

Understanding the relationship between user emotion and latent musical features

Master's Thesis

Aishwarya Shastry

Understanding the relationship between user emotion and latent musical features

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK DATA SCIENCE AND TECHNOLOGY

by

Aishwarya Shastry
born in Balaghat, India



Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
<http://wis.ewi.tudelft.nl>

Understanding the relationship between user emotion and latent musical features

Author: Aishwarya Shastry
Student id: 4743016
Email: a.shastri@student.tudelft.nl

Abstract

With the advent of Internet and resulting data boom, Recommender Systems have come to rescue by filtering the information available on the internet by providing us with relevant information. These systems come handy when one wants to listen to songs, watch movies or even buy products on the Internet. Primarily, these recommender systems used content based or collaborative filtering techniques to recommend items. More recent research has studied the importance of contextual features in recommender systems. Music preference has always been associated with the contextual feature emotion. However, few studies study the mood congruence effect in the domain of music recommender systems. The field of music emotion recognition also remains unexplored with recommendations being made with limited features.

This master thesis analyses the relationship between few latent musical features and user emotion through our interface Moodify. It is a music recommendation system that incorporates emotion in a user using emotion induction techniques and investigate the effect of their emotional state on satisfaction and unexpectedness when presented with songs curated to specific musical features. To achieve this, we analysed the enjoyment and unexpectedness ratings for recommendations specific to latent musical features for a given emotional state. We have been able to achieve some interesting results through this study which has been discussed later in this work.

Thesis Committee:

Chair: Prof. dr. ir. Geert-Jan Houben, Faculty EEMCS, TUDelft
University supervisor: Dr. Nava Tintarev , Faculty EEMCS, TUDelft
Committee Member: Dr. Cynthia Liem, Faculty EEMCS, TUDelft

Preface

First and foremost, I would like to express my gratitude to my thesis supervisor Dr Nava Tintarev for her constant support and guidance throughout my thesis. Her critical yet insightful feedback has helped shape my work and has developed in me the confidence of conducting independent research. She was never out of ideas and always encouraged me to do something novel which helped me to stay motivated during this journey. I would also like to thank all the thesis committee members for taking time to assess my work.

My journey towards a successful Master Thesis would not have been possible without the feedback of my fellow Epsilon members. Our meetings have always been a pleasure and are things I would always cherish. Our monthly discussions, feedback on each other's work, and suggestions have helped me conduct a better study. I would also like to thank the participants in my study who provided me with the necessary input to validate my research. A special thanks to the participants who expressed deep interest in the study and made me believe that I am doing something exciting.

Finally, I would like to thank my family and my friends for their unconditional love and support throughout my journey as a masters student. Thank you for always having my back.

This Master Thesis has been an invaluable experience and is something I would always treasure.

Aishwarya Shastry
Delft, The Netherlands
August 15th 2019

Contents

Preface	iii
Contents	v
List of Figures	ix
List of Tables	xiii
1 Introduction	1
1.1 Problem Statement	1
1.2 Research Objectives	3
1.3 Contributions	3
1.4 Thesis Outline	4
2 Literature Review	5
2.1 Introduction to Recommender Systems	5
2.2 Existing Context-Aware Recommender Systems	6
2.2.1 User Centric context-based approach	6
2.2.2 Environment context-based approach	6
2.3 Emotion in Recommender Systems	7
2.3.1 Why Emotion?	7
2.3.2 Interactive Emotion-Based Recommender Systems	8
2.4 Features for Music Recommendation	8
2.4.1 Features used in Existing Music Recommender Systems	8
2.4.2 Features in the field of Music Emotion Recognition	9
2.4.3 Emotion and music preference	10
2.5 Evaluation	10
2.5.1 Discovery oriented Metrics	10
Diversity	10
Novelty	11
Serendipity	11
Coverage	11
Others	11
2.5.2 Evaluation of Emotion-based Recommender Systems	12

2.6	Emotion- Definition, Models, Induction and Detection techniques . .	12
2.6.1	Emotion-Definition and Models	12
2.6.2	Emotion Induction Techniques	13
2.6.3	Emotion Detection in the domain of Recommender System . .	14
	Facial recognition	14
	Skin sensors	14
2.7	Conclusions	14
2.8	Research Gaps and Motivation	15
2.8.1	Research Gaps	15
2.8.2	Motivation	15
3	Data Exploration and Selection	17
3.1	Introduction	17
3.2	Dataset Selection	17
3.2.1	Selection of Features	17
3.2.2	Requirements from Dataset	18
3.2.3	Dataset Analysis	18
	Million Song Dataset	18
	Spotify	19
	DEAM	20
3.2.4	Final Decisions about Dataset	21
3.3	Extraction of Audio features	22
3.3.1	Audio Analysis	22
	Librosa	22
	Essentia	22
3.3.2	Features extracted from Essentia	23
	Loudness	23
	Danceability	23
	Beats Per Minute	23
	Valence values	24
3.3.3	Emotion Classification of DEAM dataset	24
3.4	Conclusions	24
4	System	27
4.1	System Overview	27
4.1.1	System Design	29
4.2	Initial Recommendation Phase	31
4.2.1	DEAM Dataset	31
	Data Merging	31
	Data Pre-processing	34
4.2.2	Initial Seed items	35
4.2.3	Recommendation Algorithm	36
	Item Item Similarity	37
4.2.4	Recommendation Process	39
4.3	Mood Information Phase	43
4.3.1	Mood Induction	43
4.3.2	Mood Detection	45

	Pilot Study	45
	Questionnaire	46
4.4	Mood based Recommendation Phase	47
4.4.1	Recommendation Algorithm	48
	K means Clustering	49
	Item Item Similarity	52
4.5	Conclusion	53
5	Evaluation 1 - To study the relationship between mood and latent musical features	55
5.1	Introduction	55
5.2	Design	55
5.3	Independent Variable	56
5.4	Dependent Variable	56
5.5	Hypothesis	56
5.6	Approach	57
5.7	Participants	58
5.8	Statistical Tests and Results	59
5.8.1	User Satisfaction for Happy Users	59
5.8.2	User Satisfaction for Sad Users	60
5.8.3	Unexpectedness 1: For Happy Users	62
5.8.4	Unexpectedness 2: For Sad Users	64
5.8.5	Additional Analysis	65
	Comparing Genre and Happy ratings	66
	Comparing Genre and Happy ratings	66
5.9	Conclusion and Discussion	68
6	Evaluation 2 - To study the relationship between user mood and valence	73
6.1	Introduction	73
6.2	Design	73
6.3	Independent Variable	73
6.4	Dependent Variable	74
6.5	Hypothesis	74
6.6	Approach	74
6.7	Participants	75
6.8	Statistical Tests and Results	76
6.8.1	User Satisfaction for Happy Users	76
6.8.2	User Satisfaction for Sad Users	77
6.9	Conclusion and Discussion	78
7	Discussion	81
7.1	Analyse the relationship between latent musical features and user satisfaction for a given emotion	82
7.2	Analyse the relationship between latent musical features and unexpectedness for a given emotion	83
7.3	Limitations	83

8 Conclusion	85
8.1 Future work	86
Bibliography	87
A System for Experiment 2	97

List of Figures

2.1	Valence Arousal Model [1]	13
3.1	Heatmap representing Correlation between different attributes in MSD subset	19
3.2	Heatmap representing Correlation between different attributes in Spotify Dataset	20
4.1	Research Goal 1	28
4.2	Research Goal 2	28
4.3	System Design of Moodify	29
4.4	Valence values information of DEAM dataset	32
4.5	Metafeatures Information of DEAM dataset	32
4.6	Acoustic features of DEAM extracted using Essentia	32
4.7	Final Merged DEAM Dataset	33
4.8	Example of a content based recommender system[2]	36
4.9	A genre-based music recommender system	38
4.10	Recommendation Process for Baseline Recommendations	39
4.11	Demographics collected from user	40
4.12	Users asked to rate songs by Moodify	41
4.13	Genre based recommendations provided by Moodify	42
4.14	Process flow in Mood Information Phase	43
4.15	Happy video	44
4.16	Responses for Happy Video via Affectiva	45
4.17	Responses for Happy Video via Questionnaire	45
4.18	Responses for Neutral Video via Affectiva	46
4.19	Responses for Neutral Video via Questionnaire	46
4.20	Responses for Sad Video via Affectiva	46
4.21	Responses for Sad Video via Questionnaire	46
4.22	Questionnaire after showing a Happy Video	46
4.23	Process flow from Mood Information to Recommendations	47
4.24	K Means Clustering Explanation[3]	49
4.25	Datframe without Labels	49
4.26	Dataframe after obtaining Valence Labels	50

4.27	Final Dataframe after obtaining Danceability Labels	50
4.28	Distribution of Songs in Valence Clusters	50
4.29	Distribution of Songs in Danceability Cluster	50
4.30	Subsets formed from the Dataset	51
4.31	Distribution of Valence in dataset	51
4.32	Distribution of Danceability in dataset	51
4.33	Recommendations generated for List 2	53
5.1	Participant Demographics by Nationality	58
5.2	Participant Demographics by Gender	58
5.3	Participant Demographics by Age	58
5.4	Participant Demographics by their consumption of Music Recommender Systems	58
5.5	Enjoyment Ratings of Users for List 2 recommendations	59
5.6	Enjoyment Ratings of Users for List 3 recommendations	59
5.7	Gaussian Plot for List 2 enjoyment ratings	59
5.8	Gaussian plot for List 3 enjoyment ratings	59
5.9	Q-Q Plot for List 2 enjoyment ratings	60
5.10	Q-Q plot for List 3 enjoyment ratings	60
5.11	Enjoyment Ratings of Users for List 4 recommendations	61
5.12	Enjoyment Ratings of Users for List 5 recommendations	61
5.13	Gaussian Plot for List 4 enjoyment ratings	61
5.14	Gaussian plot for List 5 enjoyment ratings	61
5.15	Q-Q Plot for List 4 enjoyment ratings	61
5.16	Q-Q plot for List 5 enjoyment ratings	61
5.17	Surprising Ratings of Users for List 2 recommendations	62
5.18	Surprising Ratings of Users for List 3 recommendations	62
5.19	Gaussian Plot for List 2 surprising ratings	63
5.20	Gaussian plot for List 3 surprising ratings	63
5.21	Q-Q Plot for List 2 enjoyment ratings	63
5.22	Q-Q plot for List 3 enjoyment ratings	63
5.23	Surprising Ratings of Users for List 4 recommendations	64
5.24	Surprising Ratings of Users for List 5 recommendations	64
5.25	Gaussian Plot for List 4 surprising ratings	64
5.26	Gaussian plot for List 5 surprising ratings	64
5.27	Q-Q Plot for List 4 enjoyment ratings	65
5.28	Q-Q plot for List 5 enjoyment ratings	65
5.29	Enjoyment ratings for genre based recommender system	66
5.30	Distribution plot for Genre ratings	66
5.31	Q-Q plot for genre ratings	66
6.1	Participant Demographics by Nationality	75
6.2	Participant Demographics by Gender	75
6.3	Participant Demographics by Age	75
6.4	Participant Demographics by number of hours they consumed a Music Recommender System in a week	75
6.5	Enjoyment Ratings of Users for List 2 recommendations	76

6.6	Enjoyment Ratings of Users for List 3 recommendations	76
6.7	Gaussian Plot for List 2 enjoyment ratings	76
6.8	Gaussian plot for List 3 enjoyment ratings	76
6.9	Enjoyment Ratings of Users for List 4 recommendations	77
6.10	Enjoyment Ratings of Users for List 5 recommendations	77
6.11	Gaussian Plot for List 4 enjoyment ratings	77
6.12	Gaussian plot for List 5 enjoyment ratings	77
A.1	Recommendations generated for List 2	97
A.2	Recommendations generated for List 3	98
A.3	Recommendations generated for List 4	99
A.4	Recommendations generated for List 5	100

List of Tables

2.1	Algorithmic approaches in Context-Aware Recommender System	7
2.2	Features in MER	9
3.1	Comparison of all datasets	18
3.2	Occurrences of Null Values in the Million Song Subset	19
3.3	Scores for MER task	24
4.1	Song Waterduct from our dataset	33
4.2	Initial Seed Items	35
4.3	Clips used for Emotion Induction	44
4.4	Characteristics of Recommendation Lists	48
4.5	Descriptive Statistics for Valence and Danceability	51
5.1	Demographic questions	57
5.2	Statistical Test for List 2 and List 3 enjoyment ratings	60
5.3	Wilcoxon’s signed rank for List 2 and List 3 ratings	60
5.4	Descriptive Statistics of recommendation List 2 and List 3	60
5.5	Statistical Test for List 4 and List 5 enjoyment ratings	62
5.6	Wilcoxon’s signed rank for List 4 and List 5 ratings	62
5.7	Descriptive Statistics of List 4 and List 5 enjoyment ratings	62
5.8	Statistical Test for List 2 and List 3 surprise ratings	63
5.9	Wilcoxon’s signed-rank for List 2 and List 3 surprising ratings	63
5.10	Descriptive Statistics of List 2 and List 3 surprising ratings	64
5.11	Statistical Test for List 4 and List 5 surprise ratings	65
5.12	Wilcoxon’s signed-rank test for List 4 and List5 surprise ratings	65
5.13	Descriptive statistics of List 4 and List 5 surprise ratings	65
5.14	Wilcoxon’s signed-rank test to compare genre ratings and ratings for list 2 and list 3(happy phase)	66
5.15	Descriptive Statistics of enjoyment ratings for list 2 and list 3(happy phase)	66
5.16	Wilcoxon’s signed-rank test to compare genre ratings and ratings for list 4 and list 5(sad phase)	67
5.17	Descriptive statistics to compare genre ratings and ratings for list 4 and list 5(sad phase)	67

5.18	Insights for Happy and Sad Phase recommendations	69
5.19	Insights for Happy and Sad Phase recommendations	69
6.1	Demographic questions	74
6.2	Characteristics of Recommendation Lists for Experiment 2	74
6.3	Wilcoxon’s signed-rank test for List 2 and List 3 enjoyment ratings	76
6.4	Descriptive statistics of List 2 and List 3 enjoyment ratings	77
6.5	Wilcoxon’s signed-rank test for List4 and List5 enjoyment ratings	77
6.6	Descriptive statistics of List 4 and List 5 enjoyment ratings	78
6.7	Relevance of Recommendation Lists to User Profile	79
6.8	Insights for Happy and Sad Phase recommendations	79

Chapter 1

Introduction

Recommendation systems are search and decision tools that filters and provides relevant information to the user. With plethora of information available on the web, it becomes important to have such filtering tools that could help users find interesting items saving both their time and energy.

Recommender systems are popular in the domain of music and help users find musical tracks that they would enjoy. Spotify, Pandora, Apple Music use recommender systems to provide users with music recommendations. These systems analyse user's listening habits and generate music recommendations that would suit user's taste.

In the recent years, research shows that contextual information like emotional state, activity, time of the day impact user's preferences [4] and thus should be considered while providing recommendations. Studies show that music listening is context dependent. Moreover, emotion highly influences user's music preferences [5]. Researchers have discussed that considering emotion in the recommender system positively influences user satisfaction [6] [7] making it an important feature to be considered while providing music recommendations. Additionally, researchers observed that current emotional state of the user directly influences their music taste. It is seen that when people are sad, they prefer listening to sad music slow music and when people are happy they prefer listening to more happy and upbeat music which is also known as *mood congruence* effect in the field of psychology. [8].

Our study focuses on the research directions discussed above. We further discuss our problem statement in the next section.

Note: We use Mood and Emotion Interchangeably in this study. We are talking about user's current emotion whenever we mention emotion/mood in this work.

1.1 Problem Statement

The boom of the internet has left us overwhelming with information. Movies, music, products, there is plenty of everything on the internet. With so many choices, the user often suffers while making a decision. This is where recommender systems come into play, helping users by providing them with relevant items. The focus of our study is music recommendation. Just like the Internet, the digital music industry has grown

exponentially in the last few decades resulting in enormous musical data and resulting in need of the recommender systems. These recommender systems use machine learning algorithms, behaviour analytics of the users for making music predictions. Some popular examples of music recommenders are Spotify, Amazon Music and Apple Music. These recommenders are known to use machine learning techniques for learning user's preferences, and provide them with enjoyable recommendations. Spotify's Discover Weekly is one such playlist known for providing a personalised playlist that seems like a curated mixedtape made just for you.

