-0.000	-0.000	-0.001	-0.002	-0.001	-0.001	-0.001	-0.002	-0.0
school	-0.002	0.010	0.004	-0.002	0.000	-0.000	-0.002	-0.005	-0.003	0.001	-0.004	-0.003	-0.0
prize	0.002	-0.000	0.002	0.010	0.001	0.002	0.001	-0.003	0.001	-0.002	0.004	0.005	-0.0
chew	-0.002	0.001	0.008	0.000	-0.001	0.001	-0.001	-0.001	-0.001	-0.001	-0.002	-0.003	-0.0
horn	-0.002	-0.000	0.001	0.001	0.003	0.003	-0.002	-0.003	0.005	-0.000	0.002	-0.003	-0.0
nails	0.000	-0.002	-0.002	0.003	0.000	0.004	-0.001	0.007	-0.002	-0.003	0.004	0.002	-0.0
second	-0.001	-0.000	0.018	-0.000	-0.001	0.007	-0.001	0.005	0.002	0.002	0.001	0.010	0.00
thunder	-0.002	-0.002	-0.000	-0.001	-0.001	0.000	0.003	-0.005	0.001	-0.002	0.001	-0.003	0.00
lightning	0.000	-0.001	-0.001	-0.001	-0.001	-0.001	0.030	0.005	-0.001	-0.003	0.001	-0.002	0.00
poison	0.006	0.009	0.009	-0.001	-0.001	-0.000	0.002	-0.003	0.003	0.002	0.005	-0.001	-0.0
haze	-0.002	0.004	0.006	-0.002	-0.001	0.007	-0.001	-0.004	-0.001	-0.001	-0.001	-0.002	0.01
toll	-0.002	-0.001	-0.001	-0.002	-0.001	0.003	-0.002	-0.003	-0.002	-0.001	0.007	-0.001	-0.0
hero	0.001	-0.001	0.005	-0.001	-0.001	0.005	-0.003	-0.000	0.001	-0.001	-0.000	-0.001	-0.0

Figure 5. 13: Correlation matrix for lyric features.

As depicted in Figures 5.12 and 5.13, the diagonal depicts complete correlation with value 1, since it represents correlation of a feature with itself. Analysis of the two matrices helped us deduce that it was possible to select a subset of audio features based on decided threshold of +/-0.75. However, none of the lyric feature pairs depicted significant correlation thus compelling us to retain all 1536 lyric words for classification. Based on analysis of audio feature correlation matrix depicted in Figure 5.12, ten audio features were selected for learning purposes. These were the standard deviation and average values of following features.

I. Spectral centroid
II. Spectral flux
III. Compactness
IV. Spectral variability
V. Fraction of low energy windows

Thus three data files were subject to refinement for reflecting this selection. Those are ARFF file for audio features, CSV file for audio features and ARFF file for hybrid (audio + lyric) features.

5.4. Final Dataset

As extensively described under section 5.2 and section 5.3, conducting various preprocessing and feature selection tasks has caused significant alterations to the files initially created (section 5.2.1). In the initial phase, the data files were created considering merely the audio and lyric attributes. The class attribute 'emotion' was excluded in the initial stages, since the music specific emotion model (discussed under Chapter 6) was required to be designed prior to incorporating class attribute. Thus, the preprocessing stages discussed under sections 5.2.1, 5.2.2, 5.2.3, and 5.2.4, together with feature selection described under section 5.3 were executed before creating the emotion model. However, preprocessing stages described in sections 5.2.5 and 5.2.6 were conducted subsequent to evaluating emotion clusters as elaborated under Chapter 6, and incorporating them in the data files.

The final data files specified herewith were used for classification tasks analyzed under Chapter 7, except the CSV tag-track file. The tag-track file was utilized for creation of music specific emotion model as described under next chapter. The audio CSV file and lyric CSV file formerly created were merely used for feature selection task. Thus, those two files were not retained for conducting experiments. Following describe the final dataset used in subsequent experiments and model creation. All these files were created using python based self-implemented programs.

File I: *CSV tag-track file*

This was used for music specific emotion model creation which is discussed under Chapter 6. Rows correspond to emotion related tags whereas columns correspond to songs. Presence or absence of a tag with respect to a song is depicted by 1 and 0 respectively. This final CSV tag-track file is similar in format to one depicted by Figure 5.2. Rest of the files specified below were created after forming emotion model, whereas those were used for classification (relates to Chapter 7)

File II: *ARFF audio feature file for multi-label classification*

This file comprises of audio features (selected 10 audio features) and emotion class attributes per song. Non-discretized audio features have been utilized. The file reflects all the emotions evoked by a particular song. A row corresponds to a song whereas columns correspond to final emotion clusters and audio features. 1 depicts that a song belongs to corresponding emotion cluster, whereas 0 depicts otherwise. Audio feature value assumed by a song for selected 10 audio features is depicted. Figure 5.14 provides with a partial view of this data file.

File III: *ARFF audio feature file for single label classification (non-discretized)*

This file comprises of 10 selected audio features and emotion class attribute per song. The most representative emotion of a song is reflected as depicted in Figure 5.15. The emotion attribute can assume values of identified emotion clusters.

File IV: *ARFF audio feature file for single label classification (discretized)*: Similar to File III, with discretized values for audio features being assigned. Figure 5.9 depicts how audio attributes are viewed when loaded on Weka.

File V: *ARFF lyrics feature file for single label classification:* Similar to File III, where 1536 lyric words are used instead of audio features. As depicted under section 5.2.6, discretization was attempted only on audio data, since the majority of lyric feature values assumed by songs were 0. Therefore this file format is as shown in Figure 5.3.

File VI: *ARFF hybrid (audio + lyric) feature file for single label classification:* This is a combination of File III and File V, where both lyric and audio information is provided in one file (file is too large to be presented as an image).

```
0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3.808,0.005712,244,0.002615,0.07392,5.531,0.003178,1643,0.005318,0.5705
1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,12.24,0.001293,215,0.001488,0.05323,14.06,0.000819,1696,0.002419,0.5659
0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,12.04,0.005138,212.7,0.002722,0.04087,16.5,0.0052,1522,0.005767,0.6383
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5.825,0.007174,204.1,0.002978,0.05591,7.657,0.00649,1598,0.00693,0.5641
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,9.872,0.00033,226.9,0.001225,0.07704,11.7,0.000228,1571,0.001771,0.5542
0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,6.603,0.004536,190.6,0.002502,0.05125,8.921,0.004631,1583,0.006002,0.5374
0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5.465,0.009556,195,0.002873,0.05884,8.569,0.006011,1585,0.006073,0.6135
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,4.812,0.005973,177.7,0.003013,0.03996,6.579,0.004026,1556,0.004829,0.6099
```

Figure 5. 14: ARFF audio feature file for multi-label classification

```
3.808,0.005712,244,0.002615,0.07392,5.531,0.003178,1643,0.005318,0.5705,uplifting_feelgood
12.04,0.005138,212.7,0.002722,0.04087,16.5,0.0052,1522,0.005767,0.6383,aggressive_angry
5.825,0.007174,204.1,0.002978,0.05591,7.657,0.00649,1598,0.00693,0.5641,joyful_danceable
9.872,0.00033,226.9,0.001225,0.07704,11.7,0.000228,1571,0.001771,0.5542,creepy
6.603,0.004536,190.6,0.002502,0.05125,8.921,0.004631,1583,0.006002,0.5374,sensual
5.465,0.009556,195,0.002873,0.05884,8.569,0.006011,1585,0.006073,0.6135,sensual
4.812,0.005973,177.7,0.003013,0.03996,6.579,0.004026,1556,0.004829,0.6099,crazy
8.149,0.005847,282.3,0.002489,0.05441,8.747,0.003447,1825,0.006316,0.5119,sensual
6.004,0.002821,257.4,0.002204,0.06488,6.953,0.002974,1710,0.005392,0.5112,nostalgic
12.27,0.002346,173.2,0.001531,0.05168,11.5,0.001452,1605,0.003262,0.5948,nostalgic
7.198,0.003126,251.1,0.002083,0.06178,8.742,0.001778,1710,0.0037,0.5693,sensual
6.872,0.001561,243.3,0.001563,0.07306,10.71,0.000715,1780,0.002899,0.5358,crazy
3.331,0.001637,216.5,0.002171,0.06701,7.191,0.001197,1776,0.004722,0.5431,witty
5.931,0.001641,200.3,0.00149,0.04485,4.301,0.001435,1645,0.003569,0.5279,sensual
3.894,0.006183,229.5,0.003299,0.06748,7.707,0.005409,1622,0.006847,0.5344,sentimental
```

Figure 5. 15: ARFF audio feature file for single-label classification; non-discretized.

5.5. Summary

This chapter provided with an overview of data preprocessing and feature selection tasks, which were conducted for preparation of data to be fed to classifiers and for creating music specific emotion model. Selection of preprocessing mechanisms and feature selection algorithms was validated via comparative analysis of several techniques. The final dataset was presented under section 5.4, where six files were created for conducting various experiments. Among these files, File I was required for creating music specific emotion model (Chapter 6) whereas the rest was utilized for classification (Chapter 7).

The next chapter describes creation of music specific emotion model utilizing emotion related tags. Despite it being presented under Chapter 6, creation of the model was a prerequisite for executing certain preprocessing tasks and file creation, which were elaborated in this chapter. However all preprocessing tasks were opted to be described under this chapter, to avoid segmentation of Feature Engineering details.

Chapter 6 - Music Emotion Model

6.1. Introduction

The former chapter extensively described the data preprocessing and feature engineering tasks executed prior to preparation of final data files. Merely ten audio features were retained subsequent to feature selection phase whereas numerous data reforming tasks as outlier removal, feature discretization were executed under data preprocessing.

This chapter is dedicated to describing the steps followed in creation of music specific emotion model, which is executed with the intent of determining the class attribute. Knowledge regarding class attribute is necessary for conducting supervised learning as extensively analyzed under chapter 7. Identification of emotion related tags which was discussed under section 5.2.4 is a prerequisite for creation of music specific emotion model. File I specified under section 5.4 was utilized for model creation.

6.2. Clustering Synonymous Emotion Related Tags

Creation of music specific emotion model corresponds to the third component of the architecture diagram depicted in Figure 3.1. The initial subcomponent of this phase was addressed in section 5.2.4, which helped retrieve emotion related tags and resolve inconsistencies among certain tags. The tag set thus retained has been utilized in this section.

In this phase, derivationally related emotion tags were initially identified using the WordNet[13] tool. The term 'derivational' has been defined as follows on word net:

Derivational: characterized by inflections indicating a semantic relation between a word and its base.

Figure 6.1 depicts an example for retrieval of derivationally related terms using WordNet tool.

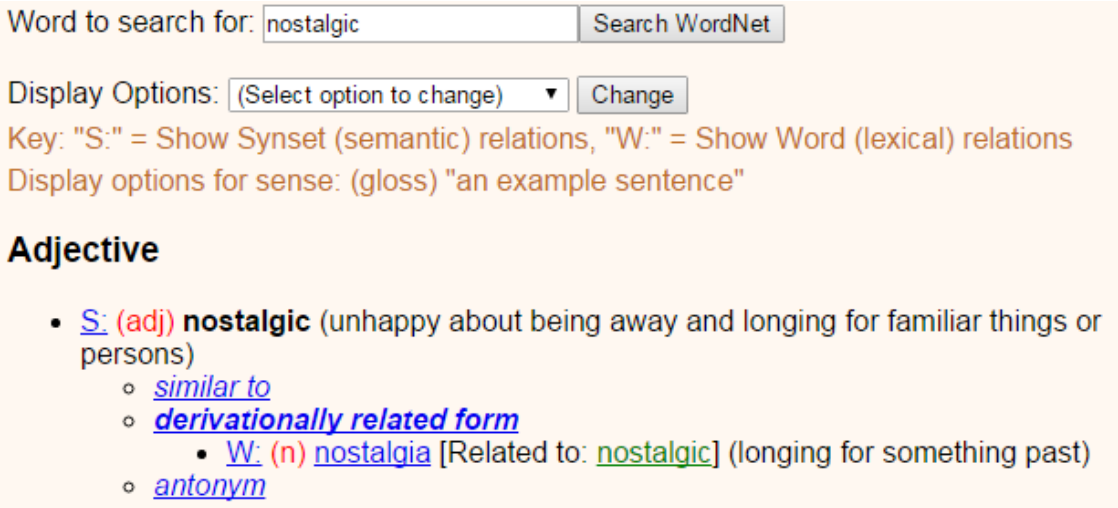

Figure 6. 1: Retrieving derivationally related terms from WordNet tool

Thus, tags such as 'nostalgic', 'nostalgia'; 'power', 'powerful' could be combined since they are derivationally related. These terms were further combined if they happened to be synonymous. As interpreted in WordNet tool, term 'synonymous' could be defined as follows.