As we know these recommendations are curated to user's music taste, but often fail to consider their current emotion which is indeed an important factor to be taken into consideration while providing music recommendations. Thus, making it an important topic of study in the recommender systems domain. The existing research on emotion-based recommender system often uses self-reported emotions via a questionnaire or colour based strategy to detect user's emotional state. The emotion induction techniques from the field of psychology are still in the nascent stages in the field of recommender systems. Moreover, music is often associated with an emotion. The field of Music Emotion Recognition solely works to identify potential musical features that are responsible for representing an emotion in a song. The most popular features studied are tempo and mode. Modulating tempo can express emotions in a song, slow tempo often represents sad music and fast tempo often represents a happy tune [9] [10]. Major and Minor Modes are associated with emotions Happiness and Sadness [11] [10]. In addition to tempo and mode, a song has many other features which form a key part in expressing emotion in music, timbre, tonality, loudness, danceability, valence and arousal being few of them. For our study, we wanted to test the relationship between current user emotion and their preference of emotional music. Due to the scope of the thesis, we concentrate on understanding the relationship between happiness/sadness in a user with happy/sad music. This is done by looking at values of features danceability and valence in the musical piece. We decided to look at these features because of the reasons: a) ease in understanding the meaning of these features b) limited study on these features and their corresponding emotional value. To summarize the research in the recommender systems domain and music emotion recognition:

- Primarily, the recommender systems use a content based or collaborative filtering approach for recommendations. Recent research has shown emotion as an important contextual features especially in the domain of music recommendations
- Studies have seen the effect of modulating tempo and mode but limited research has been done to see the effect of modulating other features and corresponding emotion expressed doing the same.
- Emotion recommender systems often use self-reported emotions by the users. Limited research has been done to bridge the gap between emotion induction techniques and recommender systems domain.

These studies bring us to problems in the field:

- Limited research to understand emotion and preference of latent musical features
- Not enough studies use the emotion induction techniques from psychology in the recommender system domain.

1.2 Research Objectives

Based on the problem statement defined, we have one main research gap which forms the essence of this study.

RQ1: How does mood affect user's consumption of latent musical features Danceability and Valence?

Objective: To answer RQ1, which is the main research objective of our study, we aim to build an interactive emotion based recommendation system. We plan to generate recommendations based on user mood and analyse user satisfaction. These recommendations will be generated using similarity between musical features (danceability & valence) for songs and hence help us analyse the relationship between user mood and satisfaction when recommended items based on the musical features.

RQ2: Does mood impact how surprising user finds the items recommended by the system?

Objective: To answer RQ2, we use our interactive system and ask questions to check if the recommendation list was surprising to the user.

Note: These recommendations are the same as the recommendations which checks user satisfaction

1.3 Contributions

Our main contributions have been the following:

- We successfully built an interactive system which would induce emotions and then provide mood specific recommendations based on latent features.
- We analyzed the relationship between latent song features (Danceability and Valence) and user mood and were able to analyse the feature values they prefer in a particular emotional state.
- Additionally, our emotion based recommender system is novel in a way that it uses unconventional techniques of emotion induction borrowed from the field of psychology. The users are shown movie clips curated for the purpose of emotion induction(here sadness and happiness) that would put them in a certain emotional state.

1.4 Thesis Outline

We begin with a study of the literature in the field in Chapter 2. In Chapter 3, We explore different datasets and validate our decisions for the dataset chosen. Next, we describe our System in Chapter 4. Chapter 5 and 6 discuss our experimental setup and Hypothesis testing. We conclude with discussion and limitations in Chapter 7 followed by conclusions and future work in Chapter 8.

Chapter 2

Literature Review

To answer our research questions on Understanding the relationship between latent musical features and mood in the domain of recommender systems, we conducted an in-depth study in the emotion recommender systems domain. We aim to answer our research questions by building an interactive interface which would help the user go through a certain emotional state and later consume the recommendations provided by our system.

In this chapter, We start with some background work in the domain of context-aware recommender systems, followed by motivation to use emotion as a contextual feature, approaches and algorithms used in emotion-based recommender system research, techniques of music recommendations, evaluation techniques and conclude with emotion induction and detection techniques which we would need in our research.

2.1 Introduction to Recommender Systems

With the advent of the World Wide Web and the resulting data boom, services and tools which would filter data and find relevant information become of great value. Recommender Systems are such search and decision tools which help us find the relevant information. These systems help overcome information overload by providing users with information that is appealing to them[12]

Recommender systems are widely used by different services like Amazon[13], Netflix[14], Spotify[15] to provide users with relevant and interesting recommendations. These systems use different techniques for the recommendation. Traditional recommendation techniques can be broadly classified into Content-Based Recommender Systems, Collaborative Filtering based Recommender Systems, Demographic-based and Knowledge-Based Recommender Systems. These recommendation techniques are briefly described next.

Content-Based systems provide recommendations by finding similar items to the items that were previously consumed by the user. Collaborative Filtering method compares the user profiles and provides recommendations liked by a similar user[16]. Demographic-based recommender systems provide recommendations based on the demographics of the user[17]. Knowledge-Based recommender system recommends items based on domain knowledge. A similarity function estimates how much the user's needs match the recommendations solutions of the problem. The similarity

score can be considered as a utility of recommendation to the user.

These algorithmic approaches have majorly contributed to the recommender system community. However, these approaches fail to consider that users can have different preferences in different contexts. It has been validated in research that contextual information like mood, time of the day, activity, presence of people and location when taken into consideration by recommender systems lead to greater user satisfaction [4]. Hence, making context-aware recommendations important to study.

2.2 Existing Context-Aware Recommender Systems

As discussed in Section 2.1, context-aware recommender systems take into account the user's contextual information while recommending the items.

In this section, we discuss context-aware recommenders in the domain of music recommender systems. Research says that music listening is context-dependent as people might prefer listening to a different kind of music in different scenarios. Context has a strong impact on one's music preference and consumption [18][19]. Eg people prefer listening to songs on which they could dance to at a party, loud-high energy music while working out and soothing music on a romantic dinner.

Context-Aware music recommender systems could be broadly divided into two categories: 1. *User Centric context-based approach* and 2. *Environmental context-based approach*. These approaches have been discussed in detail.

2.2.1 User Centric context-based approach

User-centric approaches study the impact of user's physical and mental state and their mood on their preferences and consumption of recommended items. Research says that the mood of the user directly influences their music preferences [5]

Wang et al. [20] develop a music recommendation system that would provide music recommendations for activities: *running, working, sleeping, walking, shopping and studying*. Deng et al. [21] recommends music that would fit their current emotional state. This emotion is extracted from user's microblogs. Yoon et al. [22] developed a personalised music recommendation system based on low-level musical features extracted from TV music program's audience rating based on emotional feelings, user's listening rating and user's current mood. Han et al. [23] proposed an emotion state transition model for their context-aware recommender system. This model would help in modelling the user's emotions and their transitions by music. This model acts as a bridge between a user's emotional state and musical features.

2.2.2 Environment context-based approach

Environment based approaches are based on the fact that the environment of the user influences their music preferences [24]. It has been seen that music recommender systems that consider environmental contextual information perform better than traditional systems which do not consider any environmental contextual information. Time, weather, location are some of the environmental-related contexts in the recommender

system research. Park et al. [25] developed an environmental-based context-aware system that would recommend music based on weather, time, noise and light level. Chen et al. [26] developed VenueMusic which would recommend relevant songs based on popular venues in our daily lives. Schedl [4] propose a geospatial model that takes into account user's GPS coordinates along with a cultural model that accounts for continent, country and state of the user to provide music recommendations. Dias et al. [27] use temporal information in session-based collaborative filtering system to improve recommender system performance.

Table 2.1: Algorithmic approaches in Context-Aware Recommender System

Article	Algorithm Technique
Improving Music Recommendation in Session-Based Collaborative Filtering by using Temporal Context [27]	Collaborative Filtering
User Geospatial Context for Music Recommendation in Microblogs [4]	Hybrid Collaborative Filtering
On Effective Location-Aware Music Recommendation [26]	Content-Based approach
A Context-Aware Music Recommendation System Using Fuzzy Bayesian Networks with Utility Theory [25]	Fuzzy Bayesian approach
Music emotion classification and context-based music recommendation	Content-Based approach
Exploring user emotion in microblogs for music recommendation	Hybrid Collaborative Filtering
Music Recommendation System Using Emotion Triggering Low-level Features	Content bases

2.3 Emotion in Recommender Systems

In this section, we discuss the importance of emotion in decision making in humans which makes it an important contextual feature to be considered while providing recommendations which we have mentioned in Section 2.2. Later, we discuss some interactive emotion-based recommender systems.

2.3.1 Why Emotion?

Decision making depends on various factors. Studies show that emotions play an important role in the decision-making process [28]. Goleman mentions that emotional intelligence plays an important role in human decision-making process [29]. Joseph proves that emotions play an absolute role in cognitive processes at a neurological level [30]. They should be considered while providing user recommendations as these solutions might help in improving the positive emotional state of the user. Polignano [31] considers emotions and personality while providing recommendations. Their work looks into the role of emotions in each decision making task namely low risk, medium risk and high-risk tasks and acts as a framework for including emotions inside the recommender system. Tkalcic et al [32] provide a framework that describes how emotions can be used to improve the quality of recommender systems in three ways namely when emotions are induced in 1) Entry Stage 2) Consumption Stage and 3) Exit Stage, and how changing the process in each stage is an issue. Gonzalez et al. [33] show that adding an emotional factor in recommender systems both content-based and collaborative approaches improve the recommendations and user satisfaction in the restaurant domain. Research also shows that emotions act an important role in the context-aware recommendation by improving predictive performance [7]. It is also seen that emotions are critical in the decision making process and decisions are always transmitted with emotions by users [6].

Thus, it becomes important to consider emotion for the recommendation process.

2.3.2 Interactive Emotion-Based Recommender Systems

E-MRS is a movie based recommender system which recommends movies based on inferences about a user's emotion and preferences and opinions of similar users. It uses a colour based strategy to detect the emotion of a user. Eg colours like Yellow, Light Orange, Blue and Green denote Joy. It incorporates emotions to recommend movies and considers novelty as a factor while evaluating system[34]. Their system provides users with explanations for the provided recommendations. CoFeel uses emotion to enhance social interaction by providing emotional feedback and thus engaging users in group interactions[35]. The results of the experiment showed that emotion serves as an effective element to elicitate users' attitude and increased user engagement in the group. Arapakis et al use an emotion recognition system which analyses the current state of the user and provides feedback which is seen to improve the recommender system's performance[36]. Moodplay integrates content and mood-based algorithm in an interactive interface to provide music recommendations. The system supports control and explanation of affective data through an interactive interface. Results show that visualization and interaction in a latent space improved acceptance of items among users. Users liked exploring moods in interactive space[37].

2.4 Features for Music Recommendation

In this section, we discuss the features that are popular in the Music Recommender System domain. We further discuss features used for music emotion recognition process.

2.4.1 Features used in Existing Music Recommender Systems

In the field of Recommender Systems, features play a key role in providing recommendations. In literature, most of the Music Recommender Systems rely on musical features like genre, song popularity and artist name to provide the user with song recommendations. Auralist uses an Artist based LDA model for providing item-based recommendation [38]. LDA is traditionally a technique used for topic modelling. Auralist uses LDA to form user communities based on their artist preferences where the artist is considered as a document and users are considered as words. Thus, the system produces a similarity value for artist topic vectors which is later used to provide item-based recommendations by calculating. F Lu and N Tintarev [39] in their work 'A Diversity Adjusting Strategy with Personality for Music Recommendation' look at the features Release date of the track, Artists, Genres, Key and Tempo for their diversification algorithm. Y Jin et al [40] studied the effect of personal characteristics on music recommender systems where users could control artist, track and genre weight in the algorithm to get personalised recommendations. Ferwerda et al [41] explore music diversity needs of users across various countries by looking at the Echonest features hotness, familiarity, discovery along with artist and genre. Schedl and Hauger look into the genre and unique track count for estimating the diversity of a user [42]. Bogdanov [43] propose three content-based recommender systems which use various features timbral, temporal features, in which some features have been extracted using Essentia.

For our study, we wanted to use rich audio features for providing useful recommendations. This required investigating musical features which are highly correlated to emotion. Thus, we looked in the field of Music Emotion Recognition to find some potential features which are often used for emotion recognition in music. This has been discussed in the next section. These features are studied so that they could be used to build an emotion classifier at later stages of the study.

2.4.2 Features in the field of Music Emotion Recognition

This section discusses the literature in the field of Music Emotion Recognition(MER) and the features used in the articles for emotion recognition in music.

Studies often rely on the standard and melodic features for emotion analysis in music. Standard features are often described as features that aim to represent attributes of audio. Timbre, rhythm, tempo, pitch, harmony are some attributes that are often extracted for emotion recognition in music. Melodic audio features are another set of features that are popular for emotional analysis of music. These features are broadly divided into three categories: pitch and duration, vibrato, and contour topology. Additionally, lyrics of the song is another important feature for emotion analysis in music. We further studied a few papers to understand these features which are further discussed below.

Panda et al[44] proposed to combine standard and melodic features extracted from audio for music emotion recognition. Their study shows that melodic features achieve better performance than standard audio. Madsen et al[45] in their study test if using multiple temporal and non-temporal representations of different features helps in modelling music structure that would help in predicting emotion in music where temporal features represent features that are dependent on time and non-temporal features are frame-based vectors that are independent in time. Fukayama and Goto [46] use an adaptive aggregation for improving emotion recognition accuracy. Jamdar [47] use lyrical and audio features for emotion classification. ANEW and WordNet knowledge is incorporated for computing valence and arousal values from lyrics and audio features are supplemented with the lyrical features. Corona [48] use lyrics to classify emotions in Million Song Dataset achieving accuracy up to 70 % for classification of some moods but the results were not statistically significant. Trohdis et al [49] focus on multi-label classification where the predictive power of various audio features is evaluated using a multi-label feature selection method.

Table 2.2 lists the features used in the papers in detail.

Table 2.2: Features in MER

Article	Features
Music emotion recognition: A state of the art review [50]	Dynamics, Timbre, Harmony, Register, Rhythm and Articulation
Music Emotion Recognition with standard and melodic audio features [44]	Standard Audio Features, Melodic Audio Features, Low-Level Descriptors
Music emotion recognition with an adaptive aggregation of Gaussian process regressors [46]	Tempo, Pitch, Loudness and Timbre
Learning Combinations of Multiple feature representations for music emotion prediction [45]	Chrona features and Loudness
Prediction of multidimensional emotional ratings in music from audio using multivariate regression models [51]	Timbre, Harmony, Register, Rhythm, Articulation and Structure
An exploration of mood classification in MSD [48]	Lyrical features
Multi-Label Classification of music by emotion [49]	Rhythm and Timbre
Emotion Analysis of songs based on lyrical and audio features [47]	BPM, Danceability, Loudness, Energy and Mode
Exploiting Genre For Music Emotion Classification [52]	Genre

2.4.3 Emotion and music preference

We looked into the literature to find out music preferred by people in their current mood. Studies were seen to be limited to two emotions happiness and sadness as these emotions are easily recognisable. Tempo and Mode were two popular features used to express emotion in a song. In literature, the fast tempo is associated with happy or positive music and the slow tempo is recognised as a sad musical piece. Additionally, major and minor modes are associated with happy(positive) and sad(negative) music respectively [53]. Studies show that people show an increased liking towards sad music when they are sad [54]. It is also seen that when people are sad, they are less likely to listen to a happy-sounding music than people who are in a happy or neutral state [55] [56]. There were studies which reported a contrast effect, where people preferred listening to sad music after they were made to hear happy tunes consecutively [57].

2.5 Evaluation

Recommender systems are an integral part of the digital era and help us providing with relevant information. The recommender system research community has been constantly working towards developing new algorithmic approaches which would help generate relevant and enjoyable recommendations. These systems are crucial and hence their evaluation extremely important. Evaluation techniques are needed to compare different algorithms, determine the performance of a recommender system, and in understanding the best approaches for a dataset.

In the traditional recommender system research, accuracy was the most popular metric for evaluating the performance of a recommender system. Thus, the most common metrics used to evaluate the system were accuracy oriented - *precision*, *recall*, *F1 score*. These metrics are easier to measure as they are mostly evaluated using a *offline experiment* by performing experiments on the existing datasets [58]. These metrics though easy to deploy in a system often do not satisfy users in real-time.

2.5.1 Discovery oriented Metrics

As discussed, accuracy oriented metrics determine if the predictions made by the recommender system are close to actual user choices and thus judge the efficiency of a recommender system. Herlocker et al. [59] in their research discuss how user satisfaction is not only dependent on the accuracy of the recommender system. There could be other factors influencing user satisfaction such as new recommendations or items that have not yet been experienced by the user [60]. This led to the development of discovery-oriented metrics. In this section, we discuss the discovery-oriented metrics: *diversity*, *novelty*, *serendipity* and *coverage*

Diversity

Diversity is one of the most popular and well-known metrics, which has been taken from the information retrieval community in recommender systems research. It is measured as a dissimilarity between recommended items of a list. Introducing diversity in

recommender system processes will provide users with interesting recommendations. Diversity was first used in recommender systems by Bradley and Smyth [61]. Intra list similarity is defined as the aggregate pairwise similarity of the items in the list and is often used in recommender systems literature as diversity metric [62]. Diversity is often measured by considering average or aggregate dissimilarity of items with different functions to calculate the item distance. When items are represented by content descriptors, the distance between items is measured through the complement of Jaccard similarity [63], the complement of cosine similarity [64] and taxonomy-based metric [62] and items are represented by rating vectors the item distance is measured by the complement of cosine similarity [65], complement of Pearson correlation [63] or Hamming Distance [66].