Synonymous: meaning the same or nearly the same.

Figure 6.2 depicts an example for retrieval of synonymous words using WordNet.

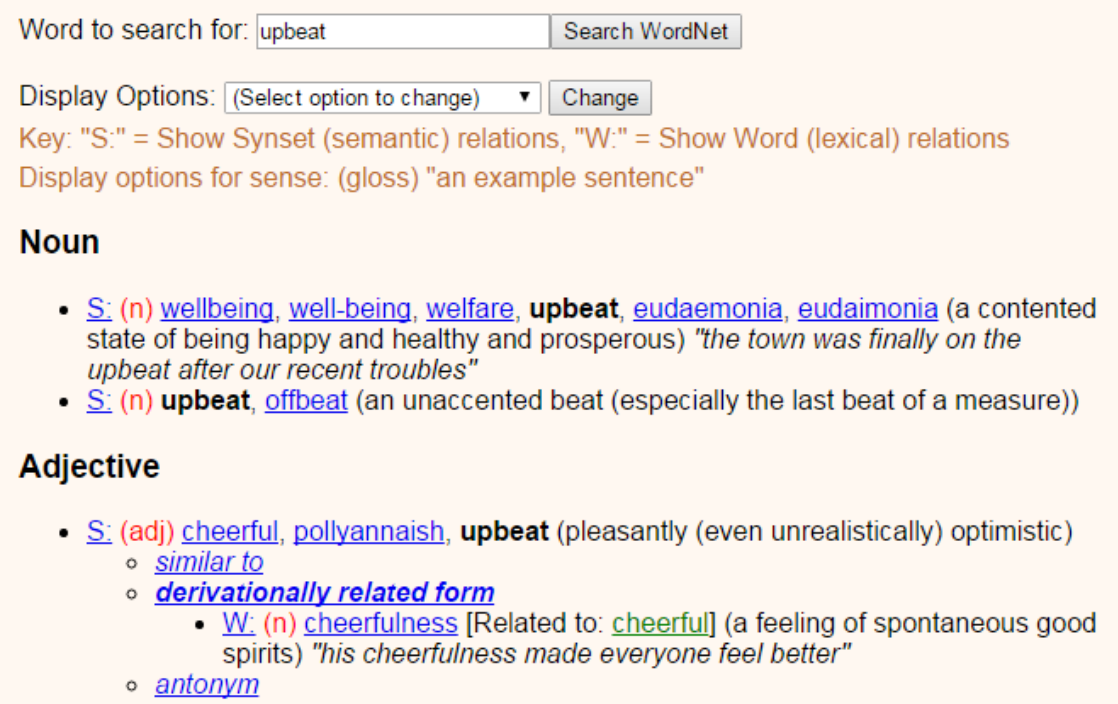

Figure 6. 2: Retrieving synonymous terms from WordNet tool

This allowed further reduction of the emotion space by combining synonymous terms as ‘cheerful’, ‘upbeat’ and ‘witty’, ‘humor’. The 37 emotion clusters identified by completion of this process is depicted in Table 6.1. Emotion tags within a single cell, highlighted belong to one emotion cluster. Highlighted words (numbered) depict the single tag chosen to represent all the tags in its emotion cluster.

Emotion Clusters Formed Via Combining Synonyms	**1) Joyful** Makes me happy Happy Happiness Fun Cheerful Upbeat
2) Witty Humor Funny Comedy	**3) Calming** Calm Chill-out Chill Relaxed Relax Relaxing At ease Mellow Peaceful Soft
4) Uplifting Positive Affirming Optimistic Hope	**5) Aggressive**
6) Moving	**7) Sensual** Erotic Sexy Passionate
8) Haunting	**9) Feel good music** I feel good Good mood feel good
10) Angst	**11) Sick**
12) Melancholic Sadness Makes me cry Sad Sad songs Melancholy	**13) Makes me smile** Smile

14) Heartbreaking Heartbreak Heartache	**15) Crazy**
16) Inspiring Inspirational	**17) Exciting**
18) Creepy Strange Eerie	**19) Ethereal**
20) Dark	**21) Cool**
22) Romantic Romance Love Loved	**23) Bittersweet**
24) Energetic Energy	**25) Soulful**
26) Dreamy Dream	**27) Angry**
28) Hypnotic	**29) Soothing**
30) Sentimental	**31) Danceable** Dancing Dance
32) Depressing Depression Depressive	**33) Powerful** Power Intense
34) Psychedelic	**35) Lonely**
36) Spiritual	**37) Nostalgic** Nostalgia Wistful Longing

Table 6. 1: Thirty seven emotion clusters formed by combining synonyms

6.3. Hierarchical Clustering

Subsequent to identifying synonymous tags and reducing emotion space to 37 clusters, this knowledge was reflected in File I specified under section 5.4. This was achieved by reforming the file using a python based self-implemented program. Rows corresponding to emotion tags from the same cluster were merged using column-wise logical OR operation. Since the columns in File I correspond to songs, executing this code results in reflecting that a song belongs to an emotion cluster, if at least one word from that cluster is associated with it. Due the emotion clusters considered being a collection of synonymous tags, the logic on which this merging is based is validated (i.e. if a song has the tag 'happy' associated with it, it's logical to say that the song expresses the emotion 'joyful').

This refined version of File I was used to execute Hierarchical clustering, with the intent of further reducing dimensionality of emotion space. Rationale behind applying Hierarchical clustering was discussed under section 4.6.1. Among many common distance measures such as Euclidean, Manhattan, etc. the Jaccard distance (explained under section 4.6.1) was the most effective measure which had the strength of reflecting what proportion of songs a pair of tags had in common. Calculation of Jaccard distance X, between two tags A and B could be depicted as follows (using Jaccard distance equation in section 4.6.1).

$$X = 1 - \frac{|songs\ with\ tag\ A\ \cap songs\ with\ tag\ B|}{|songs\ with\ tag\ A\ \cup songs\ with\ tag\ B|}$$

A tag in this instance relates to one of the 37 emotion clusters retained subsequent to synonyms combination. The R tool was used for executing five Hierarchical clustering experiments as described herewith. Jaccard distance measure was used in each of these instances. Linkage method used in each stage was described under section 4.6.1. Dendrograms could be used to graphically represent hierarchical clustering. The leaf numbers 1,2,….37 of each dendrogram representation corresponds to emotion clusters 1,2,…37 depicted in Table 6.1.

Hierarchical clustering based on single linkage method is depicted by Figure 6.3

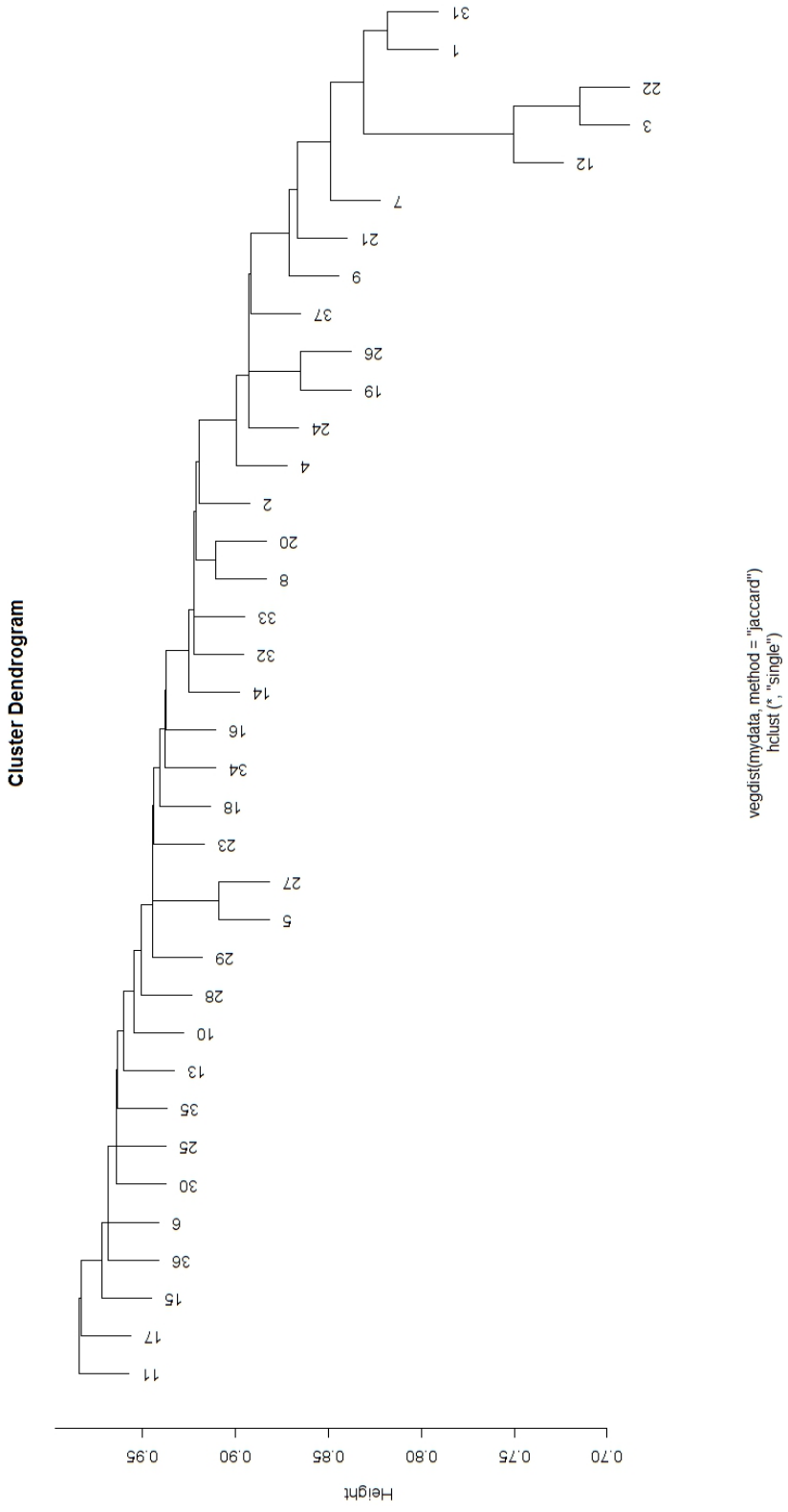

Figure 6. 3: Dendrogram for single linkage hierarchical clustering

Hierarchical clustering based on complete linkage method is depicted by Figure 6.4

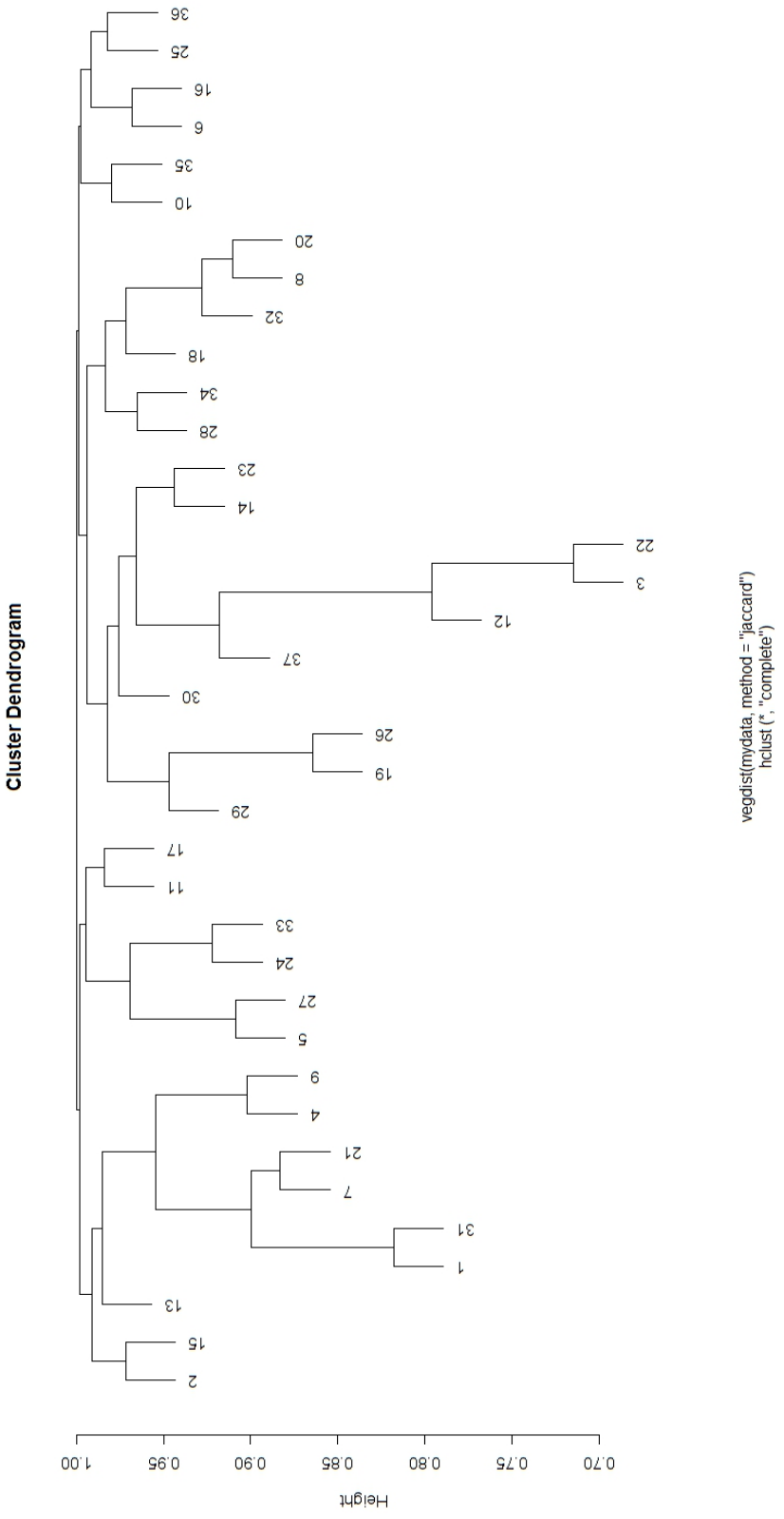

Figure 6. 4: Dendrogram for complete linkage hierarchical clustering

Hierarchical clustering based on group average method is depicted by Figure 6.5

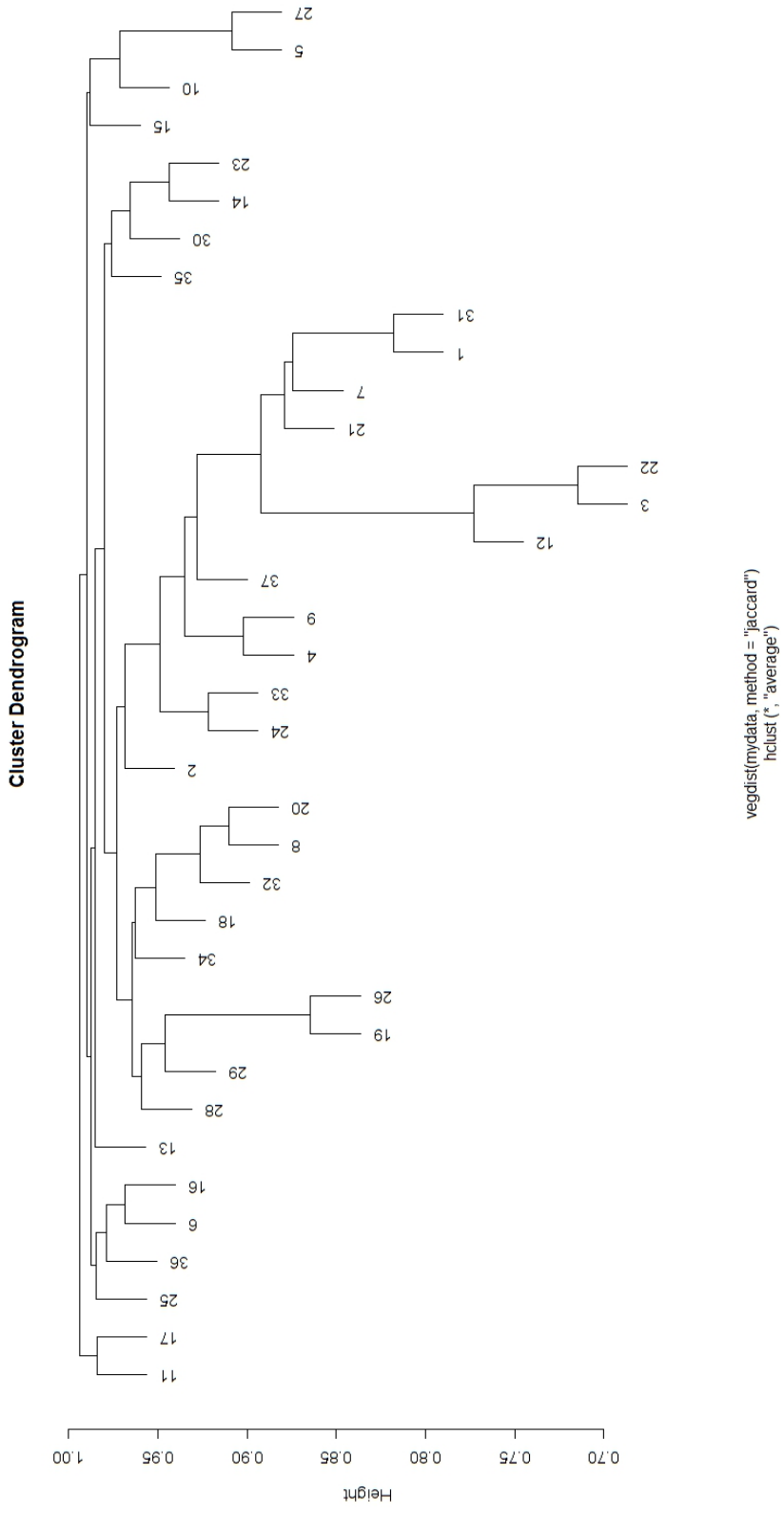

Figure 6. 5: Dendrogram for group average hierarchical clustering

Hierarchical clustering based on Ward's criterion (Ward1) is depicted by Figure 6.6

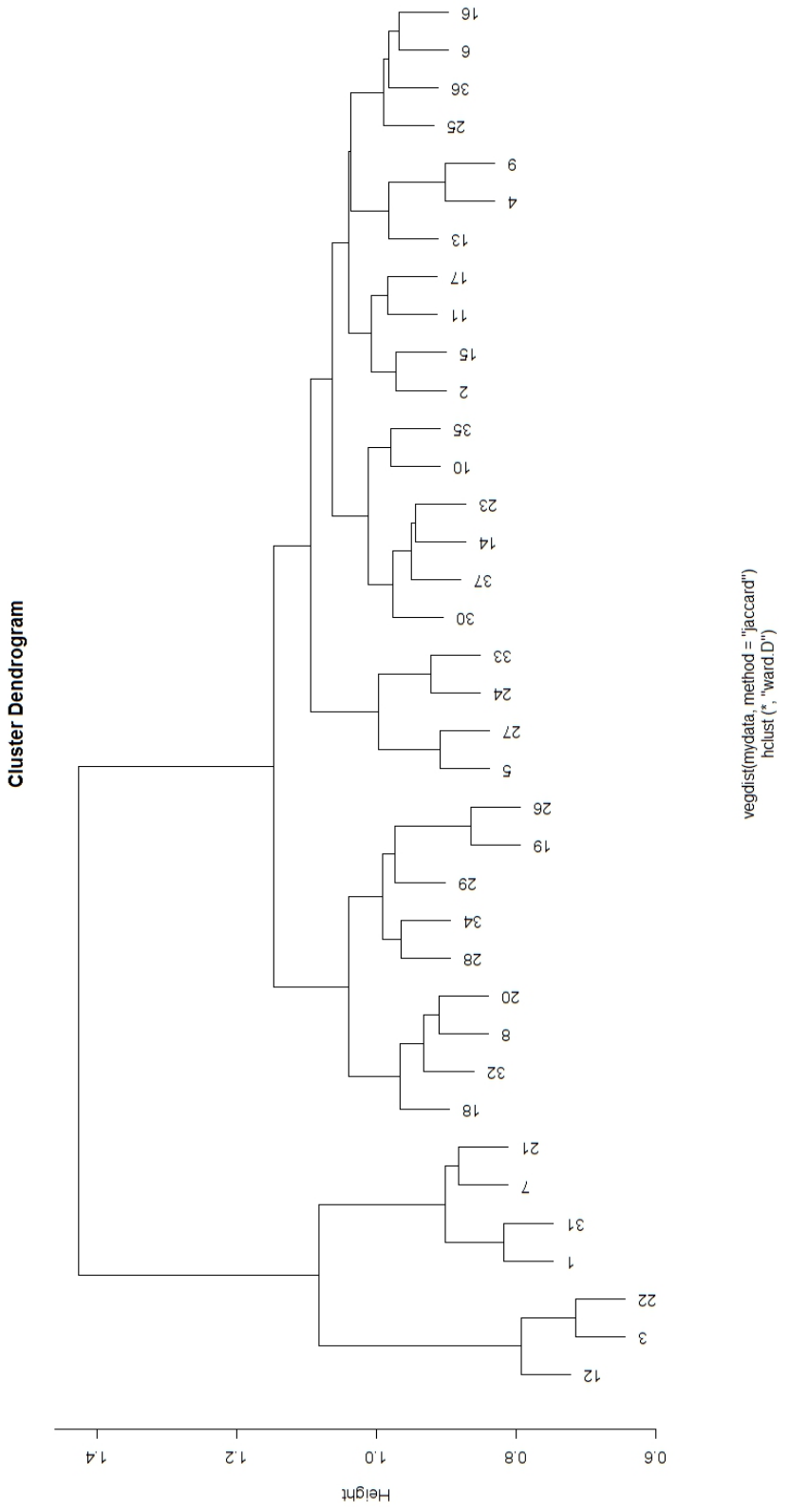

Figure 6. 6: Dendrogram for hierarchical clustering using Ward1 measure

Hierarchical clustering based on Ward's criterion (Ward2) is depicted by Figure 6.7

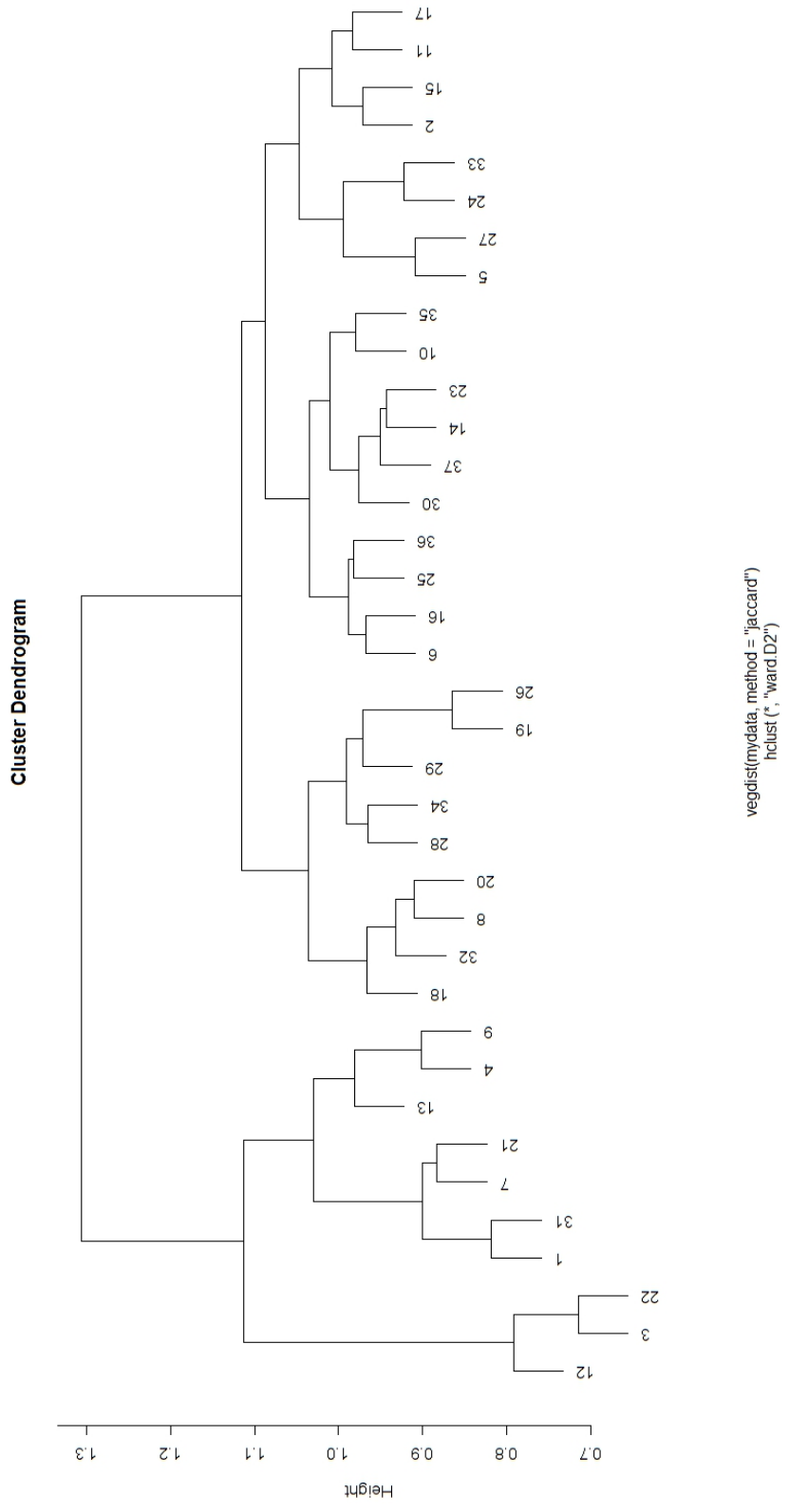

Figure 6. 7: Dendrogram for hierarchical clustering using Ward2 measure

Hierarchical clustering is an iterative process as described under section 4.6.1. Thus emotion clusters considered to be most similar, depending on the linkage measure used, are iteratively combined in this algorithm, to form larger clusters. The clusters combined at lower levels of the Dendrogram represent clusters with higher similarity.