Novelty

Novelty often refers to how different information is from what has been previously seen by the user. The concept of novelty mainly focuses on two aspects: an item is unknown to the user and it being different from a user's profile. Literature has different variants of Novelty metric, Castells [63] use both different and unknown characteristic into their novelty metric, Yang et al [67] and Nakatusji et al [68] consider item's distance from user's profile, Zhang [69] considers an item to be novel if it fulfils three qualities: is unknown to the user, relevant to the user and is different from the items in the user profile.

Serendipity

The term serendipity has no formal definition in terms of recommender systems [70] [71]. Few pieces of research define an item to be serendipitous if it is relevant, novel and unexpected to the user [70]. Serendipitous items are unexpected by the user and yet leave them in a positive emotional state.

Zhang et al. [38] define serendipity as something unusual and surprising. Maksai et al. [72] describe serendipity as the quality of being both unexpected and useful. McNee [73] defines serendipity in a recommender system as the experience of receiving an unexpected and fortuitous item. Sridharan [74] defines Serendipity as the accident of finding something good or useful while not especially searching for it

Coverage

Coverage refers to the degree to which a recommender system can cover the set of available items and generates recommendations for all potential users. This metric is defined at a system level instead of just item or user-level [75]. It is also defined as the percentage of the dataset for which the recommender systems can predict [76] [77]. Recommender systems with Higher coverage are preferred because of their higher predictability of items in the dataset.

Others

Apart from the above-mentioned metrics, few works also propose some other metrics for recommender system evaluation. Hijikata [78] discusses discovery ratio which is a

measurement of the number of unknown items in the recommendation list. This metric is independent of a user's preference and hence is different from novelty.

Avazpour et al [79] discuss other metrics, correctness, trustworthiness, recommender confidence, being few of them.

2.5.2 Evaluation of Emotion-based Recommender Systems

Emotion-based recommender systems often use accuracy oriented metrics for evaluation purposes. Prediction accuracy metrics like precision, recall, F1 score are popular in the domain of emotion-based recommender systems [21] [23]. Mean absolute error is another popular metric in the field [34].

These metrics evaluate a recommender system but fail to see the effect of emotion successfully. Eg Foster et al. [80] discusses serendipity as a difficult concept to measure as it includes some emotional dimension. It would be interesting to see how mood impacts this discovery-oriented metric.

2.6 Emotion- Definition, Models, Induction and Detection techniques

This section discusses literature on emotion and emotion-based models (Section 2.6.1), followed by a study on existing emotion induction techniques in the field of psychology (Section 2.6.2) and concludes with emotion detection techniques in the domain of recommender systems (Section 2.6.3)

For our study, we found it important to study the existing emotion induction approaches as it was important for our experiment which would involve putting the participants in a certain emotion in the simulated environment. We decided to go with this approach to avoid the small sample size obtained from participants for emotion happy and sad when they would self report it.

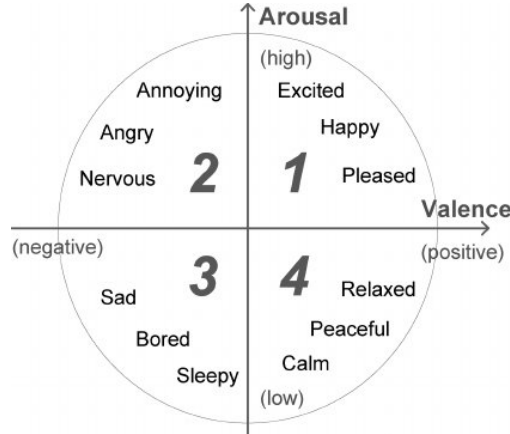
2.6.1 Emotion-Definition and Models

Oatley et al in their book *Understanding Emotions* term emotion as a state which is caused by an important event to the person. It includes a) conscious mental state with an identifiable feeling directed towards some object b) some kind of bodily perturbation c) noticeable expressions on the face, the tone of voice and gesture) readiness to involve in certain types of actions [81]. Picard et al [82] describes the emotion as a sequence of changes of state such that these states are inter-dependent and synchronized in a way in response to evaluations of the relevance of an external or internal stimulus. In our study, we induce an emotional state in the participant thus following Picard's definition of emotion [82]. People experience music in their everyday life and react to the kind of music they are listening to. They become happy after listening to concert music and dance numbers and sad after listening to sad songs [83]. Representing emotion is often difficult as induced emotion from music is often different from perceived emotion [84].

Different models have been studied on perceived emotion [85]. Some popular models are- categorical and dimensional models of emotion. The categorical approach

describes the emotion with limited categories. It proposes the existence of six distinct emotions- happiness, sadness, anger, disgust, surprise and fear [86]. Dimensional model, on the other hand, considers various affective terms which arise from independent neurophysiological systems: valence and arousal. Valence ranges from negative to positive and Arousal from calm to exciting [87].

Figure 2.1: Valence Arousal Model [1]



2.6.2 Emotion Induction Techniques

As previously discussed, Emotion plays an important role in human decision-making processes. Emotional states also affect one's music preferences. Our work looks into the impact of an emotional state on consumption of mood specific recommendations. Thus, In this section, we look at various Emotion Induction strategies discussed in the literature.

Emotional elicitation is often required in psychological research to be created in laboratories for scientific purposes. Various methods have been discussed in the literature for emotional elicitation that includes music [88], imagery [89], hypnosis [90], drugs and sleep deprivation [91] [92]. Few methods could have ethical issues and some hence use of pictures and movies is an acceptable method for emotional induction purposes. These methods involve deception and might not be standardised. Films, on the other hand, are standardised, dynamic and have a higher ecological validity which is desirable for emotional elicitation.

We further discuss approaches of emotional elicitation using films and video clips.

Hewig [93] discusses the capacity of emotional film clips for emotional elicitation. He introduces a set of films from commercially available film clips where participants are asked how they felt while watching the film on a scale from 0-9. Subjects are provided with enough time to recover from the induced emotional state before showing them a new film clip.

Gross et al [94] create a set of films that elicit eight emotional states-amusement, anger, contentment, disgust, fear, neutral, sadness and surprise. A number of 494 ethnically diverse, English speaking participants were employed for this task. Subjects were shown movies in groups ranging from size 3-30. Before each movie clip, a blank screen was shown to bring back the subjects to a normal state.

Their work was compared against Philippot's work on *Inducing and assessing differentiated emotion-feeling states in the laboratory* [95] and it was found that Philippot

looked for six target emotional states and the movies elicited less discrete emotions than Gross[94]. Fernandez et al[96] introduce a software program to recognise discrete emotions through psychological and physiological responses. The participants are shown movie clips and their facial expressions and physiological responses are evaluated at the same time. In addition to this, they are asked to fill in the questionnaires after each clip. A distracting task of showing different figures for a minute is performed by the participants to prevent the accumulation of emotions. After each session, a neutral clip is presented for recovery of the subject.

As part of our work, we follow Hewig's emotional film clips for emotional elicitation as they are commercially available films and easy to retrieve.

2.6.3 Emotion Detection in the domain of Recommender System

Emotion plays a vital role in the decision-making process inspiring Emotion-based Recommender Systems. These systems consider emotions and provide recommendations based on one's emotional state. Here, emotion detection becomes important. Various techniques to detect emotion exist. For this work, we chose to study Facial recognition and Skin Sensors for emotion detection.

Facial recognition

Facial recognition is widely used in Recommender Systems for emotion detection. Pauly and Sankar[97] use a facial recognition system as emotional feedback for their online product recommendation system. The system detects the emotional state of the user after each recommendation. Pessemier[98] propose facial detection techniques to know the demographics of the user which are used to solve the cold start problem in TV applications. Emotions depicted by the user during item consumption is taken as feedback.

Skin sensors

Guo[99] discusses a novel method of e-commerce recommender systems in virtual reality environments. Prepurchase ratings of users are used to provide recommendations and emotions are captured through EEG signals while users interact with these virtual recommendations.

Ayata et al[100] uses wearable physiological sensors for recommending music. Galvanic Skin Response(GSR) and Photo Plethysmography(PPG) are used for this purpose. The proposed method on real data provides better accuracy for emotion classification and thus can be integrated into recommender systems.

2.7 Conclusions

In this chapter, we discussed context-aware and emotion-based recommender systems. Additionally, we discussed discovery-oriented metrics and different techniques used for emotion induction and detection. We also discussed features used in existing recommender systems to train the model and briefly described the musical features in the

domain of music emotion recognition.

Emotion is an important contextual feature and it has been seen in research that people have different musical preferences for a certain mood. Most of the emotion-based recommender systems are seen to be still using accuracy oriented metrics for evaluation. Additionally, the existing music recommender systems look into features such as genre, artist, song popularity or low-level features for providing recommendations. After reading the literature we found some gaps which have been discussed in the next section.

2.8 Research Gaps and Motivation

This section discusses the research gaps identified in the literature and motivation behind the research questions and goals of this research.

2.8.1 Research Gaps

The following gaps have been identified in the literature from our survey:

- Limited work which uses high-level features for music recommendations
- The study of how the mood has an impact on one's need for high-level musical features in the recommender system domain hasn't been explored properly
- A gap between the field of Music Emotion Recognition and the field of Music Recommender Systems persists.

2.8.2 Motivation

In this research work, our main motivation was to understand the relationship between latent song features and user's preference for them in their current emotion. This would help us in understanding the songs they would enjoy the most in a given emotional state. Additionally, we also wanted to check if the recommendations based on these latent features are rather surprising to the user.

Chapter 3

Data Exploration and Selection

3.1 Introduction

In this chapter, we discuss data exploration and data enrichment which was done to achieve the final dataset used in our research.

We start with describing prospective datasets for our study (Section 3.2) and their analysis, followed by our motivation to decide a dataset. We further discuss enriching the dataset with musical features with the help of audio analysis tools (Section 3.3.2) and conclude with classification approaches (Section 3.3.3) that were carried on to increase the training dataset.

3.2 Dataset Selection

3.2.1 Selection of Features

We have discussed features used in the music recommender system domain in Section 2.4.1. Features such as genre, artist name, the popularity of the song tracks are the most popular features considered when generating musical recommendations.

For our study, we wanted to use some different features and hence we decided to look in the domain of Music Emotion Recognition (Section 2.4.2) for studying features that are correlated with mood.

After reading the literature, some features were selected that would be used by our recommender system to make music recommendations. The prospective features selected were namely: tempo (beats per minute), loudness, danceability and valence. These features were selected because a) they were easy to understand for someone without any background in music b) were known to be available in music recommender datasets and c) time constraints to conduct this study. We look for these features in the publicly available datasets so we could further exploit them to build a music recommender system.

3.2.2 Requirements from Dataset

In the previous section, we discussed prospective features that could be used to train our recommender system model. Next, we discuss the requirements this places on the selected dataset. Further, we discuss the properties of three candidate datasets. We conclude with motivation for our final choice of dataset.

While looking for a dataset, we concentrate if the dataset provides us with rich audio features discussed in the previous section, meta-features like track title, artist name and genre, audio files of the songs and emotion of the musical piece. We look for emotional tags happy, sad or valence arousal values for a given track. We look into different major datasets and compare them for features. Finally, we focus on three major datasets Spotify Dataset, Million Song Dataset [101] and Dataset for Emotional Analysis of Music(DEAM) dataset [102]. In addition to these datasets, we also looked at some other major datasets used in the field of Music Recommender Systems, like last.fm. We explored the dataset last.fm and found it unsuitable even for our analysis as it provides with no meta-features which is important for our research. Spotify dataset has been scraped from the web for the data available until November 2018. These datasets have been compared below:

Table 3.1: Comparison of all datasets

Dataset	MSD	Spotify	DEAM
Number of Songs	10,000	116,373	1802
Audio Features	Danceability,Tempo	Danceability,Tempo, Loudness,Valence..	Valence
Meta Features	Song & Artist Name, Song ID..	Song & Artist Name, Song ID, Genre	Song & Artist Name, Song ID, Genre
Audio Clips	No	No	Yes

3.2.3 Dataset Analysis

This section discusses the analysis that was done on the three datasets MSD, Spotify and DEAM. The purpose of the analysis was to identify missing and duplicate values in the dataset. Additionally, we wanted to find features with high correlations values which would also be used for the recommender system model.

Million Song Dataset

The Million Song Dataset [101] is a freely available dataset which is a collection of audio and meta-features for a million popular songs.

The dataset consists of a million songs and is 300 GB in size. Due to logistic constraints, we decided to first test a subset of the Million Song Dataset. The subset consists of 10,000 songs selected at random and is provided by Echonest. The dataset is in the form of HDF5 files and needs to be extracted to be used in a CSV/text form. With the help of the code provided by echonest, these files are extracted in the form a CSV file. The final result is a CSV file with 10,000 songs and all the meta-features and audio features provided by the echonest. This CSV file is further used for some Exploratory data analysis. The dataset is searched for duplicate values of song ids provided by echonest, null values of attributes, and correlation between different attributes. The dataset was found to have no duplicate values.. Table 3.2 lists some important attributes and the number of null values present for that attribute.

Attribute	Number of Null Values
Danceability	10,000
Energy	10,000
Song Hotness	4352

Table 3.2: Occurrences of Null Values in the Million Song Subset

Correlations between different attributes were found in the dataset. Figure represents correlation between different attributes present in the dataset excluding danceability and energy as they were all null values. From figure 3.1, one can see that no attributes have a high correlation value which would make merging the attributes not plausible in case we want to enrich our dataset. Another limitation of the dataset is that it doesn't come with genre tags for the tracks provided.

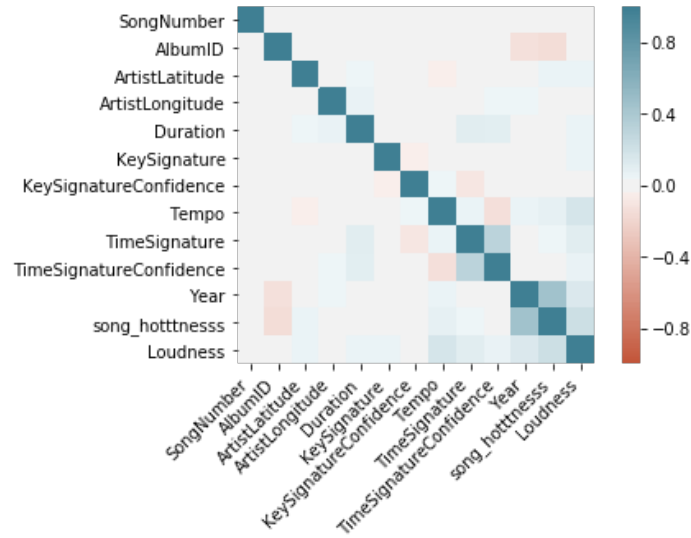


Figure 3.1: Heatmap representing Correlation between different attributes in MSD subset

Spotify

Spotify is a popular audio streaming platform that provides access to millions of songs. Spotify provides us with a Web API that allows us to extract the songs along with its audio features and meta features in a form of JSON or CSV file. A number of 116,373 songs were extracted from Spotify API till November 2018 and were further analysed. The dataset consists of songs with each row having values for artist name, track name, track id and the audio features.

This dataset was then searched for duplicate song ids, null values of any attributes and correlation between the attributes provided by the dataset.

The dataset was found to have no duplicate and null values for any of the features. A heatmap was plotted that describes the correlation between different features in the dataset.

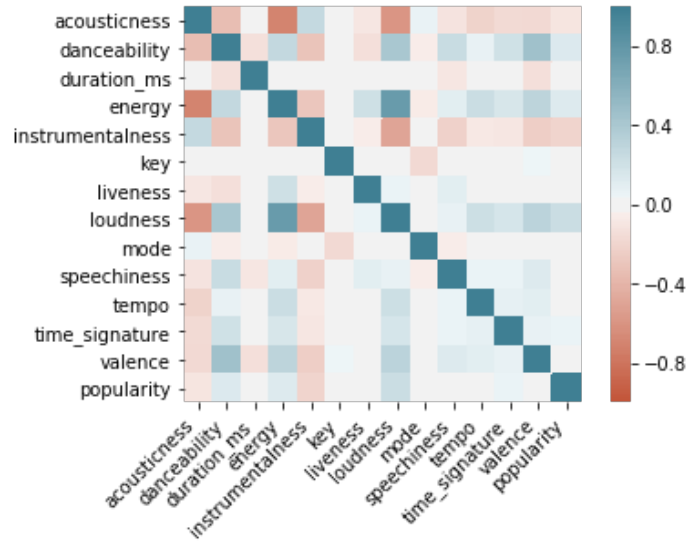


Figure 3.2: Heatmap representing Correlation between different attributes in Spotify Dataset

From Figure 3.2, one can see that few attributes have a high correlation between them. The attribute *acousticness* is seen to have a high negative correlation with *energy* and *loudness*. *Energy* has a high positive correlation with *loudness*. These correlations indicate that the feature *energy*, *loudness* and *acousticness* are correlated. This aligns with our study of musical features used in Music Emotion Recognition field (Section 2.4.2).