The criterion used for selecting clusters to merge was them being at the lowest level of the Dendrogram in 80% of the hierarchical experiments. Thus, out of the five dendrograms (Figures 6.3, 6.4, 6.5, 6.6 and 6.7) four had to depict merging of clusters at lowest height, for them to be combined.

This allowed further combining emotion clusters formerly identified (Table 6.1) as depicted in Table 6.2.

Label in dendrogram corresponding to clustered emotions	**Corresponding emotions**
1, 31	Joyful, danceable
3, 12, 22	Calming, melancholic, romantic
19, 26, 29	Ethereal, dreamy, soothing
8, 20, 32	Haunting, dark, depressing
4, 9	Uplifting, feel good
5, 27	Aggressive, angry
6, 16	Moving, inspiring
14, 23	Heartbreaking, bittersweet
24, 33	Energetic, powerful

Table 6. 2: Selection of emotions for clustering

Even though forming these clusters was entirely automated via usage of a clustering algorithm, analysis of the semantic meaning of combined emotions helped us ascertain their acceptability (i.e. it's logical to consider that joyful songs make one feel like dancing and for aggressive songs to make us angry.

An interesting observation was the second cluster depicted in Table 6.2. Emotions 'calming', 'melancholic' and 'romantic' have been grouped to form a single cluster. Despite this knowledge not being evident, this cluster formed by hierarchical clustering was considered acceptable, since the rest of the clusters formed using same mechanism are semantically meaningful. Based on this cluster formation it's logical to infer that the three emotions 'calming', 'melancholic' and 'romantic' are generally evoked together when listening to certain songs. The final 25 emotion clusters resultant subsequent to merging those specified in Table 6.2 are as follows. Emotions within parentheses depict emotion clusters.

1: (Joyful, danceable), 2: witty, 3: (calming, melancholic, romantic), 4: (uplifting, feel good), 5: (aggressive, angry), 6: (moving, inspiring), 7: sensual, 8: (haunting, dark, depressing), 9: angst, 10: sick, 11: makes me smile, 12: (heartbreaking, bittersweet), 13: crazy, 14: exciting, 15: creepy, 16: (ethereal, dreamy, soothing), 17: cool, 18: (energetic, powerful), 19: soulful, 20: hypnotic, 21: sentimental, 22: psychedelic, 23: lonely, 24: spiritual, 25: nostalgic.

These final emotion clusters were incorporated in File II, File III, File IV, File V and File VI in the final dataset creation stage described under section 5.4. Thus, these 25 emotion clusters serve as class attribute values in single-label clustering, whereas they are represented as an array of class attributes in the multi-label scenario. In creation of files for single-label clustering, a song in training set was associated with an emotion cluster, only if all the emotions in that cluster were evoked by it.

6.4. Summary

This chapter focused on creation of the music specific emotion model based on tags associated with songs in the Million Song Dataset. Synonyms combination and hierarchical clustering were adopted to achieve this objective. 25 music specific emotion clusters were thus identified, which were incorporated in the files used for classification.

Next chapter discusses the results obtained for each classification experiment attempted.

Chapter 7 - Results and Discussion

7.1. Introduction

The former chapter was dedicated to elaborating the mandatory steps followed in creation of the music specific emotion model. The model creation was necessary to determine the values of class attribute of the dataset. Classification algorithms would consider these class values as the 'true' emotions evoked by a song when training classifiers (i.e. classifier evaluates its performance utilizing these class values as ground truth)

This chapter discusses the evaluation of classifier performance with respect to their aptitude for accurate prediction of emotions evoked by music. Algorithms, evaluation procedure, evaluation metrics, etc. discussed under Chapter 4 are extensively referred in this chapter, since their practical application is associated with classification tasks. This chapter corresponds to the last two components of the overall architecture depicted in Figure 3.1. Classification was initially attempted only using audio features due to time constraint. The rationale of this selection was based on the literature where audio based classification has generally outperformed lyric based classification. Thus the 5^{th} component of architecture was attempted only in instances where audio based classification produced satisfactory results.

7.2. Multi-label Classification Attempt

The initial learning approach attempted was multi-label classification based on audio features. The algorithms attempted in most cited research paper on audio based multi-label classification [29]were applied in this approach. Multi-label classification algorithms described under section 4.6.3 were attempted using the SVM (implemented as SMO in Meka) algorithm as base classifier. Classification was initially attempted using default parameters of SMO (using polynomial kernel as the kernel).

File II (refer section 5.4) which was created for executing multi-label classification was utilized in this experiment. Meka tool which supports multi-label classification was used, on which the algorithms Binary Relevance (BR), Label Power-set (LP), Multi Label K Nearest Neighbor (MLKNN) and Random K Label-sets (RAKEL) were run. Performance of each classifier utilizing evaluation metrics defined in section 4.5, is depicted by Tables 7.1, 7.2 and 7.3. RAKEL was attempted with parameter k set to 5.

Individual Evaluation Measures

Label: emotion class	**Accuracy**			
	BR	**LP**	**MLKNN**	**RAKEL**
joyful_danceable	0.883	0.883	0.883	0.883
witty	0.902	0.902	0.902	0.902
calming_melancholic_romantic	0.848	0.845	0.848	0.848
uplifting_feelgood	0.979	0.979	0.979	0.979
aggressive_angry	0.994	0.994	0.994	0.994
moving_inspiring	0.998	0.998	0.014	0.998
sensual	0.730	0.587	0.730	0.730
haunting_dark_depressing	0.997	0.997	0.997	0.997
angst	0.983	0.983	0.983	0.983
sick	0.990	0.990	0.990	0.990
makes	0.971	0.971	0.971	0.971
heartbreaking_bittersweet	0.995	0.995	0.995	0.995
crazy	0.978	0.978	0.978	0.978
exciting	0.986	0.986	0.986	0.986
creepy	0.970	0.970	0.970	0.970
ethereal_dreamy_soothing	0.997	0.997	0.997	0.997
cool	0.739	0.585	0.739	0.739
energetic_powerful	0.983	0.983	0.983	0.983
soulful	0.965	0.965	0.965	0.965
hypnotic	0.977	0.977	0.977	0.977
sentimental	0.970	0.970	0.970	0.970
psychedelic	0.861	0.798	0.861	0.861
lonely	0.989	0.989	0.989	0.989
spiritual	0.985	0.985	0.985	0.015
nostalgic	0.861	0.861	0.861	0.861

Table 7. 1: Class-wise accuracy in multi-label classification

	BR	LP	MLKNN	RAKEL
Average Accuracy	0.941	0.926	0.902	0.902
Average Precision	0.937	0.94	0.954	0.951

Table 7. 2: Average accuracy and precision in multi-label classification

When analyzing the values obtained for individual evaluation measures (Table 7.1 and Table 7.2), audio based multi-label classification depicts extremely good performance. Accuracy and precision values obtained for each classification algorithm are greater than 0.9 in almost all the instances, which is much close to the ideal classification performance of obtaining 1.

Combined and Graphical Evaluation Measures

	BR	LP	MLKNN	RAKEL
F-measure	0.97	0.961	0.948	0.949
Hamming Loss	0.059	0.073	0.098	0.098
AUC	0.5	0.509	0.439	0.5

Table 7. 3: F-measure, Hamming loss and AUC in multi-label classification

Analysis of the values obtained for F-measure, hamming loss and AUC as depicted in Table 7.3 provides contradicting views regarding the performance of multi-label classifiers. F-measure is much satisfactory since it's close to 1. While obtaining a value close to 0 for hamming loss is considered positive, the values obtained for Area under the Curve (AUC) made us question the validity of other measures. As specified in Table 4.2, obtaining a value between 0.5 and 0.6 is considered as a failure with relation to classifier performance. This compelled us to analyze the dataset more closely to realize which evaluation measures have provided with accurate review regarding performance of multi-label classifiers.

For studying this issue with multi-label classification, one emotion class (among the 25 class attributes) was retained. The considered emotion class was Joyful_danceable.

When merely the Joyful_danceable class attribute is retained from File II (using which multi-label classification was executed), the data distribution among positive and negative classes is depicted as shown in Figure 7.1.

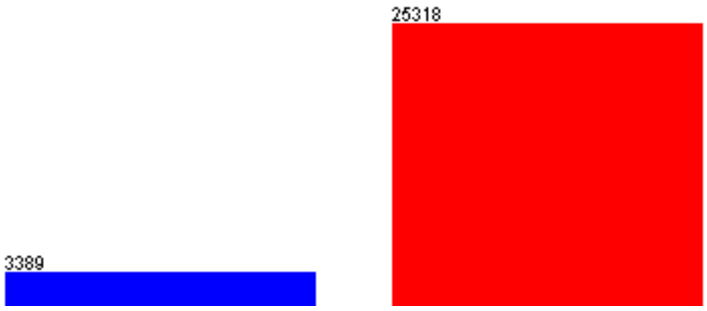

Figure 7. 1: Data distribution for joyful_danceable class in File II

Number of songs with which joyful_danceable emotion is associated is 3389 whereas 25 318 songs don't have this emotion associated with them. Hence, the number of data points for negative class is immensely greater than the number of records for positive class, thus resulting in significant class imbalance. This issue was observed with relation to rest of the 24 emotion classes as well. As described in section 4.6.3, Multi-label classification algorithms rely partially (LP, MLKNN, RAKEL) or entirely on the individual classifications using each class attribute separately. The final result is often a combination of results obtained for each individual binary classification. Therefore, evaluating the base classifier's (SMO) performance on a single class attribute allows one to understand the results depicted for multi-label classification. Table 7.4 depicts values obtained for evaluation metrics, when SMO was executed using joyful_danceable attribute only. Based on which class we consider to be positive (positive as positive or negative as positive in confusion matrix) evaluation measures give different class-wise values except for accuracy and AUC.

	Individual evaluation measures				Combined and graphical measures	
	Accuracy	**Precision**	**Recall**	**FP-rate**	**F-measure**	**AUC**
Positive	0.883	0	0	0	0	0.5
Negative		0.883	1	1	0.938	0.5

Table 7. 4: Evaluation of SMO performance on joyful_danceable class

FP-rate: False Positive rate

It could be easily deduced by studying values in Table 7.4, that the base classifier depicts commendable results for negative class only. This phenomenon occurs due to the classifier learning only from the majority class data. Accuracy is a misleading metric in the presence of extreme class imbalance, since commendable accuracy is obtained due to majority class being correctly predicted. This problem with base classification for each of the 25 class attributes has dire effects on multi-label classification as well. A solution for this would be dealing with class imbalance of the individual attributes.

Of the many mechanisms discussed for handling class imbalance (section 4.7), oversampling and undersampling are not effective in the multi-label scenario. This is caused by the representation of attribute values in multi-label classification (i.e. when data is undersampled to achieve class balance for one class attribute, it would be detrimental to another class attribute). Thus, merely bagging, boosting and cost sensitive learning were attempted with the base classifier for joyful_danceable class.

	Individual evaluation measures				Combined and graphical measures	
	Accuracy	**Precision**	**Recall**	**FP-rate**	**F-measure**	**AUC**
Positive	0.883	0	0	0	0	0.5
Negative		0.883	1	1	0.938	0.5

Table 7. 5: Evaluation of SMO performance with bagging on joyful_danceable class

As depicted in Table 7.5, application of bagging technique has yielded no improvement on base classifier performance. Thus, bagging technique would not be beneficial for multi-label classifier either.

	Individual evaluation measures				**Combined and graphical measures**	
	Accuracy	**Precision**	**Recall**	**FP-rate**	**F-measure**	**AUC**
Positive	0.883	0	0	0	0	0.673
Negative		0.883	1	1	0.938	0.673

Table 7. 6: Evaluation of SMO performance with boosting on joyful_danceable class

According to Table 7.6, boosting technique has effected slight improvement on the AUC measure. However, it has failed to improve other accuracy measures with relation to positive class. Hence application of boosting with multi-label too was disregarded.

The third mechanism of handling class imbalance was cost sensitive learning. Since a natural 'cost' for misclassifications was unknown, an attempt was made to realize this value empirically. Despite empirical learning of cost measure being feasible to a certain extent, in binary classification, such learning becomes impractical when number of class attribute values increase. Table 7.7 depicts performance evaluation of base classifier for different costs with respect to joyful_danceable class.

Due to class imbalance in this scenario, positives often get misclassified as negatives, rather than negatives as positives. Thus the cost of False Negatives (FN) must be greater than cost incurred by False Positives (FP).

Cost ratio FN/FP		Individual evaluation measures				Combined and graphical measures	
		Accuracy	**Precision**	**Recall**	**FP-rate**	**F-measure**	**AUC**
5	**Positive**	0.799	0.262	0.396	0.147	0.316	0.624
	Negative		0.914	0.853	0.604	0.882	0.624
6	**Positive**	0.747	0.236	0.516	0.222	0.323	0.647
	Negative		0.924	0.778	0.484	0.845	0.647
7	**Positive**	0.69	0.217	0.619	0.296	0.321	0.661
	Negative		0.933	0.704	0.381	0.802	0.661
8	**Positive**	0.637	0.199	0.695	0.37	0.31	0.662
	Negative		0.94	0.63	0.305	0.754	0.662
9	**Positive**	0.59	0.19	0.764	0.433	0.304	0.666
	Negative		0.948	0.567	0.236	0.71	0.666
10	**Positive**	0.54	0.179	0.809	0.493	0.293	0.658
	Negative		0.952	0.507	0.191	0.662	0.658
15	**Positive**	0.34	0.146	0.944	0.729	0.254	0.608
	Negative		0.973	0.271	0.056	0.424	0.608
20	**Positive**	0.11	0.117	1	1	0.21	0.5
	Negative		0	0	0	0	0.5

Table 7. 7: Performance for different cost ratios

As depicted in Table 7.7, the best cost ratio of FN:FP is 10:1 with respect to AUC measure. However adhering to this technique was disregarded due to contradiction among other measures, where the best measure for each class has been highlighted.

As depicted with empirical evidence under this section, class imbalance within each class attribute has dire impact on the results depicted for holistic multi-label classification. Multi-label evaluation metrics, except AUC, fail to reflect this matter and could be considered misleading. Due to problems inherent to multi-label classification approaches, the experiment was not further pursued.

7.3. Tuning Parameters of Classifiers

Subsequent to the failed attempt at multi-label classification of music into emotions, the single-label approach was attempted. Prior to commencing with classification it was necessary to tune parameters of certain classifiers applied in this phase. Parameter tuning was executed as elaborated under section 4.6.4. Of the four single label classification techniques applied, it was necessary to tune parameters of three, which are C4.5, SVM (SMO) and Random Forest. Naïve Bayes classifier is a probabilistic method with no parameters associated. Since Weka was used to conduct single-label experiments, the same was used for parameter tuning.

Parameter Tuning of C4.5.

Weka offers two pruning mechanism; C4.5 pruning and reduced error pruning. The pruning mechanism could be altered via setting the parameter 'reduced error pruning' to either true or false. The former would apply reduced error pruning on the tree whereas latter would execute C4.5 pruning. Root Mean Squared Error (RMSE) value was considered for selecting best parameter.

Parameter value	RMSE	Time to build model (sec)	Size of tree
Reduced error pruning = false	0.2365	5.01	11337
Reduced error pruning = true	0.1995	2.41	1225

Table 7. 8: C4.5 parameter tuning

Based on the values obtained for each parameter as depicted in Table 7.8, using reduced error pruning was opted for due to RMSE being comparatively less. Had the C4.5 pruning been chosen, we would have had to tune another parameter 'confidence factor'. However the parameter is associated with C4.5 pruning only, thus making it unnecessary for it to be tuned since reduced error pruning was chosen.

Parameter Tuning of SMO

Sequential Minimal Optimization (SMO) implementation of SVM was applied in this experiment. Among the many kernel functions offered four kernels which support classification of numerical attributes were considered since audio, lyric attributes are continuous. Table 7.9 RMSE values obtained for each kernel.

Kernel	RMSE	Time to build model (sec)
Polynomial kernel	0.1915	11.06
Normalized polynomial kernel	0.1915	1039.61
Radial basis function kernel	0.1918	1441.28
Pearson VII function-based universal kernel	0.1914	1225.58

Table 7. 9: SMO parameter tuning

Analysis of RMSE values and time to build model in Table 7.9 helped us determine applying polynomial kernel would yield best results in minimal time.

Parameter Tuning of Random Forest

Number of trees used in random forest has considerable impact on its overall performance. Apart from RMSE, Out of Bag Error (OOBE) is another measure which could be utilized for parameter tuning in Random forest, as specified in section 4.6.4.

Number of trees	OOBE	RMSE	Time to build model (sec)
60	0.7747	0.1866	42.36
80	0.7697	0.1862	54.16
100	0.7637	0.1859	69.67
200	0.7545	0.1854	137.5
500	0.7456	1.852	336.5

Table 7. 10: Random Forest parameter tuning

As depicted in Table 7.10, both OOBE and RMSE are reduced when increasing number of trees. Thus 500 trees were utilized in each classification instance using Random Forests technique.

7.4. Single-label Classification Attempt for 25 Emotion Classes

Prediction of the most representative emotion of a music piece was attempted at this phase. Figure 5.6 depicts distribution of data among the 25 emotion classes identified in Chapter 6. The initial classification attempt was for non-discretized audio attributes, for which File III specified under section 5.4 was used.

C4.5 classification for non-discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.05	0.023	0.126	0.05	0.072	0.616
witty	0.164	0.054	0.23	0.164	0.192	0.597
calming_melancholic_romantic	0.256	0.086	0.214	0.256	0.233	0.647
uplifting_feelgood	0	0.001	0	0	0	0.522
aggressive_angry	0	0.001	0	0	0	0.598
moving_inspiring	0	0	0	0	0	0.496
sensual	0.437	0.309	0.256	0.437	0.323	0.571
haunting_dark_depressing	0	0	0	0	0	0.533
angst	0	0.001	0	0	0	0.507
sick	0	0.001	0	0	0	0.601
makes	0	0.001	0	0	0	0.512
heartbreaking_bittersweet	0	0	0	0	0	0.534
crazy	0.004	0.001	0.04	0.004	0.007	0.563
exciting	0	0.001	0	0	0	0.618
creepy	0	0.001	0	0	0	0.536
ethereal_dreamy_soothing	0	0	0	0	0	0.489
cool	0.326	0.258	0.23	0.326	0.27	0.541
energetic_powerful	0	0	0	0	0	0.621
soulful	0	0.001	0	0	0	0.559
hypnotic	0.005	0.001	0.077	0.005	0.009	0.51
sentimental	0.006	0.002	0.054	0.006	0.01	0.554
psychedelic	0.326	0.134	0.249	0.326	0.282	0.632
lonely	0	0	0	0	0	0.516
spiritual	0	0.001	0	0	0	0.499
nostalgic	0.049	0.035	0.123	0.049	0.07	0.55

Table 7. 11: C4.5 classification: non-discretized: 25 emotion classes

Overall accuracy obtained for C4.5 classification was 0.23. (C & G: Combined and graphical evaluation metrics)

Naïve Bayes classification for non-discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.126	0.04	0.171	0.126	0.145	0.693
witty	0.116	0.029	0.283	0.116	0.164	0.65
calming_melancholic_romantic	0.444	0.151	0.212	0.444	0.287	0.729
uplifting_feelgood	0	0	0	0	0	0.573
aggressive_angry	0	0.001	0	0	0	0.762
moving_inspiring	0	0	0	0	0	0.737
sensual	0.326	0.188	0.297	0.326	0.311	0.611
haunting_dark_depressing	0	0	0	0	0	0.426
angst	0	0	0	0	0	0.649
sick	0.017	0.008	0.019	0.017	0.018	0.728
makes	0	0	0	0	0	0.572
heartbreaking_bittersweet	0	0	0	0	0	0.523
crazy	0	0.002	0	0	0	0.636
exciting	0.011	0.005	0.021	0.011	0.015	0.712
creepy	0.003	0	0.333	0.003	0.005	0.592
ethereal_dreamy_soothing	0	0	0	0	0	0.544
cool	0.307	0.238	0.233	0.307	0.265	0.569
energetic_powerful	0.095	0.015	0.048	0.095	0.064	0.753
soulful	0	0	0	0	0	0.66
hypnotic	0	0	0	0	0	0.592
sentimental	0.003	0.002	0.03	0.003	0.005	0.667
psychedelic	0.448	0.19	0.242	0.448	0.314	0.701
lonely	0	0	0	0	0	0.558
spiritual	0.005	0.001	0.083	0.005	0.01	0.6
nostalgic	0.022	0.018	0.109	0.022	0.037	0.594

Table 7. 12: Naïve Bayes classification: non-discretized: 25 emotion classes

Overall accuracy obtained for Naïve Bayes was 0.23

Random Forest classification for non-discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.037	0.014	0.148	0.037	0.06	0.716
witty	0.175	0.036	0.322	0.175	0.226	0.676
calming_melancholic_romantic	0.256	0.074	0.242	0.256	0.249	0.753
uplifting_feelgood	0	0	0	0	0	0.541
aggressive_angry	0	0	0	0	0	0.746
moving_inspiring	0	0	0	0	0	0.662
sensual	0.46	0.313	0.263	0.46	0.335	0.619
haunting_dark_depressing	0	0	0	0	0	0.483
angst	0	0	0	0	0	0.615
sick	0.006	0	0.2	0.006	0.011	0.661
makes	0	0	0	0	0	0.541
heartbreaking_bittersweet	0	0	0	0	0	0.58
crazy	0.012	0	0.5	0.012	0.023	0.598
exciting	0	0	0	0	0	0.713
creepy	0	0	0	0	0	0.561
ethereal_dreamy_soothing	0	0	0	0	0	0.465
cool	0.381	0.287	0.238	0.381	0.293	0.575
energetic_powerful	0	0	0	0	0	0.738
soulful	0	0	0	0	0	0.67
hypnotic	0	0	0	0	0	0.617
sentimental	0	0.001	0	0	0	0.632
psychedelic	0.381	0.133	0.28	0.381	0.323	0.723
lonely	0	0	0	0	0	0.572
spiritual	0	0	0	0	0	0.591
nostalgic	0.057	0.033	0.149	0.057	0.083	0.606

Table 7. 13: Random Forest classification: non-discretized: 25 emotion classes

Overall accuracy obtained for Random Forest was 0.25

SVM (SMO) classification for non-discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0	0	0	0	0	0.699
witty	0.003	0	0.5	0.003	0.007	0.659
calming_melancholic_romantic	0.112	0.032	0.243	0.112	0.154	0.742
uplifting_feelgood	0	0	0	0	0	0.558
aggressive_angry	0	0	0	0	0	0.784
moving_inspiring	0	0	0	0	0	0.543
sensual	0.553	0.37	0.267	0.553	0.36	0.605
haunting_dark_depressing	0	0	0	0	0	0.472
angst	0	0	0	0	0	0.644
sick	0	0	0	0	0	0.736
makes	0	0	0	0	0	0.533
heartbreaking_bittersweet	0	0	0	0	0	0.445
crazy	0	0	0	0	0	0.653
exciting	0	0	0	0	0	0.715
creepy	0	0	0	0	0	0.592
ethereal_dreamy_soothing	0	0	0	0	0	0.5
cool	0.443	0.356	0.227	0.443	0.3	0.558
energetic_powerful	0	0	0	0	0	0.768
soulful	0	0	0	0	0	0.68
hypnotic	0	0	0	0	0	0.6
sentimental	0	0	0	0	0	0.68
psychedelic	0.428	0.149	0.281	0.428	0.339	0.71
lonely	0	0	0	0	0	0.563
spiritual	0	0	0	0	0	0.635
nostalgic	0	0	0	0	0	0.553

Table 7. 14: SVM classification: non-discretized: 25 emotion classes

Overall accuracy obtained for SVM(SMO) was 0.25

Analysis of Tables 7.11, 7.12, 7.13 and 7.14 yields that single label classification for 25 emotion classes using non-discretized audio attributes doesn't produce satisfactory results for most of the emotion classes. TP rate, Precision, Recall and F measure are 0 for majority of the emotion classes for all four classifiers. The effect of discretizing audio attributes was thus analyzed in the next experiment. Audio feature discretization was discussed under section 5.2.6.