DEAM

DEAM dataset [102] is a MediaEval database for emotional music analysis of music and is a benchmark dataset for the task of Music Emotion Recognition task. The DEAM dataset is the largest available dataset with dynamic annotations for valence and arousal attributes.

The DEAM dataset consists of royalty-free music from various sources like jamendo.com, mendleyDB dataset and FMA. The dataset consists of 1,744 audio clips of 45 seconds retrieved from FMA and 58 full-length audio songs from both Jamendo and MedleyDB.

The audio files from FMA were from various genres like- pop, rock, blues, soul, electronic, classical, hip-hop, international, experimental, jazz, folk, country and other genres. The audio clips from MedleyDB had songs from other genres like rap and Jamendo also had songs that were from reggae music in addition to the genres from FMA. The dataset was created in such a way that no more than 5 songs from a single artist were included in the dataset. The full-length songs from MedleyDB and Jamendo were selected in such a way that they had emotional variation in them. This was done by using a dynamic Music Emotion Recognition algorithm for filtering and finally selecting the songs manually.

The annotations for these audio clips were collected through crowdsourcing using Mturk. The researchers took various steps to obtain high-quality annotations by designing tasks to filter out poor quality workers. For the years 2013 and 2014, each audio file was annotated by a minimum of 10 crowd workers. For the year 2015, each

audio file was annotated by 5 crowd workers. The dynamic annotations were collected through a web interface on a scale of -10 to 10 where the crowd workers could annotate the song for valence and arousal while the song was played. The static annotations were derived from the dynamic annotations on a 9 point scale for both valence and arousal. For our task, we consider the static annotations which are the average of the dynamic annotations.

DEAM was checked for duplicate values, null values. It was found to have no duplicate and null values. DEAM dataset had limited features *song title*, *artist name*, *genre*, *valence* and hence there was no need to perform a correlation analysis. Additionally, the dataset had audio clips.

3.2.4 Final Decisions about Dataset

After comparing the datasets on various metrics, one could see that MSD and Spotify provide a lot of songs but DEAM dataset is limited to only 1802 audio excerpts. Though the dataset provides us with a lot of songs, it doesn't provide us with the audio clips which are required to build our content-based recommender system. The Spotify and MSD dataset provides us with a lot of rich audio features making them potentially good datasets that would serve our purpose. MSD misses the values for our relevant features and hence was discarded in the initial stage of data selection. We further looked to know how these features were obtained by Spotify and unfortunately could not find any documentation on techniques used to extract these features. This made Spotify an unreliable source for our research. On the other hand, DEAM provides with the freedom to use the audio clips for audio analysis and other methods by providing audio clips in the dataset as well as explicit mood information in the form of Valence Arousal values annotated by humans. With the use of proper tools, Machine Learning techniques and/or audio analysis one could extract the required features and enrich the dataset. These features act as mood specific features and are later used for mood classification and feature vectors to train our recommender system. The Valence Arousal values for each song have been annotated by more than 10 people on average. The annotations are on a scale of 1 to 9 with 1 being the lowest and 9 being the highest. Thus, we decided to use the DEAM dataset as it gives freedom to use the audio clips for feature extraction ourselves which could be validated along with valence values that have been annotated by humans.

To extract the required features, we explored Librosa and Essentia. These are audio processing tools which support Python and are extensively used in the research domain for audio analysis. These are further discussed in the next section.

3.3 Extraction of Audio features

This section discusses audio processing tools for extracting musical features from the audio clips. It later discusses the algorithms that were used to retrieve these features.

3.3.1 Audio Analysis

Audio analysis is used to extract patterns and meaning from audio signals. It helps to extract the audio features from an audio signal. Audio analysis is used for classification, analysis, synthesis, storage and retrieval of audio files. Audio processing is easily supported in Python. It provides various libraries for audio processing like PyAudio and Librosa. For our study, we explored Librosa and another open-source library used in the research domain called Essentia. The libraries are discussed in detail in the sections below.

Librosa

Librosa [103] is a python module designed especially for audio and signal processing for music. The goal to build Librosa was to develop a stable package in python for Music Information Retrieval Applications. The Librosa package provides a basic tutorial and documentation of various functions that could be used with the help of the package. It provides with options to extract features like **tempo, beats, onset, mfccs, chromagram**.

Essentia

Essentia [104] is an open-source C++ library used for audio analysis and audio-based music information retrieval. It provides python bindings and thus could be used extensively for extracting audio features from an audio file. The library contains a collection of reusable algorithms that have implemented like standard digital processing blocks, statistical characterization of data, and music descriptors for spectral, tonal and high-level extraction of features.

Essentia was designed to focus on robustness, optimality and performance of the algorithms included in the library and has been made user-friendly. It provides with algorithms for **reading/writing audio files, signal processing tasks, filters, statistical descriptors, time-domain descriptors, rhythm descriptors, spectral descriptors and other high-level descriptors**. These algorithms help to retrieve both low level and high-level musical features. With the help of Essentia, we could retrieve features such as **danceability, loudness and tempo**

After exploring the two libraries, we decided to use Essentia for our study. This was done due to the following reasons:

- It allows flexibility to change the algorithms according to our needs
- It allows extraction of some high-level features of our interest, namely danceability, loudness which Librosa failed to provide.

3.3.2 Features extracted from Essentia

In this section, we describe the features we extract to represent mood. Additionally, we briefly discuss the algorithm used by Essentia to derive the features.

Loudness

Description

Loudness is described as a quality of a song/audio file that which is physical resonance to sound pressure and intensity. It is calculated in decibels(dB) and is averaged over the entire song to retrieve the loudness of the song.

Algorithm

Essentia comes with inbuilt algorithms to extract loudness from a song. The algorithm calculates the loudness of an audio signal that is defined by Steven's power law. The law states that loudness of an audio signal is equivalent to the energy of the audio signal raised to the power of **0.67**. In the domain of signal processing, Energy is defined mathematically as in equation 3.1 for a continuous-time signal

$$E_s = \langle x(t), x(t) \rangle = \int_{-\infty}^{\infty} |x(t)|^2 dt \quad (3.1)$$

Danceability

Description

Danceability is defined as an element which tells how suitable a song/audio file is to dance based on various audio features tempo, rhythm stability, beat strength and overall regularity.

Algorithm

The package has an inbuilt algorithm to compute danceability of an audio signal. The algorithm is derived from the method DFA described by Streich et al[105] in the paper Detrended Fluctuation Analysis of Music Signals: Danceability Estimation and further Semantic Characterization. To perform DFA, one needs to define the minTau and maxTau over which the DFA is performed. The algorithm outputs danceability of the input audio signal which ranges from 0 to 3. The higher value of danceability, more danceable is the song.

Beats Per Minute

Description

Tempo or beats per minute is the speed of a given song/audio file and thus gives some insights to the mood of the song. Eg Songs with higher BPM would be more energetic or happy than the songs with lower BPM values [9].

Algorithm

The Beats Per Minute(BPM) can be obtained using Rhythm Extractor module provided by Essentia. The algorithm in Rhythm Extractor extracts beat positions and estimates bpm and their confidence for a given audio signal. It requires the sample rate of the input signal to be 44100 Hz to be run correctly.

Valence values

Unfortunately, Essentia doesn't provide algorithms to extract valence from a given audio file. The valence values annotated by crowd workers were therefore considered for our dataset. The dataset provides annotations both per song and per user. As we needed valence values for an audio file, we considered the song level annotations for our study. Within that, we look for static annotations rather than dynamic annotations. The static annotations have feature values valence mean and valence standard deviation. Valence means is selected among the two as the song valence and is further used for building our recommender system. This was done because we wanted to focus on the mean valence of the entire audio clip and dispersion of the values from mean was not the focus of our study.

3.3.3 Emotion Classification of DEAM dataset

DEAM dataset consists of only 1802 audio files which are small to build a good music recommender system. To enrich the dataset, we wanted to do some emotion classification. This way we could get valence values for some additional audio files and create a larger dataset. To perform emotion classification, we looked at various state of the art methods for music emotion classification. Popular methods in the field of emotion classification are Linear Regression, Random Forests, SVM, Boosting methods[106][107][108]. These algorithms were used in our dataset for emotion classification. Algorithms like linear regression, XGB boost and SVM were tried for the DEAM dataset. The algorithm was trained on the features tempo, loudness and danceability and target feature was valence mean for the DEAM dataset. We used a 10 Fold cross-validation method and the reported scores are the mean of all the scores obtained from fitting the model.

Algorithm	MSE	MAE
XGB Boost	0.02448	0.126
SVM	0.0238	0.124
Linear Regression	0.024	0.125

Table 3.3: Scores for MER task

The scores obtained from the regression task were very poor. Hence, emotion classification task for the DEAM dataset was not feasible. We decided to use the DEAM dataset with 1802 music files for our study considering its limitations and the affect of the small dataset on our study.

3.4 Conclusions

In this chapter, We describe the flow and procedure that was carried out to obtain the enriched DEAM dataset used in our study. The conclusion of this chapter is briefly discussed below:

- We successfully studied the mood-based features which could help in recommending songs specific to a mood.
- We studied and compared different datasets for these features and decided to use the DEAM dataset for our study.

- We further obtained values for features danceability, loudness and tempo using *essentia* for the DEAM dataset. This is further used to enrich the dataset which has been discussed in the next chapter.
- Now, after deriving the necessary features we build our recommender system which has been discussed in the next chapter.

Chapter 4

System

In this chapter, we look into the methodology and our system **MooDify** that was developed to answer our research questions discussed in Chapter 1 in detail. Additionally, we discuss the design choices made for MooDify.

We begin with the chapter with providing an overview of the system architecture (Section 4.1) which is followed by a detailed explanation of different components of our recommender system (Section 4.1.1). Further, we discuss the three recommendation phases in our experiment (Section 4.2, 4.3 and Section 4.4).

4.1 System Overview

In our background work, we have discussed both emotional-based recommendation systems and context-aware based recommender systems. From our previous discussions, we understand how emotion-based recommender systems differ from traditional recommender systems. Emotions play a key role in decision making and thus has a great impact on user satisfaction. Traditionally, recommender systems do not consider irrational elements for computing recommendations but recent studies have shown that incorporating emotion can improve both quality and accuracy recommendations thus increasing system satisfaction. Taking these attributes into account, our study provides recommendations for emotions happy and sad. Another critical aspect of our research is measuring the unexpectedness of the recommendations for users. This is novel in the aspect that most of the work look into offline metrics precision, recall and accuracy. In order to answer our research questions defined in Chapter 1, we focus on the emotion of the user and metrics user satisfaction and unexpectedness (surprising).

We define the following research goals for our system:

- RG1: Our goal is to design a system which can help us study the relationship between latent musical features(danceability & valence) and user satisfaction for a given emotion
- RG2: Our goal is to design a system that would study the relationship between latent musical features(danceability & valence) and unexpectedness for a given emotion

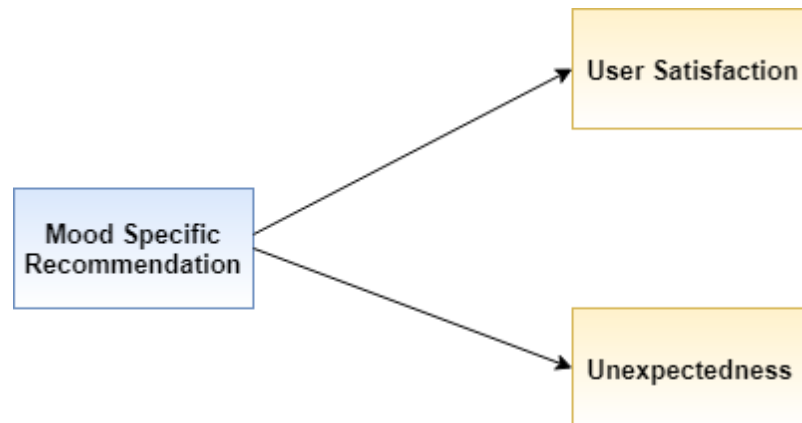


Figure 4.1: Research Goal 1

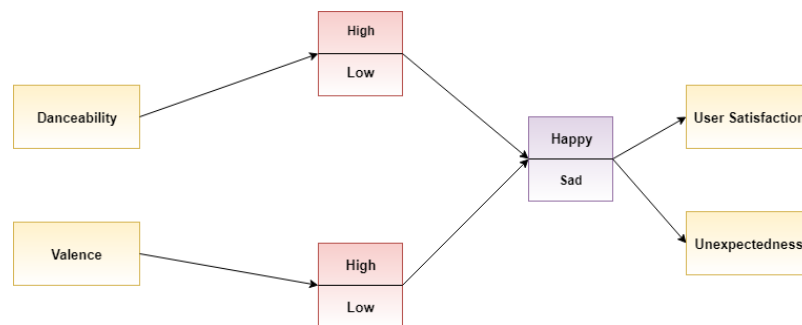


Figure 4.2: Research Goal 2

Figure 4.1 and Figure 4.2 provides with an overview of our research goals. We desire to measure the following:

- Analyse the relationship between latent musical features and user satisfaction for a given emotion
- Analyse the relationship between latent musical features and unexpectedness for a given emotion

4.1.1 System Design

In this section, we discuss the design of our recommendation system - MoodDify, a music recommendation system that provides mood specific recommendations.

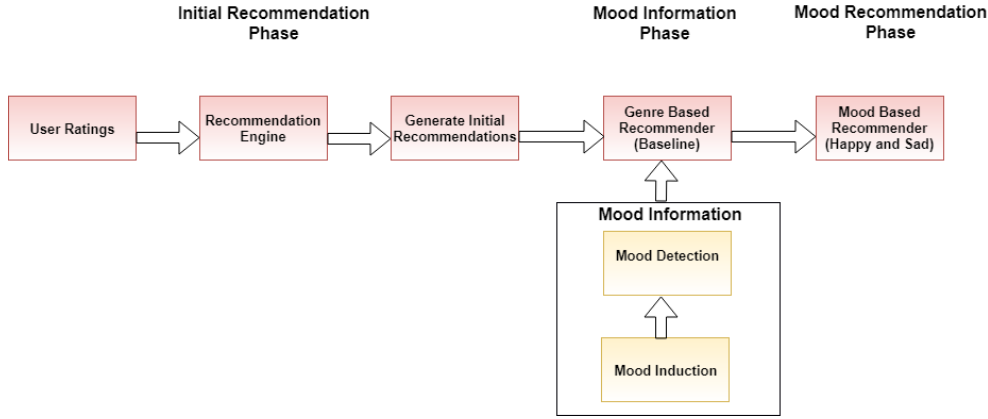


Figure 4.3: System Design of MoodDify

Figure 4.3 provides a basic system design of MoodDify, our mood based music recommender system. We can roughly describe the processes taking place in our system in three steps namely:

- **Initial Recommendation Phase** - The initial phase consisted of three basic steps. First, users were provided with a set of songs from different genres from our dataset and were asked to rate them on a scale of 1 to 5, 1 being least enjoyable and 5 being the most enjoyable. After getting the ratings from users, a user profile is created and genre-based recommendations that are most relevant to a user's musical taste are generated. This has been discussed in detail in section 4.2
- **Mood Information Phase** - The mood information phase induces and collects the mood of the user. After providing users with our first set of recommendations-genre based recommendations, users are shown a set of videos for inducing happy/sad mood in them. After this step, they are asked about their current mood and this information is given as an input for our next phase. This section has been discussed in detail in section 4.3
- **Mood Recommendation Phase** - This is the final set of recommendations provided to the user based on their mood information provided from the previous phase. It involves providing users with mood specific recommendations to test our hypothesis. These recommendations although being mood specific, are still consistent with a user profile for a few musical features. This has been discussed in detail in section 4.4.

The entire process can be summarized as below:

1. First, a user(participant) is provided with a consent form where the experiment is explained. If the participant agrees to it, he/she is asked to proceed with the demographics form where they are asked basic demographic information.

2. After collecting basic demographic data, the user is provided with an initial set of songs and is asked to rate them on a scale of 1(least enjoyable) to 5(most enjoyable). These ratings are used to build a user profile and learn user preferences.
3. After this, the user is provided with genre-based recommendations based on genres most enjoyable by them. For this, we set a threshold of 3 and the ratings are given by the user above it were considered enjoyable for the user. The user is also asked a set of questions based on recommended songs.
4. Next, the user enters a happy emotion phase where he/she is presented a random happy clip from our pool of happy clips for emotion induction and later is asked to self-report his/her emotional state.
5. After this, the user is presented recommended songs specific for a happy user. This is done by providing songs from the pool of songs with features mentioned in Table 4.4 for a happy user, by calculating the similarity between user profile and songs in these subsets. The user is also asked to answer a few questions to measure his satisfaction with the recommendation list.
6. Next, the user is shown a neutral video clip. In our case, this is the clip from *Hannah and her sisters*
7. For the next step, the user enters a sad phase, where he/she is presented with a random sad clip from our pool of sad clips and later is asked to report his current emotional state.
8. After this step, the user is provided with songs curated for a sad mood. This is done by selecting the most similar songs to the user profile in our subset data of sad songs. These subsets have features as shown in Table 4.4 for a sad user. The user is again asked a few questions based on the recommendation list provided to him.