C4.5 classification for discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.035	0.016	0.127	0.035	0.055	0.655
witty	0.142	0.05	0.216	0.142	0.171	0.612
calming_melancholic_romantic	0.25	0.085	0.212	0.25	0.23	0.683
uplifting_feelgood	0	0	0	0	0	0.498
aggressive_angry	0	0	0	0	0	0.6
moving_inspiring	0	0	0	0	0	0.463
sensual	0.461	0.319	0.26	0.461	0.333	0.594
haunting_dark_depressing	0	0	0	0	0	0.473
angst	0	0	0	0	0	0.577
sick	0	0	0	0	0	0.649
makes	0	0.001	0	0	0	0.515
heartbreaking_bittersweet	0	0	0	0	0	0.489
crazy	0	0	0	0	0	0.588
exciting	0	0	0	0	0	0.636
creepy	0	0	0	0	0	0.535
ethereal_dreamy_soothing	0	0	0	0	0	0.483
cool	0.357	0.277	0.233	0.357	0.282	0.555
energetic_powerful	0	0	0	0	0	0.665
soulful	0.003	0	0.1	0.003	0.005	0.609
hypnotic	0	0	0	0	0	0.537
sentimental	0	0	0	0	0	0.582
psychedelic	0.343	0.133	0.26	0.343	0.296	0.674
lonely	0	0	0	0	0	0.539
spiritual	0	0	0	0	0	0.512
nostalgic	0.039	0.024	0.144	0.039	0.062	0.573

Table 7. 15: C4.5 classification: discretized: 25 emotion classes

Overall accuracy obtained for C4.5 was 0.238

Naïve Bayes classification for discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.089	0.035	0.144	0.089	0.11	0.684
witty	0.126	0.039	0.238	0.126	0.164	0.644
calming_melancholic_romantic	0.428	0.149	0.208	0.428	0.28	0.723
uplifting_feelgood	0	0	0	0	0	0.584
aggressive_angry	0	0	0	0	0	0.733
moving_inspiring	0	0	0	0	0	0.708
sensual	0.355	0.206	0.295	0.355	0.322	0.616
haunting_dark_depressing	0	0	0	0	0	0.537
angst	0	0	0	0	0	0.635
sick	0	0	0	0	0	0.721
makes	0	0	0	0	0	0.561
heartbreaking_bittersweet	0	0	0	0	0	0.542
crazy	0	0	0	0	0	0.635
exciting	0	0	0	0	0	0.719
creepy	0	0	0	0	0	0.574
ethereal_dreamy_soothing	0	0	0	0	0	0.718
cool	0.338	0.247	0.244	0.338	0.283	0.571
energetic_powerful	0	0	0	0	0	0.764
soulful	0	0	0	0	0	0.652
hypnotic	0	0	0	0	0	0.605
sentimental	0	0	0	0	0	0.675
psychedelic	0.448	0.205	0.229	0.448	0.303	0.697
lonely	0	0	0	0	0	0.599
spiritual	0	0	0	0	0	0.618
nostalgic	0.009	0.008	0.106	0.009	0.017	0.588

Table 7. 16: Naïve Bayes classification: discretized: 25 emotion classes

Overall accuracy obtained for Naïve Bayes was 0.24

Random Forest classification for discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.052	0.028	0.109	0.052	0.071	0.632
witty	0.142	0.063	0.181	0.142	0.159	0.607
calming_melancholic_romantic	0.256	0.087	0.212	0.256	0.232	0.673
uplifting_feelgood	0	0.002	0	0	0	0.517
aggressive_angry	0	0.001	0	0	0	0.583
moving_inspiring	0	0	0	0	0	0.547
sensual	0.332	0.236	0.255	0.332	0.289	0.57
haunting_dark_depressing	0	0	0	0	0	0.475
angst	0	0.002	0	0	0	0.538
sick	0	0.002	0	0	0	0.624
makes	0.003	0.004	0.012	0.003	0.005	0.491
heartbreaking_bittersweet	0	0	0	0	0	0.504
crazy	0	0.003	0	0	0	0.581
exciting	0	0.001	0	0	0	0.633
creepy	0.005	0.004	0.023	0.005	0.009	0.529
ethereal_dreamy_soothing	0	0	0	0	0	0.487
cool	0.333	0.274	0.223	0.333	0.267	0.535
energetic_powerful	0	0.001	0	0	0	0.66
soulful	0.014	0.005	0.046	0.014	0.021	0.583
hypnotic	0.005	0.003	0.016	0.005	0.007	0.564
sentimental	0.006	0.004	0.027	0.006	0.01	0.577
psychedelic	0.32	0.136	0.243	0.32	0.276	0.664
lonely	0	0.001	0	0	0	0.575
spiritual	0.005	0.003	0.02	0.005	0.008	0.555
nostalgic	0.088	0.062	0.125	0.088	0.103	0.56

Table 7. 17: Random Forest classification: discretized: 25 emotion classes

Overall accuracy obtained for Random Forest was 0.21

SVM (SMO) classification for discretized audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0	0	0	0	0	0.674
witty	0	0	0	0	0	0.637
calming_melancholic_romantic	0.236	0.085	0.202	0.236	0.218	0.71
uplifting_feelgood	0	0	0	0	0	0.565
aggressive_angry	0	0	0	0	0	0.736
moving_inspiring	0	0	0	0	0	0.588
sensual	0.457	0.357	0.238	0.457	0.313	0.58
haunting_dark_depressing	0	0	0	0	0	0.463
angst	0	0	0	0	0	0.621
sick	0	0	0	0	0	0.721
makes	0	0	0	0	0	0.535
heartbreaking_bittersweet	0	0	0	0	0	0.595
crazy	0	0	0	0	0	0.627
exciting	0	0	0	0	0	0.706
creepy	0	0	0	0	0	0.585
ethereal_dreamy_soothing	0	0	0	0	0	0.476
cool	0.474	0.394	0.221	0.474	0.302	0.556
energetic_powerful	0	0	0	0	0	0.764
soulful	0	0	0	0	0	0.673
hypnotic	0	0	0	0	0	0.614
sentimental	0	0	0	0	0	0.678
psychedelic	0.286	0.093	0.295	0.286	0.29	0.698
lonely	0	0	0	0	0	0.594
spiritual	0	0	0	0	0	0.625
nostalgic	0	0	0	0	0	0.565

Table 7. 18: SVM classification: discretized: 25 emotion classes

Overall accuracy obtained for SVM was 0.23

Tables 7.15, 7.16, 7.17 and 7.18 depict no considerable improvement on the classification accuracies obtained for discretized classification. AUC measure further attests to this fact since that measure has failed to exceed 0.7 in most instances. A potential reason for this phenomenon would be the extreme class imbalance of dataset used for single-label classification as depicted in Figure 5.6.

When attempting to handle the class imbalance issue in single-label scenario, application of techniques discussed under section 4.7 were attempted. However, application of undersampling and oversampling on the complete dataset, using all 25 emotion classes was not feasible, since it resulted in retention of either too few data points or too many replications. Hence, the emotion classes which comprised of more than 1000 songs alone were used to conduct this experiment. Therefore rest of the experiments were based on emotion classes depicted in Table 7.19.

Seven emotion classes with more than 1000 records per class
joyful_danceable
witty
calming_melancholic_romantic
sensual
cool
psychedelic
nostalgic

Table 7. 19: Seven emotion classes

Data distribution among these seven classes could be graphically depicted as shown in Figure 7.2. The order of bars in histogram corresponds to order in Table 7.19.

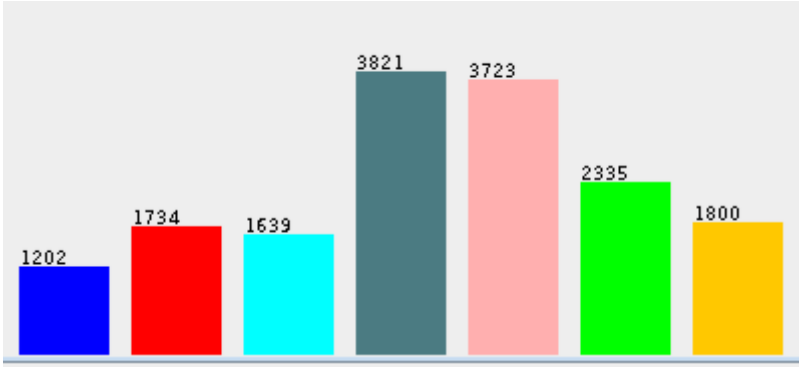

Figure 7. 2: Data distribution among seven emotion classes

Undersampling and oversampling were applied on this data space as elaborated under section 4.7. Weka implements undersampling as *SpreadSubsample* whereas oversampling is implemented as SMOTE.

Application of *SpreadSubsample* algorithm randomly removes data points from majority classes to equal the number of records in minority class. Thus 1202 records were retained from each class subsequent to application of undersampling. Figure 7.3 depicts distribution of data subsequent to application of undersampling.

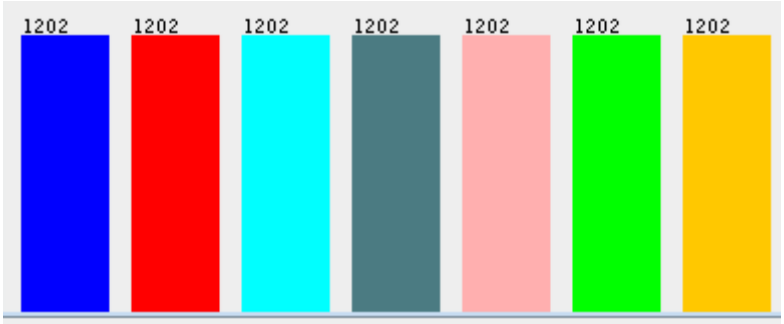

Figure 7. 3: Data distribution among seven emotion classes after undersampling

SMOTE algorithm was applied to oversample the data. The mechanism followed in *SMOTE* is the creation of synthetic records to deal with minority classes. The minority class is oversampled by specified percentage in this method. Since the majority class comprised of 3821 data points as depicted in Figure 7.4, our target was to achieve resampling of data in other classes, such that they comprised of records close to this number. The appropriate % was thus defined for each class as depicted in Table 7.20.

Class	**Original no of records**	**Selected %**	**Resultant number of records**
joyful_danceable	1202	200	3606
witty	1734	100	3468
Calming_melancholic_romantic	1639	100	3278
psychedelic	2335	50	3502
nostalgic	1800	100	3600

Table 7. 20: Application of SMOTE

The resultant dataset subsequent to applying SMOTE is depicted in Figure 7.4

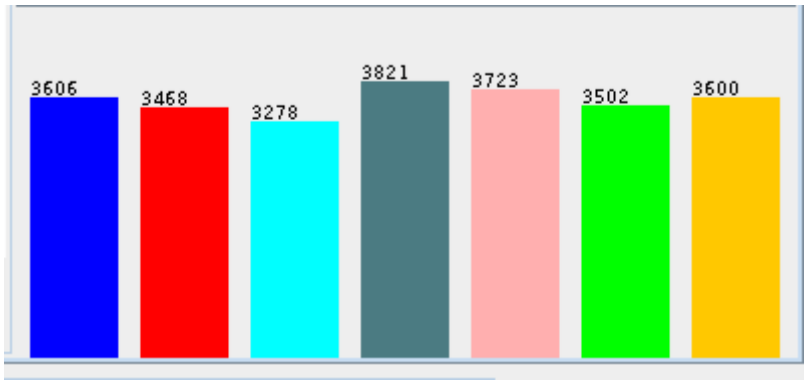

Figure 7. 4: Data distribution among seven emotion classes after oversampling

Experiments were conducted on this dataspace to realize whether application of class balancing techniques helped yield better results.

7.5. Single-label Classification Attempt for 7 Emotion Classes

The initial experiments were conducted on the data space depicted in Figure 7.2, which was utilized as baseline for evaluating implications of using class imbalance handling techniques. All these experiments utilized non-discretized audio attributes, since discretization failed to yield much improvement in classification accuracy in former experiment.

C4.5 classification for audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.058	0.03	0.135	0.058	0.081	0.609
witty	0.165	0.053	0.271	0.165	0.205	0.604
calming_melancholic_romantic	0.254	0.084	0.254	0.254	0.254	0.66
sensual	0.424	0.3	0.303	0.424	0.353	0.574
cool	0.328	0.251	0.28	0.328	0.302	0.544
psychedelic	0.335	0.138	0.29	0.335	0.311	0.641
nostalgic	0.049	0.037	0.142	0.049	0.073	0.548

Table 7. 21: C4.5 classification: 7 emotion classes

Overall accuracy obtained for C4.5 was 0.27

Naïve Bayes classification for audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.137	0.048	0.186	0.137	0.158	0.691
witty	0.117	0.029	0.324	0.117	0.172	0.657
calming_melancholic_romantic	0.451	0.145	0.258	0.451	0.329	0.735
sensual	0.334	0.192	0.349	0.334	0.341	0.62
cool	0.335	0.236	0.297	0.335	0.315	0.588
psychedelic	0.454	0.19	0.286	0.454	0.351	0.707
nostalgic	0.026	0.018	0.147	0.026	0.044	0.594

Table 7. 22: Naïve Bayes classification: 7 emotion classes

Overall accuracy obtained for Naïve Bayes was 0.29

Random Forest classification for audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.047	0.018	0.169	0.047	0.073	0.7
witty	0.173	0.04	0.342	0.173	0.23	0.676
calming_melancholic_romantic	0.251	0.072	0.282	0.251	0.266	0.753
sensual	0.446	0.308	0.308	0.446	0.365	0.613
cool	0.377	0.275	0.289	0.377	0.327	0.584
psychedelic	0.367	0.124	0.333	0.367	0.349	0.723
nostalgic	0.063	0.035	0.182	0.063	0.094	0.597

Table 7. 23: Random Forest classification: 7 emotion classes

Overall accuracy obtained for Random Forest was 0.29

SVM (SMO) classification for audio attributes

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0	0	0	0	0	0.686
witty	0.001	0	0.167	0.001	0.001	0.662
calming_melancholic_romantic	0.106	0.029	0.293	0.106	0.156	0.746
sensual	0.552	0.373	0.313	0.552	0.399	0.604
cool	0.446	0.342	0.279	0.446	0.344	0.567
psychedelic	0.43	0.141	0.337	0.43	0.378	0.717
nostalgic	0	0	0	0	0	0.548

Table 7. 24: SVM classification: 7 emotion classes

Overall accuracy obtained for SVM(SMO) was 0.3

Analysis of results depicted in Table 7.21 to Table 7.24 convey that except in application of SVM, the other classifiers haven't shown extreme poor performance with getting 0 values for TP rate, F-measure, etc. However the performance of 7 class scenario cannot be compared with that of 25 class instance, due to varying number of classes and data points. However, evaluation of ROC area (AUC) yield that generic classification (with no mechanism to handle class imbalance) is not commendable in the 7 class experiment as well.

The next series of experiments dealt with the undersampled dataset depicted in figure 7.3.

C4.5 classification for audio attributes: undersampled dataset

Class	**Individual metrics**				**C & G**	
	TP Rate	**FP Rate**	**Precision**	**Recall**	**F Measure**	**ROC Area**
joyful_danceable	0.384	0.169	0.275	0.384	0.32	0.638
witty	0.239	0.12	0.249	0.239	0.244	0.597
calming_melancholic_romantic	0.418	0.166	0.296	0.418	0.347	0.667
sensual	0.16	0.099	0.212	0.16	0.182	0.55
cool	0.152	0.091	0.218	0.152	0.179	0.563
psychedelic	0.35	0.136	0.301	0.35	0.324	0.641
nostalgic	0.107	0.084	0.175	0.107	0.133	0.538

Table 7. 25: C4.5 classification: 7 emotion classes: undersampled

Overall accuracy obtained for c4.5 was 0.25

Naïve Bayes classification for audio attributes: undersampled dataset

Class	**Individual metrics**				**C & G**	
	TP Rate	**FP Rate**	**Precision**	**Recall**	**F Measure**	**ROC Area**
joyful_danceable	0.496	0.214	0.278	0.496	0.357	0.714
witty	0.169	0.053	0.349	0.169	0.228	0.655
calming_melancholic_romantic	0.542	0.194	0.318	0.542	0.4	0.741
sensual	0.127	0.066	0.242	0.127	0.167	0.602
cool	0.152	0.112	0.184	0.152	0.167	0.59
psychedelic	0.386	0.149	0.301	0.386	0.339	0.705
nostalgic	0.072	0.054	0.182	0.072	0.104	0.577

Table 7. 26: Naïve Bayes classification: 7 emotion classes: undersampled

Overall accuracy obtained for Naïve Bayes was 0.27

Random Forest classification for audio attributes: undersampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.411	0.16	0.3	0.411	0.347	0.733
witty	0.295	0.102	0.326	0.295	0.31	0.677
calming_melancholic_romantic	0.471	0.15	0.343	0.471	0.397	0.766
sensual	0.151	0.095	0.208	0.151	0.175	0.585
cool	0.171	0.097	0.227	0.171	0.195	0.585
psychedelic	0.391	0.129	0.336	0.391	0.362	0.719
nostalgic	0.131	0.097	0.184	0.131	0.153	0.594

Table 7. 27: Random Forest classification: 7 emotion classes: undersampled

Overall accuracy obtained for Random Forest was 0.28

SVM (SMO) classification for audio attributes: undersampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.389	0.136	0.323	0.389	0.353	0.725
witty	0.255	0.068	0.383	0.255	0.307	0.678
calming_melancholic_romantic	0.606	0.203	0.332	0.606	0.429	0.75
sensual	0.146	0.069	0.26	0.146	0.187	0.609
cool	0.207	0.12	0.224	0.207	0.215	0.589
psychedelic	0.473	0.176	0.31	0.473	0.374	0.716
nostalgic	0.049	0.041	0.166	0.049	0.076	0.586

Table 7. 28: SVM classification: 7 emotion classes: undersampled

Overall accuracy obtained for SVM (SMO) was 0.3

Comparison of Tables 7.24 and 7.28 help deduce that classification of joyful_danceable and nostalgic classes has improved with undersampling, when TP rate and ROC Area (AUC) are taken into account. However, not all classes have achieved fair classification performance (reference section 4.5.3) when AUC values of Table 7.25 to Table 7.28 are considered.

The next set of experiments were conducted on oversampled dataset depicted in Figure 7.4

C4.5 classification for audio attributes: oversampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.453	0.159	0.324	0.453	0.378	0.705
witty	0.295	0.111	0.3	0.295	0.297	0.638
calming_melancholic_romantic	0.401	0.13	0.318	0.401	0.354	0.689
sensual	0.207	0.109	0.255	0.207	0.228	0.575
cool	0.154	0.098	0.215	0.154	0.179	0.55
psychedelic	0.346	0.131	0.3	0.346	0.321	0.651
nostalgic	0.163	0.095	0.223	0.163	0.188	0.584

Table 7. 29: C4.5 classification: 7 emotion classes: oversampled

Overall accuracy obtained for c4.5 was 0.28

Naïve Bayes classification for audio attributes: oversampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.563	0.238	0.285	0.563	0.378	0.738
witty	0.173	0.055	0.336	0.173	0.228	0.673
calming_melancholic_romantic	0.551	0.185	0.31	0.551	0.397	0.758
sensual	0.162	0.056	0.344	0.162	0.22	0.627
cool	0.12	0.078	0.212	0.12	0.153	0.592
psychedelic	0.399	0.141	0.316	0.399	0.353	0.716
nostalgic	0.103	0.073	0.192	0.103	0.134	0.607

Table 7. 30: Naïve Bayes classification: 7 emotion classes: oversampled

Overall accuracy obtained for Naïve Bayes was 0.29

Random Forest classification for audio attributes: oversampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.615	0.135	0.434	0.615	0.509	0.848
witty	0.401	0.077	0.455	0.401	0.426	0.789
calming_melancholic_romantic	0.545	0.117	0.412	0.545	0.469	0.83
sensual	0.252	0.097	0.32	0.252	0.282	0.66
cool	0.198	0.088	0.281	0.198	0.232	0.631
psychedelic	0.465	0.114	0.4	0.465	0.43	0.782
nostalgic	0.299	0.081	0.385	0.299	0.337	0.741

Table 7. 31: Random Forest classification: 7 emotion classes: oversampled

Overall accuracy obtained for Random Forest was 0.39

SVM (SMO) classification for audio attributes: oversampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.409	0.135	0.338	0.409	0.37	0.737
witty	0.24	0.063	0.379	0.24	0.294	0.672
calming_melancholic_romantic	0.561	0.173	0.329	0.561	0.414	0.759
sensual	0.227	0.094	0.303	0.227	0.26	0.628
cool	0.203	0.117	0.233	0.203	0.217	0.58
psychedelic	0.471	0.164	0.319	0.471	0.381	0.721
nostalgic	0.089	0.058	0.206	0.089	0.124	0.599

Table 7. 32: SVM classification: 7 emotion classes: oversampled

Overall accuracy obtained for SVM(SMO) was 0.3

Comparison of undersampling and oversampling methods help deduce the fact that oversampling outperforms undersampling with respect to overall accuracy in all instances except in SVM. Comparatively, the most commendable accuracy has been obtained for Random Forest, as depicted in Table 7.31. Execution of Random Forest on oversampled dataset has yielded 'good' results for joyful_danceable and calming_melancholic_romantic classes with relation to AUC measure (refer section 4.5.3). Except for 'sensual' and 'cool' classes, fair level of accuracy has been obtained for other emotion classes as well. Furthermore, the overall accuracy measure obtained for it is the best so far among attempted experiments for 7 classes. A random guess with relation to predicting a class would have yielded probabilistic accuracy of 1/7 = 0.14. Since 0.39 has been obtained with Random Forest + Oversampling, the accuracy of it is fairly good.

Since Random Forest classification + oversampling for audio attributes yielded fair performance for 5 out of 7 classes, the same was attempted for lyrics and hybrid features separately. Tables 7.33 and 7.34 depict results obtained for lyrics and hybrid classification respectively.

Random Forest classification for lyrics attributes: oversampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.4	0.121	0.328	0.4	0.412	0.524
witty	0.245	0.012	0.113	0.245	0.127	0.532
calming_melancholic_romantic	0.423	0.25	0.249	0.423	0.391	0.6
sensual	0.52	0.2	0.244	0.52	0.204	0.56
cool	0.13	0.55	0.119	0.13	0.213	0.549
psychedelic	0.313	0.223	0.315	0.313	0.273	0.523
nostalgic	0.254	0.095	0.283	0.254	0.109	0.623

Table 7. 33: Random Forest classification for lyric attributes

Overall accuracy obtained for Random Forest was 0.2

Random Forest classification for hybrid attributes: oversampled dataset

Class	Individual metrics				C & G	
	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area
joyful_danceable	0.455	0.109	0.384	0.455	0.49	0.64
witty	0.3	0.21	0.181	0.3	0.241	0.61
calming_melancholic_romantic	0.49	0.3	0.264	0.49	0.4	0.656
sensual	0.602	0.209	0.251	0.602	0.319	0.6
cool	0.251	0.591	0.215	0.251	0.251	0.553
psychedelic	0.4	0.239	0.331	0.4	0.281	0.613
nostalgic	0.352	0.104	0.315	0.352	0.128	0.682

Table 7. 34: Random Forest classification for hybrid attributes

Overall accuracy obtained for Random Forest was 0.29

Comparison of results in Table 7.33 and Table7.34 with that of Table 7.31 yields that lyrics and Hybrid classification has not outperformed audio feature based classification in this research. Despite Random Forest + Oversampling mechanism not yielding fair accuracy for all the classes, it provides with the comparative best results for experiments conducted.

7.6. Summary

This chapter extensively analyzed the results obtained for classification experiments conducted. A series of experiments helped in concluding that Random Forest + Oversampling technique provides in yielding comparatively better results.

A summary of the experiments conducted are as follows

I. Multi-label classification experiment for BR, RAKEL, LP and MLKNN
II. Classification experiment for deducing underlying class imbalance problem in Multi-label classification
III. Single-label classification for Naïve Bayes, SVM (SMO), Random Forest and C4.5 algorithms: Non-discretized audio features and 25 emotion classes utilized
IV. Single-label classification for Naïve Bayes, SVM (SMO), Random Forest and C4.5 algorithms: Discretized audio features and 25 emotion classes utilized
V. Single-label classification for Naïve Bayes, SVM (SMO), Random Forest and C4.5 algorithms: Non-discretized audio features and 7 emotion classes utilized
VI. Single-label classification with undersampling for Naïve Bayes, SVM (SMO), Random Forest and C4.5 algorithms: Non-discretized audio features and 7 emotion classes utilized
VII. Single-label classification with oversampling for Naïve Bayes, SVM (SMO), Random Forest and C4.5 algorithms: Non-discretized audio features and 7 emotion classes utilized
VIII. Single-label classification with oversampling for Random Forest algorithm: Non-discretized lyric features and 7 emotion classes utilized
IX. Single-label classification with oversampling for Random Forest algorithm: Non-discretized hybrid features and 7 emotion classes utilized

The following chapter concludes this thesis specifying inferences based on research study and highlights future work.