To control the quality of ratings, we measured the time taken by the participants to do our study. We estimated a time of 30-45 mins to do the experiment. If the participants took less time than the estimated time, they were ignored for our analysis. We considered participants who took more than 45 mins to complete the experiment.

4.2 Initial Recommendation Phase

This section discusses in detail the initial phase of our system MooDify. Initial Recommendation phase was responsible for performing the following tasks:

- Collecting user demographics via a questionnaire
- Building a user profile by asking participants to rate Initial seed items
- Providing users with relevant recommendations based on genre

4.2.1 DEAM Dataset

DEAM dataset has been discussed in detail in section 3.2.3. We have also discussed our motivation to use DEAM instead of other popular datasets in the previous chapter. We can recall that the DEAM dataset was enriched with musical features *tempo*, *danceability* and *loudness*. This enriched dataset consists of 1802 musical excerpts provided by the dataset, meta-features like the artist, track name, genre and acoustic features valence, tempo, danceability and loudness. DEAM dataset consists of 3 subsets of data that were released from 2013 to 2015.

Data Merging

The DEAM dataset [102] has various subsets, each providing us with different information. The dataset has different subset with audio files, meta-features, valence and arousal annotations both song and user level and features that were extracted using openSMILE [109] for a 500ms window. In addition to this, acoustic features were extracted to enrich the dataset using Essentia [104]. This has been explained in detail previously in section 3.3.2.

We considered the following subsets provided by the DEAM dataset for our study. Each subset was in a form of CSV file.

- Valence- This subset contained song id, valence mean and valence standard deviation values.
- Meta features- This subset contains Song Id, Artist name, Song title and Genre of the song
- Acoustic Features- This subset contains the features that were extracted using Essentia. It contains Song Id, Beats per minute, Danceability and Loudness.

This data is hence divided into three data models as presented below. Figure 4.4, 4.5 and 4.6 describe the data models in detail.

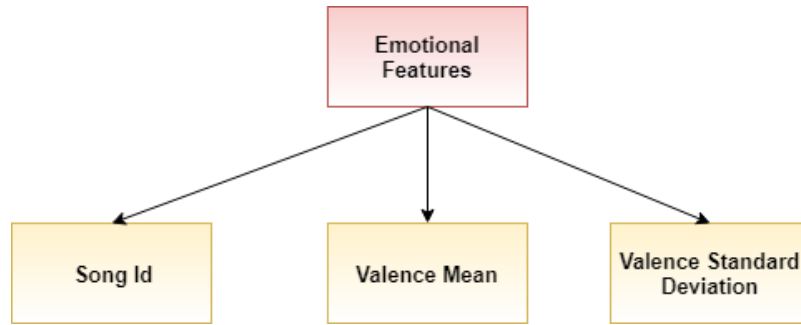


Figure 4.4: Valence values information of DEAM dataset

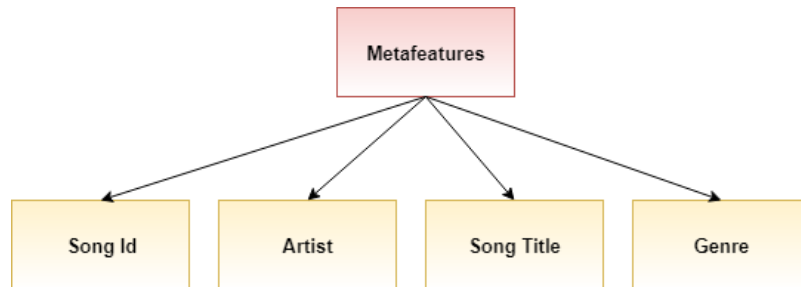


Figure 4.5: Metafeatures Information of DEAM dataset

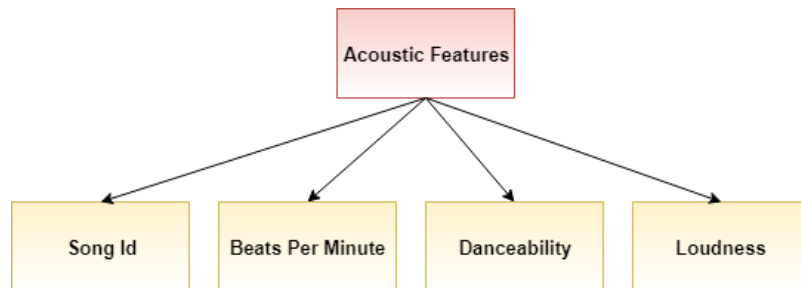


Figure 4.6: Acoustic features of DEAM extracted using Essentia

We can see that Song Id is common in Fig 4.4, 4.5 and 4.6. Thus, on the basis of song id different subsect were merged to form a complete dataset which was later used to train our recommendation system.

Merged Dataset

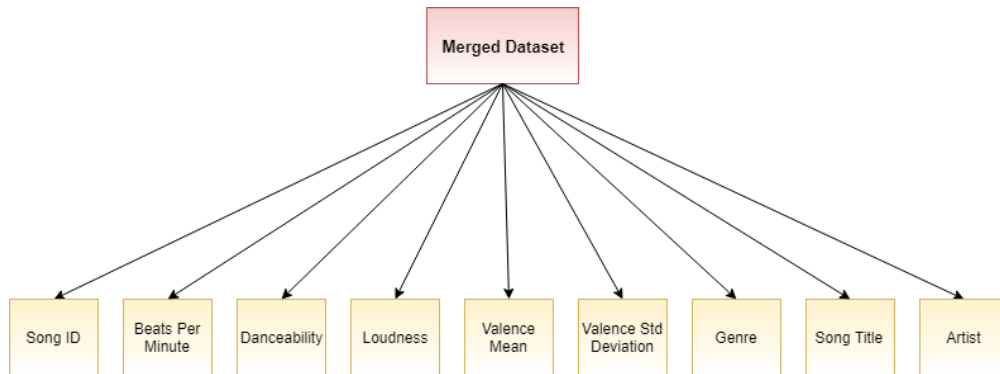


Figure 4.7: Final Merged DEAM Dataset

The final dataset should look like this after performing the necessary steps. Let's look at an example from the merged dataset.

Consider the song **Waterduct** that has been provided by DEAM.

Song Id	BPM	Loudness	Danceability	Valence Mean	Valence Std	Artist	Song Title	Genre
2005	114.89	0.26	1.01	3.8	1.17	Ava Luna	Waterduct	Rock

Table 4.1: Song Waterduct from our dataset

Data Pre-processing

The enriched dataset was preprocessed for genres, the title of the tracks and valence values. In this section, we discuss how the dataset was further preprocessed to be used.

- **Song Title details:** The dataset provided song titles in an inconsistent format. The song titles had values that contained all capital letters, special characters and other inconsistencies. This was brought to a structured format by text processing. We decoded basic Latin Unicode characters, changed the text from uppercase and lowercase to Titlecase where the initial character is capitalised and the other characters are lowercase.
- **Genre details:** The dataset contained songs from different genres. As discussed the dataset contained songs with many genre tags. These songs were processed to contain only the first genre tag as they represented the most probable genre of the song. The dataset had additional last.fm labels with its 2014 release. These labels were ignored for our study as our study did not focus on genre-based recommender systems.
- **Valence values:** As discussed in section 3.3.2, we know that dataset provided annotations for valence per song and per user. We considered valence per song as that met our requirements. In addition to that, we consider valence mean instead of considering valence standard deviation because we consider valence mean to represent valence of the song better. Moreover, the standard deviation is a measure of volatility and it is not the focus of our study.

After performing the above step, our final dataset had pre-processed song titles, genres of the songs and valence mean of the songs. In addition to this, our dataset also had features extracted from Essentia namely tempo, danceability and valence. We would like to bring it to notice that though for our experiment we modulated valence and danceability, the features tempo and loudness were used to obtain a coherent playlist. The reasons for taking this decision is described in Section 4.4.1

4.2.2 Initial Seed items

Selecting an initial set of songs that would build a basic user profile was an extremely important step. For our study, we build a basic content-based recommender system. The first set of recommendations are genre-based which will be later discussed in section 4.2.3. To select these songs we read some literature. One interesting work was that of Rashid et al [110] which talks about the importance of learning new user preferences in recommender systems. They have used MovieLens dataset for their study[111].

Rashid et al [110] discusses four important dimensions that one would need to form a strategy for forming an initial seed. These strategies are 1) User effort 2) User Satisfaction 3) Recommendation accuracy and 4) System utility. They also discuss other strategies like random selection and popularity.

The songs in our dataset are not popular and are song excerpts. In addition to this, our baseline recommender system is a genre-based recommender which will be discussed in detail in section 4.2.3. Thus, one of our goals was to learn about the user's genre preferences. To make this happen, we provide an initial seed of songs with different genres to learn about user's genre preference. In addition to this, we also learn about their preference for acoustic features beats per minute and loudness.

Song ID	Title	Artist Name	Genre
2	Tonight A Lonely Century	The New Mystikal Troubadours	Blues
2005	Waterduct	Ava Luna	Rock
346	Tennessee Hayride	Jason Shaw	Country
137	I	Aaron Dunn-Sonatina	Classical
378	Deep Sky Blue	Graphiqs Groove	Electronic
520	Broken Spell	Fit and the Conniptions	Folk
638	Winter Sunshine	Evgeny Grinko	Jazz
865	Love Me Like You	The Mythics	Pop
1606	Allhou	Salam	International
1530	Want You	Deal The Villain	Hip Hop

Table 4.2: Initial Seed Items

4.2.3 Recommendation Algorithm

In this section, we discuss the importance of a good recommendation technique. We later discuss content-based recommender systems and our baseline recommendation algorithm.

As discussed earlier, Recommendation Systems are search and decision tools which help us find relevant information. These systems achieve this by looking at the items that were rated/consumed by the user and hence calculating probable items that would be liked by the user. These recommendations can be generated using various techniques. Some of the basic recommendation techniques are popularity based recommendation technique, content-based recommendation technique and collaborative Filtering technique [112].

In our study, we use a basic content-based recommendation engine to generate our recommendations. Our baseline model is a genre-based system and we utilise a content-based recommendation model based on musical features later in our study.

For our study, we decided to the user a simple content-based recommender system that would calculate item similarities based on song genres and acoustic features that will be discussed later in the chapter 4.4. This decision was made because of the following reasons:

- Our dataset did not have any user ratings for songs. As discussed, the dataset had audio files, meta-features and valence arousal values. This limited us to use a content-based technique instead of a collaborative filtering technique. Studies show that collaborative Filtering is a more sophisticated technique than content-based but due to limited time to collect song ratings, we chose to go with content-based algorithm anyway for our baseline system.
- Our study tries to understand the relationship between user mood and latent song features. This required us to build a content-based system that would recommend songs based on musical features for our Mood specific recommender system.

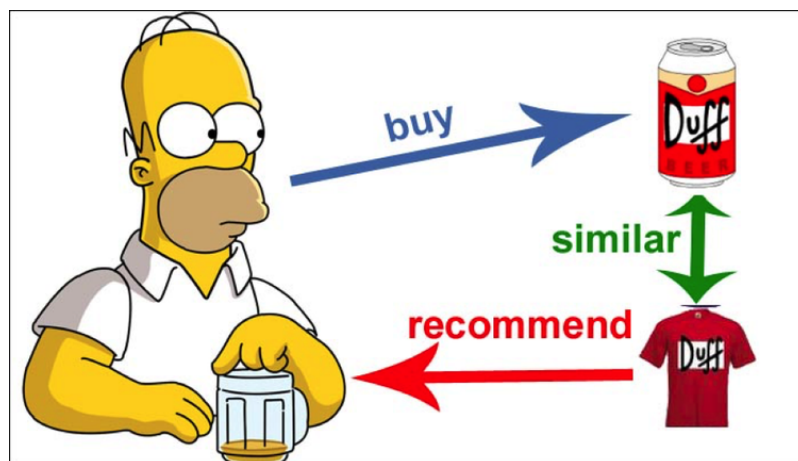


Figure 4.8: Example of a content based recommender system[2]

Figure 4.8 represents a content based recommender system. This an example of a content based system. Here, *Homer* buys *Duff* beer. So, a recommendation system

recommends it a Duff Tshirt because he might buy it as well.

Item Item Similarity

To provide recommendations in our study, we can compute similarity between two items based on some features and hence calculate a similarity score between items. Usually, similarity between items is calculated with help of similarity metrics cosine, euclidean distance or jaccard similarity.

For our baseline recommender system, we build a genre based recommender system. This is done in the following steps:

- Calculation of Tfidf vectors for genres
- Multiplying Tfidf vectors give cosine similarity scores between different songs based on genres. So, these vectors are multiplied to calculate similarity scores
- With a song as an index, similarity scores are calculated and a Top N list is built.
- Based on initial user ratings, we calculate similarities of each user rated song with a value greater than 3 to songs in our database. This list is then sorted based on similarity score and a Top N list is provided which is displayed in our genre recommender system

We further explain a basic genre-based recommender system where a user provides ratings for a few songs. Based on these ratings, similarity scores are computed between the songs rated and songs in our database. Next, a recommendation list with Top N score is presented by the genre recommender.

Note: Cosine Similarity metric was used for our genre-based recommender system because this metric has been proven to provide better similarity scores in case of text data. As our genres were in the form of a text, this was a natural choice.

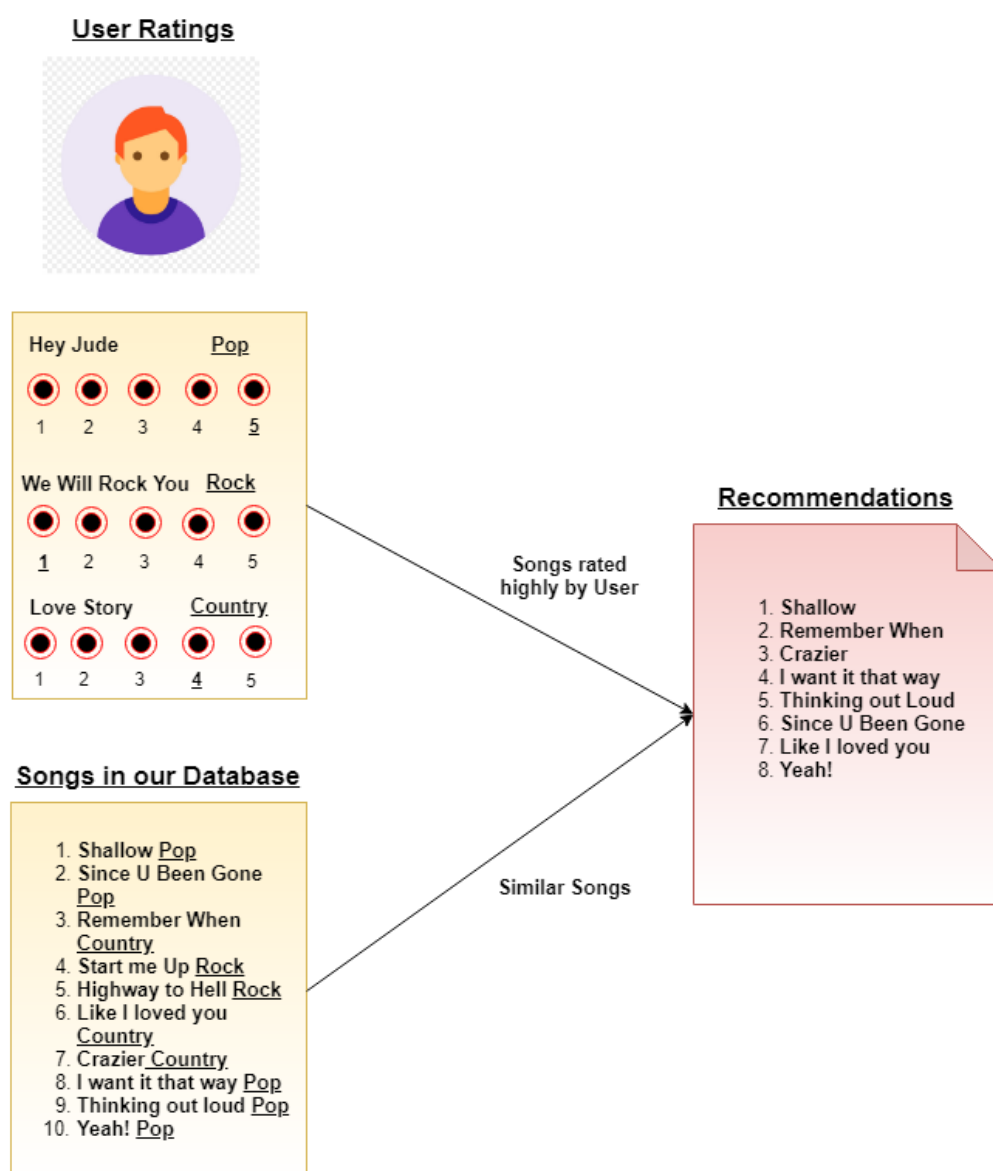


Figure 4.9: A genre-based music recommender system

Figure 4.9 provides us with an example of genre-based music recommender system. In the above example user provides ratings to songs *Hey Jude*, *We Will rock you* and *Love Story* which belong to Pop, Rock and Country genre respectively. Our database contains a list of songs from Pop, Rock and Country genres. Based on the rating provided by the user, one can see that the user likes Pop and country music and thus gives a rating of 5 and 4 to the music. On the other hand, user dislikes Rock genre and gives a rating of 1 to the song *We will rock you*.

Based on the user ratings, a user profile is created and recommendations are provided based on the genre ratings provided by him. We can notice that the recommendation list contains songs which are *Pop* and *Country* thus satisfying user's taste.

Note: For our experiment, we set a threshold of rating 3, and ratings above it were considered to be enjoyable for the users. This was done to create a user profile of songs the user actually enjoys listening and to reduce the complexities of the system.

4.2.4 Recommendation Process

In this section, we discuss the first phase of recommending songs to our participants in detail. Figure 4.10 shows the process flow of our baseline recommendation system.

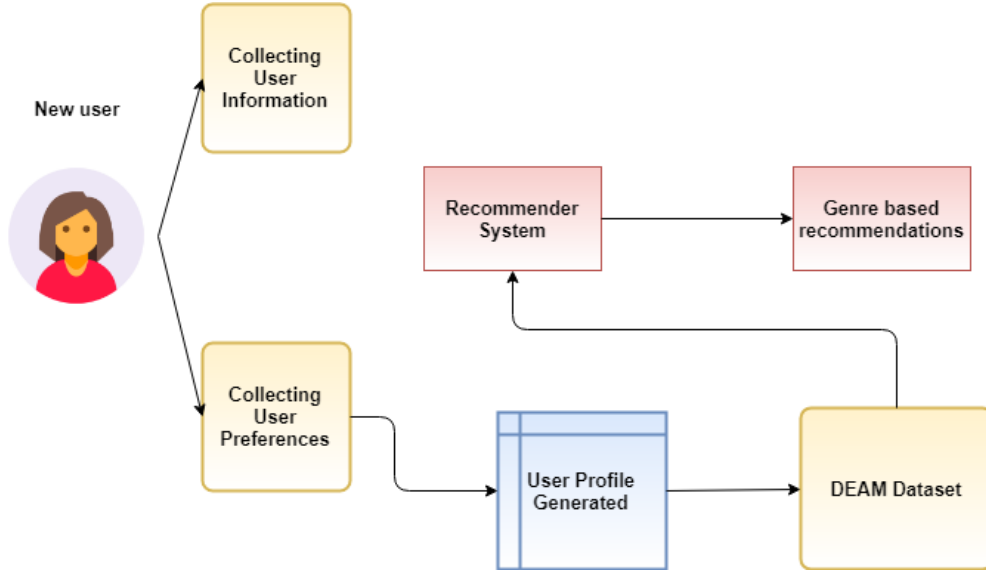


Figure 4.10: Recommendation Process for Baseline Recommendations

MooDify was developed using flask and is an interactive system which utilises the local webserver to navigate through the various steps of the system. Participants could first register into our system by answering some basic demographic questions. This is being displayed in Fig 4.11. After answering these questions, the user is asked to rate a set of 10 songs on a scale of 1 to 5 where 1 being least enjoyable by the user and 5 being most enjoyable by them as one can see in Fig 4.12. Users are informed about this in the Consent form. After collecting the ratings, a user profile is built for the active user using the system. Then, our recommender system uses Item-Item Content-based similarity based on the genre to calculate the similarity between songs rated by the user and the songs present in the DEAM data. Once the similarity values are computed, the user is recommended *Top 5* songs based on his *genre interests* which can be seen in Figure 4.13 .

MooDify

User Profile Questionnaire

Please answer the questions below to help us understand your user profile.

Gender

☐ Male
☒ Female
☐ Prefer not to answer

Age

☒ 18-24
☐ 25-34
☐ 35-44
☐ 45-49
☐ 50-55
☐ 56+

Nationality

Mention the country you identify yourself most with

Indian

We would like to understand your music behavior. How many hours do you use a Music Recommender System in a week(Spotify,Amazon Prime..)?

☐ I rarely use any Music Recommender System
☐ 1-3
☒ 4-6
☐ 7-10
☐ More than 10

Proceed

2/14

Figure 4.11: Demographics collected from user

MooDify

Kindly rate all the songs below on a scale from 1 to 5! The ratings should be provided on the basis of your music preference at the moment

Tonight A Lonely Century
Genre: Blues
1 2 3 4 5

Waterduct
Genre: Rock
1 2 3 4 5

Tennessee Hayride
Genre: Country
1 2 3 4 5

I
Genre: Classical
1 2 3 4 5

Deep Sky Blue
Genre: Postmodern
1 2 3 4 5

Broken Spell
Genre: Punk
1 2 3 4 5

Winter Sunshine
Genre: Jazz
1 2 3 4 5

Love Me Like You
Genre: Pop
1 2 3 4 5

Allhou
Genre: International
1 2 3 4 5

Want You
Genre: Hip_Hop
1 2 3 4 5

3/14

Proceed

Figure 4.12: Users asked to rate songs by MooDify

Recommendation List 1

Listen to all the songs from the Recommendation List 1 and Kindly form an opinion on how much you like the entire list. You will be asked few questions about it at the end of this page.

First

Genre: Classical

00:00 00:45

Impressions of Saturn

Genre: Classical

00:00 00:45

2

Genre: Pop

00:00 00:45

The Passing of Time

Genre: Classical

00:00 00:45

After Christmas

Genre: Classical

00:00 00:45

Kindly answer few questions on recommendation list 1 in your present state.

1. On a scale of 1 to 5, How much did you enjoy listening to songs in Recommendation List 1?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

2. On a scale of 1 to 5, How surprising were the songs in Recommendation List 1 for you?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

3. Kindly give your opinion on Recommendation List 1. If you do not have any opinion, type NaN in the textbox below.

Proceed

Figure 4.13: Genre based recommendations provided by MooDify
We see the different phases of our process as seen by the user in the Figures

4.3 Mood Information Phase

In this section, we discuss the second phase of our system Moodify. This phase is responsible for the following tasks:

- Inducing mood in participants
- Detecting mood induced in the participants from emotionally latent video clips via a Questionnaire

The process flow of Mood Information is described in Fig 4.14.

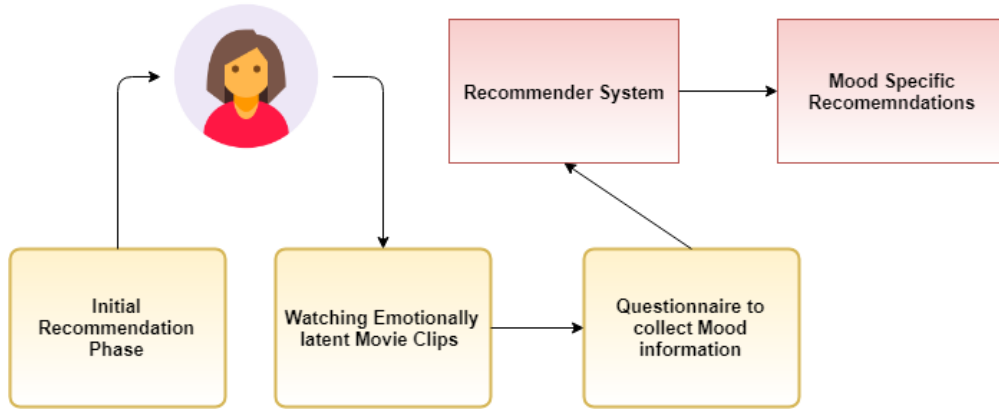


Figure 4.14: Process flow in Mood Information Phase

Users enter this phase after consuming items recommended by Genre-based recommender. They are shown a random happy clip from our data of video clips initially and are then provided with recommendations specific to happy emotion. In the second phase of mood information, the users are shown a sad video and later provided with recommendations specific to sad emotion. We discuss about this in detail in the next section 4.4

4.3.1 Mood Induction

There are different techniques for mood induction which are used in research. We have briefly discussed them in section 2.6.2.

In this section, we discuss the mood induction process that was used in our study. For our study, we use the technique of showing emotionally latent video clips to our participants for emotion induction.

We make use of Hewig's [93] clips for our study. As we are focusing on two basic emotions happy and sad, we consider the clips provided for happy and sad along with the neutral clip for bringing back people to a normal state. We used *Hannah and her Sisters* as our baseline stimuli inspiring from the work of Ferwerda et al [113]

Table 4.3 lists the clips which were used for our study

Video Clip	Corresponding Emotion
On Golden Pond	Amusement
An Officer and a Gentlemen	Amusement
When Harry met Sally	Amusement
Hannah and her Sisters	Neutral
An Officer and a Gentlemen	Sadness
The Killing Fields	Sadness
The Champ	Sadness

Table 4.3: Clips used for Emotion Induction

The clips with emotion *Amusement* were used to induce *Happiness* in users. We considered *Amusement* as emotion *Happiness* same as Ferwada et al [113] because we considered these clips to induce happiness. Finally, we had a set of 3 Happy clips, 3 Sad clips and 1 Neutral clip. These clips were randomly selected by our system for Mood Induction.

We can see some snippets of our system displaying the emotion video clip. Fig 4.15 represents a snapshot of An Officer and a Gentleman ie. happy video by our system.

Note: A different clip from An Officer and a Gentleman is also used for Sadness

MooDify

Video Player

Kindly watch the entire video before proceeding to the next page

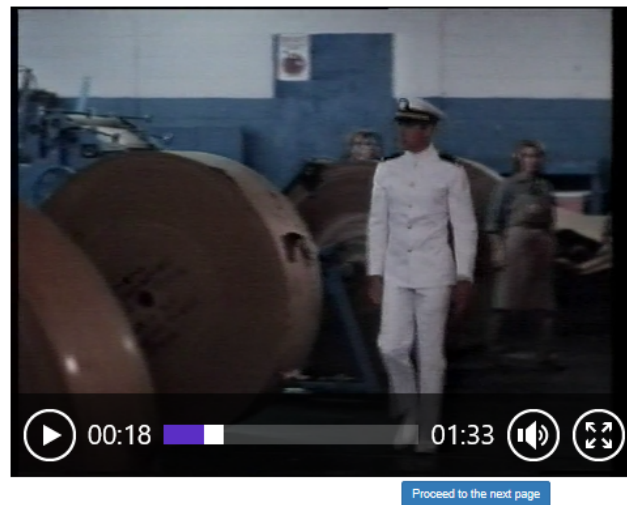


Figure 4.15: Happy video

4.3.2 Mood Detection

After the previous step of Mood Induction, our next step was to detect the emotion of the user. This would act as an input for our mood-based recommender system. We have previously discussed various mood detection techniques in section 2.6.3. We initially decided to work with an emotion recognition tool Affectiva due to its simplicity and cutting edge technology of facial emotion recognition. This tool could be integrated with our system or used as an application on mobile phones.

We did a bit of research before integrating Affectiva in our system for emotion detection. Some people suffer from a condition called Flat affect[114] where a person suffers from low expressiveness. In these conditions, people may not show normal signs of facial emotion. In addition to this, we realised a lot of people might not show their emotion on face especially when interacting with a system in a simulated environment.

Hence, we decided to conduct a pilot study which is discussed below.

Pilot Study

In this section, we briefly describe our pilot study which was conducted to decide if we are going to use Affectiva or a Questionnaire for Emotion Detection. The pilot study lasted 15-20 mins and had a number of 3 participants. These participants were students at TU Delft, between the age group 22-26, belonged to countries China, Taiwan and the United Kingdom and were female. The videos shown to the participants were chosen randomly from our video pool. The steps conducted in our pilot study can be described below :

- First, a set of videos are shown to the participants. We start with showing a *Happy Video*, then a *Neutral video* to bring back the participants to a normal state and lastly a *Sad Video*. The videos were selected from the set of videos in Table 4.3
- Then the reactions of the participants are recorded while they are watching the video and also after they have finished watching the video.
- After they have finished watching the video, they are given a questionnaire where they are asked *How are you feeling after watching the video clip?*. Their options are Ekman's basic emotions *Fear, Anger, Disgust, Sad, Happy and Surprise*.

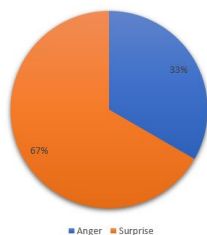


Figure 4.16: Responses for Happy Video via Affectiva

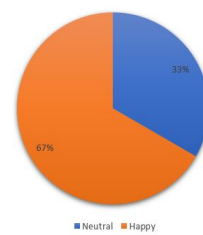


Figure 4.17: Responses for Happy Video via Questionnaire

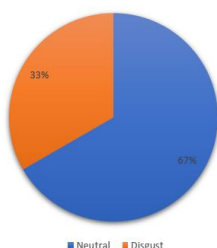


Figure 4.18: Responses for Neutral Video via Affectiva

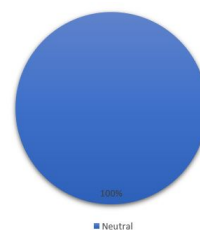


Figure 4.19: Responses for Neutral Video via Questionnaire

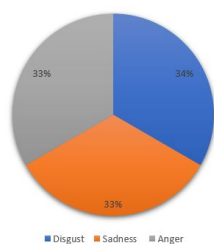


Figure 4.20: Responses for Sad Video via Affectiva

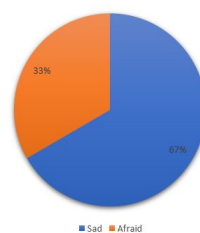


Figure 4.21: Responses for Sad Video via Questionnaire

After seeing the responses made by the participants for the video, we decided to use a Questionnaire for the emotion detection process. It seemed more reliable and was able to convey their emotion more than the emotion recognition tool Affectiva.

Questionnaire

As discussed before, we decided to use a questionnaire for emotion detection. After showing the movie clip, we asked the participants *How are they feeling?* This can be seen in Figure 4.22.

MooDify

Questionnaire

Kindly answer the below question

After watching the previous movie clip, How do you feel? Select the option that best suits your current mood.

☒ Happy ☐ Neutral ☐ Sad ☐ Angry ☐ Disgusted

Next

6/14

Figure 4.22: Questionnaire after showing a Happy Video

Note: The participants were provided with only 4 out of 6 Ekman's emotion. This was done because the movie clips in Hewig's clips for emotion induction had only clips for 5 emotions (It excluded surprise). Moreover, our pilot study indicated that people perceived sad clips as anger and disgust which made it important to include them. But, It was not the same for fear.

4.4 Mood based Recommendation Phase

In this section, we discuss in detail the process flow from the Mood Information Phase to Recommendation Phase. Then we discuss, our algorithm which generates mood specific recommendations in detail.

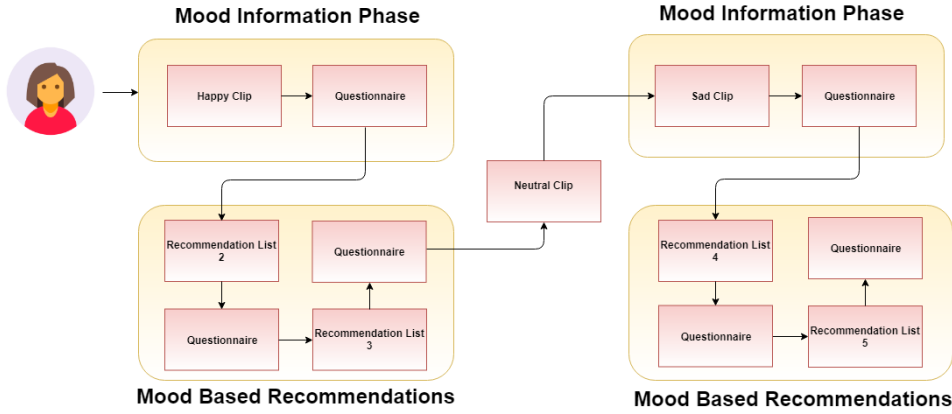


Figure 4.23: Process flow from Mood Information to Recommendations

Figure 4.23 describes the process flow from Mood Information Phase to Mood based Recommendation Phase.

This can be briefly explained in the following steps:

1. After consuming genre-based recommendations, the user enters the Mood Information Phase where he/she first sees a happy video
2. Next, they are asked a questionnaire about *How are you feeling after watching the video?*
3. After obtaining their current mood information, they enter our Mood Based Recommendation phase where they are presented with our recommendations specific to the happy user. These recommendations list is generated in a way that would test out Hypothesis for Happy users.
4. Next, a Neutral video is shown to the users to bring them back to a normal state before showing sad videos.
5. Now, the user is shown a sad video and is asked a question after that about *How are you feeling after watching the video*
6. After this, our final set of recommendations tailored for Sadness are shown to the user.
7. **Note:** The recommendations are specific to mood *Happy* and *Sad* irrespective of the user's actual mood after watching the video.

4.4.1 Recommendation Algorithm

We can recall from 1.2 our Research Objectives and corresponding Hypothesis that answers our Research questions.