Chapter 8 - Conclusion

Potential for music to evoke emotion in individuals has forever been an inexplicable, yet evident phenomenon. While numerous research work has been conducted to deduce whether the incurrence of emotion when listening to music is scientifically explicable, the literature survey conducted helped realize that there existed numerous limitations with existent mechanisms. One such limitation was usage of generic emotion models for predicting emotions in music. It has been observed that emotions incurred when listening to music takes a different form than generic emotions. For instance, the sadness evoked by music is not necessarily as intense as sadness evoked from real life experience. In fact, sadness (melancholic) belonged to the same cluster as calming and romantic in the music specific emotion model created based on the MSD dataset. Thus creation of a music specific emotion model for classification of music was viewed to be a positive aspect of this study.

A series of experiments were conducted to attempt classification of music into emotions recognized. The initial multi-label classification attempt helped deduce that multi-label classification of music was not reliable due to limitations inherent to those algorithms. Another array of single-label experiments were conducted whereas algorithms for handling class imbalance were applied. Of the numerous classification attempts, the best music emotion prediction model was provided as a combination of Random Forest with oversampling. Performance of this model was best for audio based classification.

Limitations and Future Work

The emotion model formed in this study was reliant on the assumption that emotion related tags associated with music express perceived emotion. This assumption may not always be valid since words such as 'love' could be conveying that the song is a love song, but not that the emotion was felt by a person. This limitation could be resolved by collection information from a number of people regarding what emotions they experienced when listening to songs from a dataset.

As opposed to the findings of former research work, hybrid classification did not outperform lyric based classification in this study. This could have potentially been caused by consideration of merely 5000 most popular lyric words for the study. As an alternative and possible future work, it's feasible to devise a mechanism which improves lyric/ hybrid classification via consideration of significant words from a song, rather than use of common words.

Reference

[1] W. J. Dowling, “The development of music perception and cognition,” in *Foundations of Cognitive Psychology*, 2002, pp. 481–502.

[2] Richard M. Ryan, “Moods of Energy and Tension that Motivate,” in *The Oxford Hand book of Human Motivation*, Peter E. Nathan, Ed. pp. 408–419.

[3] A. Kania, “The Philosophy of Music,” *The Stanford Encyclopedia of Philosophy*. 22-Oct-2014.

[4] P. Ekman and R. J. Davidson, “Moods, Emotions and Traits,” in *The Nature of Emotion*, 1994, p. 512.

[5] R. de Sousa, “Emotion,” *The Stanford Encyclopedia of Philosophy*. 03-Feb-2003.

[6] W. B. Cannon, “The James-Lange theory of emotions: A critical examination and an alternative theory,” *Am. J. Psychol.*, vol. 39, no. 1/4, pp. 106–124, 1927.

[7] D. G. Myers, “Theories of emotion,” in *Psychology*, 7th ed., no. 7, New York, NY: Worth Publishers, 2004, p. 500.

[8] Schachter, Stanley, and J. Singer, “Cognitive, social, and physiological determinants of emotional state,” *Psychol. Rev.*, vol. 69, no. 5, p. 379, 1962.

[9] R. Lazarus, “Cognition and Motivation in Emotion,” *Am. Psychol.*, vol. 46, no. 4, p. 352, 1991.

[10] H. Leventhal and K. R. Scherer, “The Relationship of Emotion Cognition: A Functional Approach to a Semantic Controversy,” *Cogn. Emot.*, vol. 1, no. 1, pp. 3–28, 1987.

[11] C. E. Izard, “Basic emotions, natural kinds, emotion schemas, and a new paradigm,” *Perspect. Psychol. Sci.*, vol. 2, pp. 260–280, 2007.

[12] M. Zentner, D. Grandjean, and K. R. Scherer, “Emotions evoked by the sound of music: characterization, classification, and measurement.,” *Emotion*, vol. 8, no. 4, pp. 494–521, Aug. 2008.

[13] Princeton University, “About WordNet - WordNet - About WordNet,” 2010. [Online]. Available: https://wordnet.princeton.edu/. [Accessed: 02-Aug-2015].

[14] C. Laurier, J. Grivolla, and P. Herrera, “Multimodal Music Mood Classification Using Audio and Lyrics,” *2008 Seventh Int. Conf. Mach. Learn. Appl.*, pp. 688–693, 2008.

[15] Q. Lu, X. Chen, D. Yang, and J. Wang, “Boosting for Multi-Modal Music Emotion Classification.,” *ISMIR*, no. Ismir, pp. 105–110, 2010.

[16] S. Davies, *Musical Meaning and Expression*. Cornell University Press, 1994, p. 417.

[17] K. R. Scherer and M. R. Zentner, “Emotional effects of music: production rules,” in *Music and Emotion: Theory and Research*, 2001, pp. 361–387.

[18] C. Radford, “Emotions and music: A reply to the cognitivists,” *J. Aesthestics Art Crit.*, vol. 47, pp. 69–76, 1989.

[19] A. Gabrielsson, “Emotion perceived and emotion felt: Same or different?,” *Music. Sci.*, vol. 5, no. 1, pp. 123–147, 2002.

[20] P. Evans and E. Schubert, “Quantification of Gabrielsson’s relationships between felt and expressed emotions in music,” *9th Int. Conf. Music Percept. Cogn.*, pp. 446–454, 2006.

[21] A. Kawakami, “Musical emotions: Perceived emotion and felt emotion in relation to musical structures,” *12th Int. Conf. Music Percept. Cogn.*, vol. 2, pp. 520–521, 2012.

[22] R. Paiva, “From Music Information Retrieval to Music Emotion Recognition,” 2012.

[23] J. E. Resnicow, P. Salovey, and B. H. Repp, “Is recognition of emotion in music performance an aspect of emotional intelligence?,” *Music Percept.*, vol. 22, no. 1, pp. 145–158, 2004.

[24] P. N. Juslin, “What does music express? Basic emotions and beyond.,” *Front. Psychol.*, vol. 4, p. 596, Jan. 2013.

[25] Y. Yang, Y. Lin, H. Cheng, and I. Liao, “Toward multi-modal music emotion classification,” *Adv. Multimed.*, pp. 1–10, 2008.

[26] C. Laurier, P. Herrera, M. Mandel, and D. Ellis, “Audio music mood classification using support vector machine,” *Music Inf. Retr. …*, pp. 2–4, 2007.

[27] E. Coutinho and N. Dibben, “Psychoacoustic cues to emotion in speech prosody and music,” *Cogn. Emot.*, pp. 1–27, 2013.

[28] C. Befus, C. Sanden, and J. Zhang, “Pyschoacoustic Feature Based Perceptual Segmentation,” pp. 22–29, 2010.

[29] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, “Multi-Label Classification of Music into Emotions.,” *ISMIR*, vol. 2008, pp. 307–315, 2008.

[30] K. MSocSc and W. Alexander, “Using Music as a Therapy Tool to Motivate Troubled Adolescents,” *Soc. Work Health Care*, vol. 3–4, no. 39, pp. 361–373, 2005.

[31] E. R. M. and J. Castet, “Emotional Response During Music Listening,” in *Guide to Brain-Computer Music Interfacing*, 2014.

[32] Y. Yang, C. Liu, and H. Chen, “Music emotion classification: a fuzzy approach,” *Proc. 14th Annu. ACM …*, pp. 81–84, 2006.

[33] P. Kanters, “Automatic mood classification for music,” Tilburg University, Netherlands, 2009.

[34] T. Li and M. Ogihara, “Detecting emotion in music.,” *ISMIR*, pp. 3–4, 2003.

[35] X. Hu and J. Downie, “Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata.,” in *ISMIR*, 2007.

[36] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, “Multi-Modal Music Emotion Recognition : A New Dataset , Methodology and Comparative Analysis,” pp. 1–13.

[37] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, “THE 2007 MIREX AUDIO MOOD CLASSIFICATION TASK : LESSONS LEARNED,” in *ISMIR*, 2008, pp. 462–467.

[38] X. Hu, M. Bay, and J. Downie, “Creating a Simplified Music Mood Classification Ground-Truth Set.,” *ISMIR*, pp. 3–4, 2007.

[39] X. Hu, J. Downie, and A. Ehmann, “Lyric text mining in music mood classification,” *Am. Music*, no. Ismir, pp. 411–416, 2009.

[40] G. Tzanetakis, “MARSYAS SUBMISSIONS TO MIREX 2012,” 2012.

[41] P. Mayring, “On Generalization in Qualitatively Oriented Research,” *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. 8, no. 3. 30-Sep-2007.

[42] J. Han and M. Kamber, *Data mining: concepts and techniques*, 2nd ed. Elsevier, 2006.

[43] D. Brain and G. Webb, “On the effect of data set size on bias and variance in classification learning,” in *Fourth Australian Knowledge Acquisition Workshop*, 1999, pp. 117–128.

[44] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC Music Database : Popular , Classical , and Jazz Music Databases,” in *Proceedings of the 3rd International Conference on Music Information Retrieval*, 2002, pp. 287–288.

[45] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.

[46] Y. Chathuranga and K. Jayaratne, “Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches,” *GSTF Int. J. Comput.*, vol. 3, no. 2, 2013.

[47] D. Wolff, S. Stober, A. Nürnberger, and T. Weyde, “A Systematic Comparison of Music Similarity Adaptation Approaches,” in *In Proceedings of the 13th International Society for Music Information Retrieval Conference*, 2012.

[48] M. Slaney, K. Weinberger, and W. White, “Learning a Metric for Music Similarity,” in *In Proceedings of the 9th International Conference on Music Information Retrieval*, 2008, pp. 313–318.

[49] J. Wang, H. Lee, H. Wang, and S. Jeng, “Learning the Similarity of Audio Music in Bag-of-Frames Representation from Tagged Music Data,” in *In Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011, pp. 85–90.

[50] E. Law, K. West, M. Mandel, M. Bay, and J. Downie, “Evaluation of Algorithms Using Games: The Case of Music Tagging.,” *ISMIR*, no. Ismir, pp. 387–392, 2009.

[51] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, “The million song dataset,” *ISMIR*, pp. 591–596, 2011.

[52] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman, “A large-scale evaluation of acoustic and subjective music-similarity measures,” *Comput. Music J.*, vol. 28, no. 2, pp. 63–76, Jun. 2004.

[53] D. Tidhar, G. Fazekas, S. Kolozali, and M. Sandler, “Publishing Music Similarity Features on the Semantic Web.,” *ISMIR*, no. Ismir, pp. 447–452, 2009.

[54] X. Hu and J. S. Downie, “Improving mood classification in music digital libraries by combining lyrics and audio,” *Proc. 10th Annu. Jt. Conf. Digit. Libr. - JCDL ’10*, p. 159, 2010.

[55] C. Mckay, “jAudio : Towards a standardized extensible audio music feature extraction system.”

[56] X. Huang, A. Acero, and H. Hon., *Spoken Language Processing: A guide to theory, algorithm, and system development.* Prentice Hall, 2001.

[57] R. D. C. Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria., 2008.

[58] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD …*, vol. 11, no. 1, pp. 10–18, 2009.

[59] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Int. Jt. Conf. Artificail Intelligece*, 1995.

[60] H. Modi and M. Panchal, "Experimental comparison of different problem transformation methods for multi-label classification using MEKA," *Int. J. Comput. Appl*, vol. 59, no. 15, pp. 10–15, 2012.

[61] POWERS and D.M.W, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[62] D. Measures and H. Clustering, "Distance Measures Hierarchical Clustering." pp. 1–59.

[63] T. Hill and P. Lewicki, *Statistics: Methods and Applications*, 2nd ed. StatSoft, Inc., 2007, p. 800.

[64] F. Murtagh and P. Legendre, "Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm," no. June, p. 20, Nov. 2011.

[65] A. Ng, "Support Vector Machines." pp. 1–25.

[66] I. Rish, "An empirical study of the naive Bayes classifier," pp. 41–46.

[67] N. V Chawla, "C4 . 5 and Imbalanced Data sets : Investigating the effect of sampling method , probabilistic estimate , and decision tree structure," 2003.

[68] L. Breiman, "Random Forests," pp. 1–33, 2001.

[69] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets : A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. December, 2006.

[70] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, 2001.

[71] D. M. Hawkins, *Identification of Outliers*. 1980.

[72] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," *ICML*, pp. 194–202, 1995.

[73] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley-Interscience, 1992.

[74] M. Dash and H. Liu, “Feature selection for classification,” *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 131–156, 1997.

[75] M. A. Hall, “Correlation-based Feature Selection for Machine Learning,” 1999.

[76] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” vol. 5, pp. 1205–1224, 2004.