Our Hypothesis questions are listed below:

- Happy Users would prefer Happy Music (Happy music- High Valence & High Danceability)
- Sad users would prefer Sad Music (Sad Music- Low Valence & Low Danceability)

For testing our Hypothesis, we designed our task in such a way that each user is provided with 4 recommendation lists in addition to the baseline Genre-based list.

The recommendation lists have features as described in Table 4.4

	Recommendation List 1	Recommendation List 2
Happy User	High Danceability & High Valence	Low Danceability & Low Valence
Sad User	Low Danceability & Low Valence	High Danceability & Low Valence

Table 4.4: Characteristics of Recommendation Lists

As described above, now our next step was to use build a Recommendation Algorithm which would help us generate the recommendations that would have characteristics as mentioned in Table 4.4

This task could be done by using either a Reranking function which would score the items according to the characteristics of the lists or by a Clustering approach where we can form clusters from songs based on the characteristics mentioned in Table 4.4.

For our task, we chose to use Clustering due to its simplicity and computational cost. Next, We describe our Clustering approach.

K means Clustering

K means Clustering is one of the simplest and popular unsupervised learning methods. Its goal is to form subsets of data by finding some underlying patterns in the data. The number of groups formed depends on the value of K given as input. It works in such a way that each data point is assigned to a group based on the features we provide as input by computing feature similarity. Figure 4.24 explains how K means clustering works

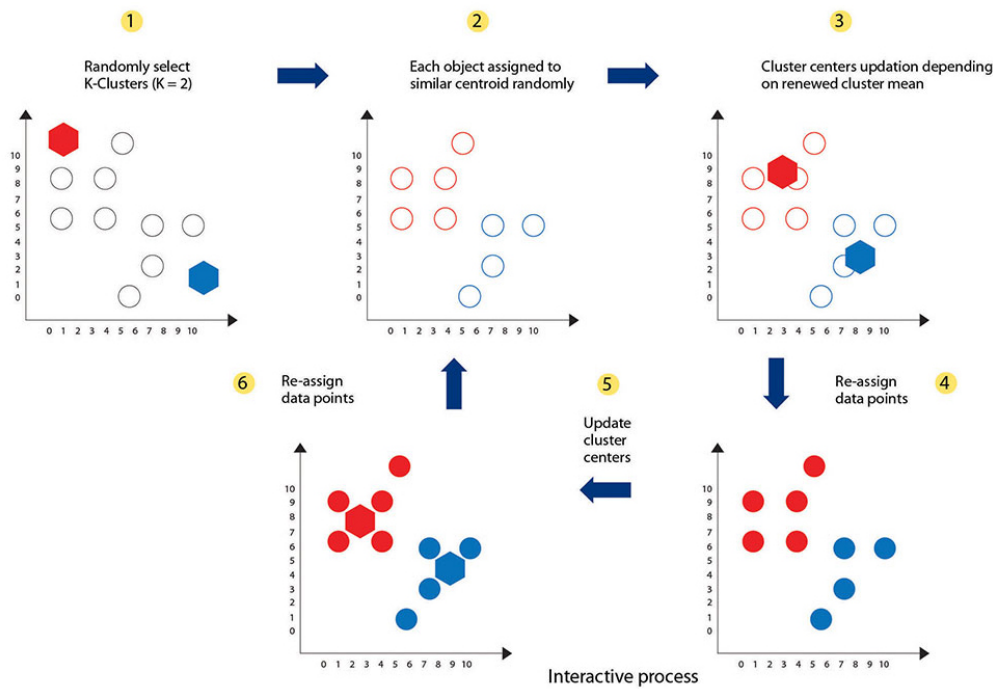


Figure 4.24: K Means Clustering Explanation[3]

We chose this technique to get labels in our dataset. We performed K means twice to obtain labels for data on features Danceability and Valence. Clusters were formed in such a way that the dataset was entire dataset was labelled *High Danceability*, *Low Danceability*, *High Valence* and *Low Valence*. This can be seen from the Figures 4.25, 4.26 and 4.27

Note: We performed the Clustering on scaled data. The data was scaled to be in the range of 0-1 before using it for our content based recommender.

	id	song_id	bpm	loudness	danceability	valence_mean	Artist	Song title	Genre
0	1	2005	0.461094	0.262009	0.104602	0.323529	Ava Luna	Waterduct	Rock
1	2	713	0.348325	0.983979	0.432686	0.485294	MarginalS	parte 1 parte 4	Jazz
2	3	1759	0.743502	0.806604	0.164567	0.558824	Terrero De Jesus	Oxum	Jazz-International-Experimental
3	4	773	0.372278	0.961554	0.183271	0.426471	U.S. Girls	The Island Song	Pop
4	5	1270	0.416378	0.972550	0.263216	0.441176	Voodoo Puppets	Intro	Country-Rock

Figure 4.25: Datframe without Labels

	id	song_id	bpm	loudness	danceability	valence_mean	Artist	Song title	Genre	Valence_Class
0	1	2005	0.461094	0.262009	0.104602	0.323529	Ava Luna	Waterduct	Rock	Low Valence
1	2	713	0.348325	0.983979	0.432686	0.485294	MarginalS	parte 1 parte 4	Jazz	High Valence
2	3	1759	0.743502	0.806604	0.164567	0.558824	Terrero De Jesus	Oxum	Jazz-International-Experimental	High Valence
3	4	773	0.372278	0.961554	0.183271	0.426471	U.S. Girls	The Island Song	Pop	Low Valence
4	5	1270	0.416378	0.972550	0.263216	0.441176	Voodoo Puppets	Intro	Country-Rock	Low Valence

Figure 4.26: Dataframe after obtaining Valence Labels

	id	song_id	bpm	loudness	danceability	valence_mean	Artist	Song title	Genre	Valence_Class	Danceability_class
0	1	2005	0.461094	0.262009	0.104602	0.323529	Ava Luna	Waterduct	Rock	Low Valence	Low Danceability
1	2	713	0.348325	0.983979	0.432686	0.485294	MarginalS	parte 1 parte 4	Jazz	High Valence	High Danceability
2	3	1759	0.743502	0.806604	0.164567	0.558824	Terrero De Jesus	Oxum	Jazz-International-Experimental	High Valence	Low Danceability
3	4	773	0.372278	0.961554	0.183271	0.426471	U.S. Girls	The Island Song	Pop	Low Valence	Low Danceability
4	5	1270	0.416378	0.972550	0.263216	0.441176	Voodoo Puppets	Intro	Country-Rock	Low Valence	High Danceability

Figure 4.27: Final Dataframe after obtaining Danceability Labels

Figure 4.25 represents the initial dataframe obtained after merging the dataset. This Data did not have any labels for both Valence and Danceability. We performed K means clustering to obtain these labels. After performing the first iteration of K means clustering we obtained a dataframe that is seen in Figure4.26. Another iteration of K Means is performed to obtain labels for Danceability. We obtain the dataframe seen in Figure4.27 after performing the second iteration.

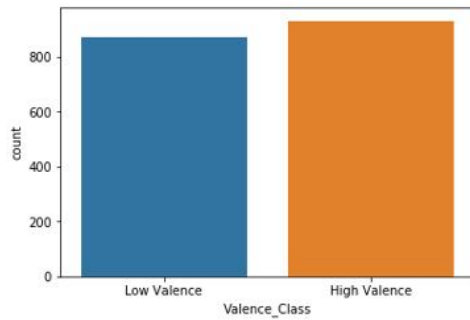


Figure 4.28: Distribution of Songs in Valence Clusters

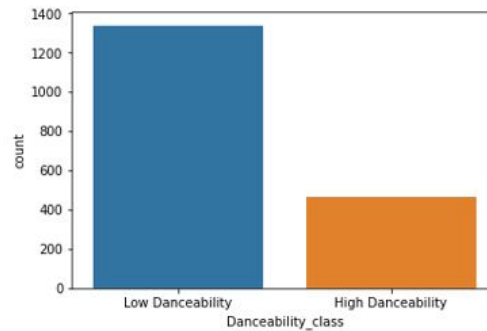


Figure 4.29: Distribution of Songs in Danceability Cluster

Figure 4.28 shows the the distribution of songs in Valence cluster and Figure 4.29 shows the distribution of sons in Danceability cluster. After obtaining the labels, we further formed subsets of the dataset such that each subset had a feature as represented in Table 4.4. Hence the dataset obtained was divided into 4 smaller datasets. This can be seen in Figure 4.30

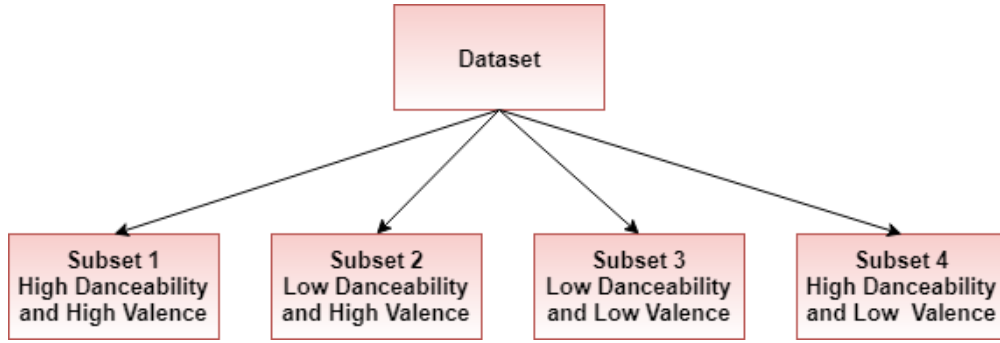


Figure 4.30: Subsets formed from the Dataset

We chose Kmeans clustering technique to divide our dataset in classes of high/low danceability and valence. This could have been done by using many other techniques. In similar scenarios, where one needs to divide the data, the most common technique is to divide it via mean or median of the data. In case of normal distribution of data, mean and standard deviation are the most popular approaches. In case of Non-normal distribution median, quartiles and tertiles are popular approaches for grouping data [115]. The distribution of valence and danceability can be seen in Figure 4.31 and 4.32. The distribution is not normal. The descriptive statistics of the data are as in Table 4.5. We decided not to go with upper and lower quartiles to preserve the data points. As discussed before, our data had only 1802 songs, and removing more songs from this list would be not feasible for training the recommender system model. Additionally, if we considered median smaller values would be a part of high danceability and valence. To avoid these scenarios, we decided to go with clustering approach which would form clusters automatically given the data points.

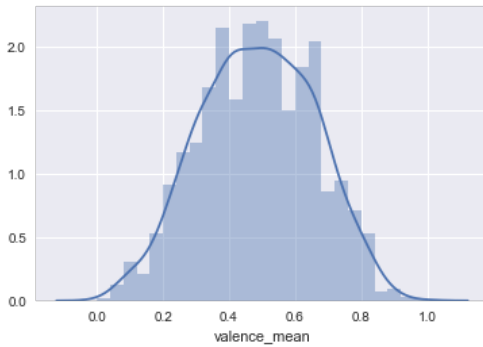


Figure 4.31: Distribution of Valence dataset

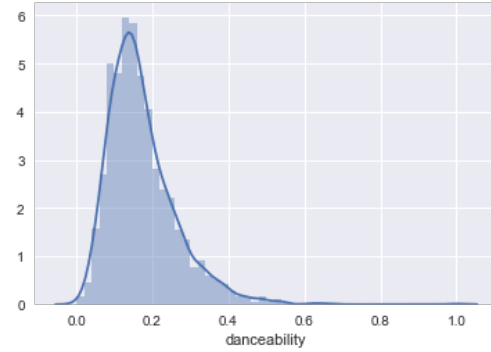


Figure 4.32: Distribution of Danceability in dataset

	Valence	Danceability
Mean	0.485	0.173
Median	0.485	0.154
Q1(0.25)	0.367	0.111
Q2(0.5)	0.485	0.154
Q3(0.75)	0.617	0.215

Table 4.5: Descriptive Statistics for Valence and Danceability

Item Item Similarity

After obtaining the subsets, our next step was to provide recommendations to the users. For this task, we decided to use an Item Item similarity algorithm.

We can recall from section 4.2.3 how Item Item similarity is calculated. For this task, we wanted to calculate the similarity between songs based on features Tempo and Loudness. This decision was made after conducting some study.

We studied the need for homogeneity of musical features in a playlist before designing our recommender system for generating Mood-based Recommendations. We look into the features of a playlist that would make it enjoyable for the user. Jannach [116] studied the characteristics homogeneity, diversity and freshness which would make a playlist enjoyable for the user. Their findings suggest homogeneity of the features loudness and tempo in a playlist. Another study by the same researchers [117] suggests the same findings.

Thus, we decided to have a playlist which will have similarity with the highly-rated Initial seed items in terms of *Tempo* and *Loudness*.

Finally, to generate our Mood Specific recommendations, We used an item-item similarity model. This was done in the following steps:


- As discussed previously, a user profile is created by user ratings from initial seed items.
- Now, the user profile is considered to look into user's musical taste.
- For similarity calculation of songs, we only consider the songs that are highly rated by the user ie a rating of 4 and 5 are considered. Let's call these set of songs as *Song_df*. This decision was made because we want to recommend songs that would be similar to the user's taste.
- Next, depending on the phase user is in, similarity values are calculated using Euclidean distance. This distance is calculated between *Song_df* and the songs in different subsets using features *Tempo* and *Loudness*.
- Then the songs are sorted based on similarity value from most similar to least similar.
- After obtaining similarity values, the user is recommended *Top 5* songs from the subset.

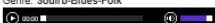
Note: The recommendations are generated based on the phase the user is in. Eg First, the user is recommended List 2 where similarity value is calculated between Song_df and Subset 1. We would also like to mention that we used euclidean distance to calculate similarity between songs as we were using the features tempo and loudness to calculate similarity which were already scaled(refer to Section 4.2.1).


The recommendation list 1 generated for the happy user can be seen in Figure 4.33.


MooDify

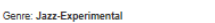
Recommendation List 2
Listen to all the songs from the Recommendation List 2 and Kindly form an opinion on how much you like the entire list. You will be asked few questions about it at the end of this page.

The River
 Genre: Folk


Sirens
 Genre: Soulrb-Blues-Folk


Benedictus
 Genre: Classical


Moonlight and Roses
 Genre: Electronic


I
 Genre: Jazz-Experimental


Kindly answer few questions on recommendation list 2 in your present mood.

1. On a scale of 1 to 5, How much did you enjoy listening to songs in Recommendation List 2?
☐ 1 ☒ 2 ☐ 3 ☐ 4 ☐ 5

2. On a scale of 1 to 5, How surprising were the songs in Recommendation List 2 for you?
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☒ 5

3. Kindly give your opinion on Recommendation List 2. If you do not have any opinion, type NaN in the textbox below.

NaN

7/14

Figure 4.33: Recommendations generated for List 2

4.5 Conclusion

In this chapter, we discussed the main phases of our recommender system. We can recall the phases to be *Initial Recommendation Phase*, *Mood Information Phase* and *Mood Recommendation Phase*.

We hereby discuss the key points of the chapter:

1. **Initial Recommendation Phase** - This phase was responsible for performing the following tasks:
 - Collecting user demographics and asking them to rate an initial set of songs to create a user profile
 - Generating genre-based recommendations (baseline) based on user profile
2. **Mood Information Phase** - This phase was responsible for performing the following tasks:
 - Showing video clips for mood induction and later asking questions to know their current mood

3. **Mood Recommendation Phase-** This phase was responsible for performing the following tasks:

- Providing user recommendations based on mood *happy and sad*
- It should be recalled that the recommendations were specific to the latent musical features. The recommendations were provided in such a way that a *Happy user* was recommended two lists *High Danceability & High Valence* and *Low Danceability & High Valence*. Similarly, *Sad users* were recommended with lists with *Low Danceability & Low Valence* and *High Danceability & Low Valence*

Additionally, we discussed the recommendation algorithm used for our system, an item-item similarity algorithm. We also discussed clustering algorithm which was performed to divide the dataset into subsets which would be then used to make mood specific recommendations. In the next chapter, we test the tests that were carried out to test our hypothesis.

Chapter 5

Evaluation 1 - To study the relationship between mood and latent musical features

5.1 Introduction

In the previous chapter, we discussed our System in detail. We have discussed previously creation of our dataset, our recommendation phases and corresponding recommendation algorithms.

In this chapter, We will try to answer our research questions restated below:

1. RQ1: How does mood affect the user's consumption of latent musical features Danceability and Valence
2. RQ2: Does mood impact how surprising the user finds the items recommended by the system?

To answer our research questions, we performed an online evaluation that could answer our Hypothesis, thus helping us answer our research question.

Hence, In this chapter, we discuss the procedure that was carried out to answer the research questions. We begin with explaining our Experimental design (Section 5.2), Variables and Hypothesis(Section 5.3 5.4 and Section 5.5). This is followed by our Procedure (Section 5.6), Demographics of Participants (Section 5.7) and conclude with Results and Discussion and Limitations of our approach in Section 5.8.

5.2 Design

In this chapter, we focus on studying the relationship between user mood and their music preference. For our task, We assume that a *Happy Song* will have musical features *High Danceability & High Valence* and a *Sad Song* will have musical features *Low Danceability & Low Valence*. We used a **within subjects** design for our experiment where each participant was presented with all recommendation lists. For each recommendation list, the participant was asked a few questions. We can recall it from Figure 4.33.

5.3 Independent Variable

For each participant, We show five different recommendation lists. The first one being a genre-based recommendation list and others specific to mood. In our study, we focus on recommendation lists that are being recommended to a *Happy and Sad* user. As discussed before, our focus is to find a relationship between Latent musical features(which are characteristics of the recommendation list) and Emotion. We achieve this by analysing user satisfaction and unexpectedness ratings given by the participants when they are provided with mood specific recommendations for a given emotional state.

Thus, the features *Danceability* and *Valence* are our Independent Variables.

5.4 Dependent Variable

As discussed before, we study the impact of mood on User satisfaction and Unexpectedness when the users are provided with Recommendation lists tailored to the Latent features. Thus, we have two Dependent variables:

1. **User Satisfaction**- User Satisfaction is measured by asking users questions like *How much did you enjoy listening to the recommendation list?*. This is answered on a Likert scale. This question answers user's satisfaction towards the recommended list.
2. **Unexpectedness**- Unexpectedness measures how surprising the recommendations were for the users. This is achieved by asking the user questions like *How Surprising were the songs in the recommendation list for you?*. This is also answered on a Likert scale.

5.5 Hypothesis

To answer our research questions, We formulated two main Hypothesis as listed below:

- **H1:** *Happy Users would prefer List 2 recommendations(High Danceability & High Valence) over List 3 recommendations(Low Danceability & High Valence)*
- **H2:** *Sad Users would prefer List 4 recommendations(Low Danceability & Low Valence) over List 5 recommendations(High Danceability & Low Valence)*
- **H3:** *Happy Users would find List 3 (Low Danceability & High Valence) recommendations more surprising than List 2 (High Danceability & High Valence) recommendations*
- **H4:** *Sad Users would find List 5 (High Danceability & Low Valence) recommendations more surprising than List 4 (Low Danceability & Low Valence) recommendations*

5.6 Approach

Each participant interacts with our system - Moodify and thus goes to a number of steps for evaluation procedure. These steps have been discussed in detail below:

1. The first step was to take consent from the participants for their free will to participate in the experiment. Thus, the participants are shown a consent form which gives them a brief overview of the experiment. After they agree to participate, they proceed to Step 2.
2. In the second step, we collect some basic demographic information about the participant by asking a few questions. These questions can be seen in Table 6.1

Index	Question
1	Gender
2	Age Group
3	Which Nationality do you identify yourself most with?
4	How many hours do you use a Music Recommender System in a week?

Table 5.1: Demographic questions

3. The third step was to form a user profile. This was achieved by showing the participants with an initial set of items and asking them to rate the songs on a scale from 1 to 5.
4. The fourth step involves a user's interaction with the genre-based recommender system.
5. The fifth step involves user seeing a video clip meant for emotion induction.
6. The sixth step involves user's interaction with the Mood specific recommender system which recommends songs based on latent features **Danceability** and **Valence**

The fifth and sixth step mentioned above is conducted for both *Happy* and *Sad* mood. Hence the process is repeated twice for each participant. Each participant is first shown a happy video clip, then he is provided with recommendations List 2 and List 3. Kindly refer to Section 4.4 for more information on these recommendation lists. Next, the participant is shown a neutral video clip to neutralise the participant. After showing a neutral clip, the participant is shown a sad video clip followed by recommendation list 4 and 5.

Kindly recall that these recommendation lists differ with respect to values of latent features *danceability* and *valence*. Refer to Table 4.4 to know about the characteristics of these lists.

5.7 Participants

For our evaluation, we recruited 15 participants. 13 out of 15 participants are students at the *Delft University of Technology*. 40% of the participants were male (n=6) and 60% of the participants were female (n=9) as seen in Figure 5.2. All of these participants belonged to the Age group 18-34. 73% of the participants (n=11) belonged to the age group 18-24 and 27% of the participants (n=4) belonged to the age group 25-34 as seen in Figure 5.3. In addition to this, we tried our best to recruit participants from different nationalities. The distribution of nationality can be seen in Figure 5.1. We had 12 different nationalities for our study.

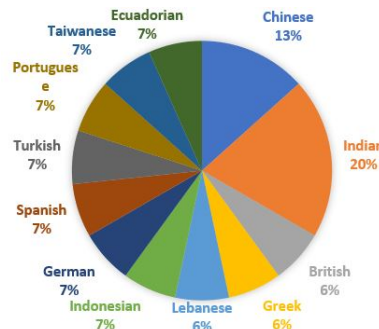


Figure 5.1: Participant Demographics by Nationality

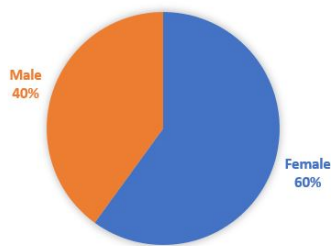


Figure 5.2: Participant Demographics by Gender

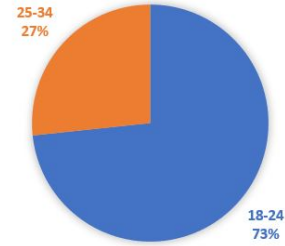


Figure 5.3: Participant Demographics by Age

We further looked into user's consumption behaviour of music recommender system. The results were very diverse with 33% of the users using a music recommender system for 4 to 6 hours a week, 27% of the users using using a music recommender for 10 hours and 1 to 3 hours per week respectively. 13% of our participants never use a music recommender system.

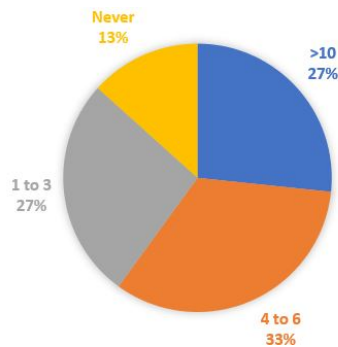


Figure 5.4: Participant Demographics by their consumption of Music Recommender Systems

5.8 Statistical Tests and Results

In this section, we discuss the results of our user-centric evaluation aimed at understanding the relationship between user mood and consumption of latent music features. We conduct some statistical tests to test our Hypothesis and discuss results for the same.

5.8.1 User Satisfaction for Happy Users

H1: *Happy Users would prefer List 2 recommendations(High Danceability & High Valence) over List 3 recommendations(Low Danceability & High Valence)*

To test our first hypothesis, we collect their answers for question: *How much did you enjoy listening to the recommendation list?* for both List 2 and List 3.

We collect ratings given by each participant for our recommendation lists and then plot a frequency vs ratings graph.

Figures 5.5 and 5.6 show the frequency of enjoyment ratings given by users to the recommendation lists List 2 and List 3

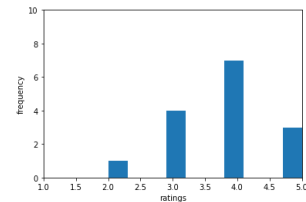
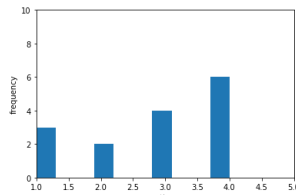


Figure 5.5: Enjoyment Ratings of Users for List 2 recommendations
Figure 5.6: Enjoyment Ratings of Users for List 3 recommendations

After obtaining the user ratings, We would like to test for our Hypothesis. Before testing our Hypothesis, we first test for normality of our ratings. This would help us in deciding the statistical test we would need to perform to test our hypothesis. In this case, the participant pool remains the same for both List 2 and List 3 recommendations. Hence, we already know that we would need a paired-samples t-test. To conduct this test, we first test the normality of the data distribution. To test the normality of enjoyment ratings of List 2 and List 3 we conducted both Visual tests and Statistical tests.

Visual Tests

As discussed, we conducted Visual tests to test the normality of the distribution.

Figure 5.7 and 5.8 shows a fitted gaussian distribution over histogram of enjoyment ratings. We can clearly notice that the histogram plot **does not follow a normal distribution**.

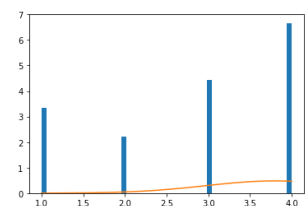
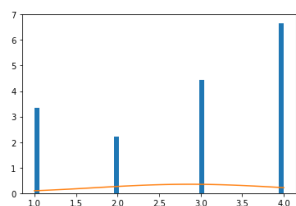


Figure 5.7: Gaussian Plot for List 2 enjoyment ratings
Figure 5.8: Gaussian plot for List 3 enjoyment ratings

We further plot a Q-Q plot for both List 2 and List 3 recommendations. From the Figures 5.9 and 5.10, we can notice that the data points do not lie on the slope for both List 2 and List 3, hence are **not normally distributed**.

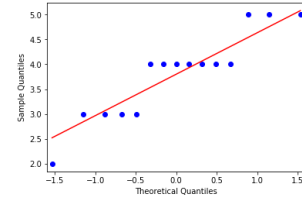
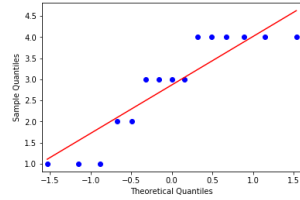


Figure 5.9: Q-Q Plot for List 2 enjoyment ratings Figure 5.10: Q-Q plot for List 3 enjoyment ratings

Statistical Tests

We further confirmed the non-normality by performing a **shapiro wilk test** [118] which has been shown in Table 5.2

Recommendation list	Statistic Value	P value	Conclusion
List 2	0.819	0.006	Sample does not look gaussian
List 3	0.882	0.05	Sample looks gaussian

Table 5.2: Statistical Test for List 2 and List 3 enjoyment ratings

After concluding that the ratings follow a non-normal distribution, We decide to perform **Wilcoxon's signed-rank** test on our data.

Table 5.3: Wilcoxon's signed rank for List 2 and List 3 ratings

		W	p	Rank-Biserial Correlation
List2	List3	19.500	0.037	-0.629

The results shown in Table 5.3 show that our p value is significant indicating that the distribution of ratings is different in both the list. We also notice that the effect size is high (-0.629) indicating that the emotion of the user had an impact on the ratings. The descriptive statistics of the ratings are shown in Table 5.4:

Table 5.4: Descriptive Statistics of recommendation List 2 and List 3

	N	Mean	SD	SE
List2	15	2.867	1.187	0.307
List3	15	3.800	0.862	0.223

Results from Table 5.4 shows that Happy Users prefer listening to recommendation List 3 than List 3 **rejecting H1**. We also see that variance in list 2 is more than list 3 indicating that some users highly disliked the recommendation list while some users really liked it.

5.8.2 User Satisfaction for Sad Users

H2: *Sad Users would prefer List 4 recommendations(Low Danceability & Low Valence) over List 5 recommendations(High Danceability & Low Valence)*

To test our first hypothesis, we collect their answers for question *How much did you*

enjoy listening to the recommendation list? for both List 4 and List 5 as we did in Section 5.8.1.

Figures 5.11 and 5.12 show the enjoyment rating distribution for List 4 and List 5 recommendations

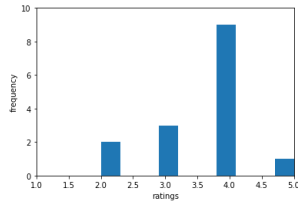


Figure 5.11: Enjoyment Ratings of Users for List 4 recommendations

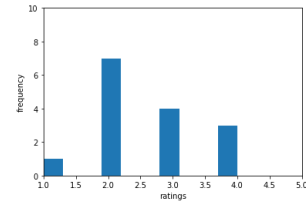


Figure 5.12: Enjoyment Ratings of Users for List 5 recommendations

After obtaining the user ratings, We would like to test for our Hypothesis. Before testing our Hypothesis, we first test for normality of our ratings. This would help us in deciding the statistical test we would need to perform to test our hypothesis. In this case, the participant pool remains the same for both List 4 and List 5 recommendations. Hence, we already know that we would need a paired samples t-test. To test this, we first performed some normality tests. This has been discussed in detail in sections Visual Tests and Statistical Tests.

Visual Tests

We conducted Gaussian Normality and Quantile plot test to test the normality of data distribution in enjoyment ratings of List 4 and List 5.

Figure 5.13 and 5.14 shows a fitted gaussian distribution over histogram of enjoyment ratings. We can clearly notice that the histogram plot **does not follow a normal distribution**.

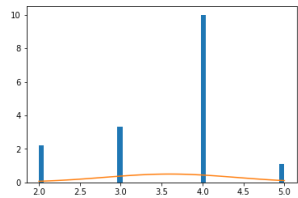


Figure 5.13: Gaussian Plot for List 4 enjoyment ratings

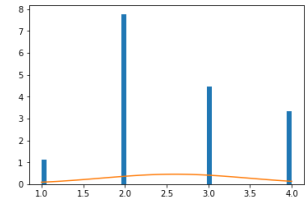


Figure 5.14: Gaussian plot for List 5 enjoyment ratings

We further plot a Q-Q plot for both List 4 and List 5 recommendations to test their normality. From the Figures 5.15 and 5.16, we can notice that that the data points do not lie on the slope for both List 2 and List 3, hence are **not normally distributed**.

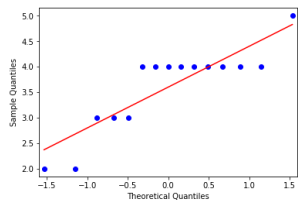


Figure 5.15: Q-Q Plot for List 4 enjoyment ratings

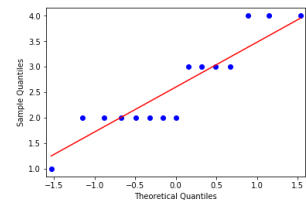


Figure 5.16: Q-Q plot for List 5 enjoyment ratings

Statistical Tests

Next, a **shapiro wilk** test was performed to test the normality of data distributions for List 4 and List 5 enjoyment ratings. Unlike List and List 3, the statistical tests in this case clearly show that both List 4 enjoyment ratings and List 5 enjoyment ratings are **Non-Normal**

Recommendation List	Statistic Value	P value	Conclusion
List 4	0.805	0.0004	Sample does not look gaussian
List 5	0.861	0.025	Sample does not look gaussian

Table 5.5: Statistical Test for List 4 and List 5 enjoyment ratings

After obtaining results from both Visual and Statistical tests, we conclude that both List 4 and List 5 enjoyment ratings follow a **Non-Normal** distribution. This was achieved by performing a **Wilcoxon's signed rank** test on our data.

Table 5.6: Wilcoxon's signed rank for List 4 and List 5 ratings

		W	p	Rank-Biserial Correlation
List4	List5	84.500	0.041	0.610

The results shown in Table 5.6 show that our p value is significant indicating that the distribution of ratings is different in both the list. We also notice that the effect size is high (0.610) indicating that the emotion of the user had an impact on the ratings.

Table 5.7: Descriptive Statistics of List 4 and List 5 enjoyment ratings

	N	Mean	SD	SE
List4	15	3.600	0.828	0.214
List5	15	2.600	0.910	0.235

These results in Table 5.7 show that the sad users enjoyed listening to recommendation List 4 more than recommendation List 5 thus accepting our Hypothesis **H2**.

5.8.3 Unexpectedness 1: For Happy Users

H3: *Happy Users would find List 3 (Low Danceability & High Valence) recommendations more surprising than List 2 (High Danceability & High Valence) recommendations*

To test our hypothesis **H3**, We consider responses given by the users to the question *How Surprising were the songs in the recommendation list for you?* for both recommendation List 2 and recommendation List 3.

Further, these responses are collected for each participant. Figure 5.17 and Figure 5.18 represent the frequency of user ratings for surprisingness of the recommendation list 2 and 3.

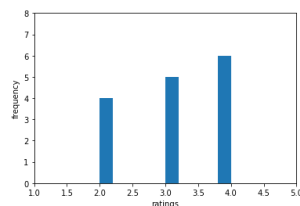


Figure 5.17: Surprising Ratings of Users for List 2 recommendations

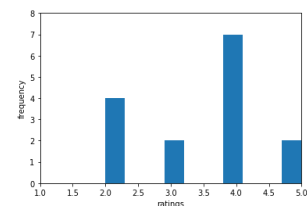


Figure 5.18: Surprising Ratings of Users for List 3 recommendations

For conducting any test, we first need to check the normality of the data distribution. This is achieved by conducting Visual and Statistical normality tests for List 2 and List 3 surprising ratings.

Visual Tests

From Figure 5.19 and Figure 5.20, we can see that the data is not normally distributed. But we do some additional tests.

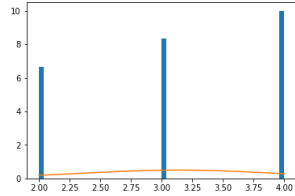


Figure 5.19: Gaussian Plot for List 2 surprising ratings

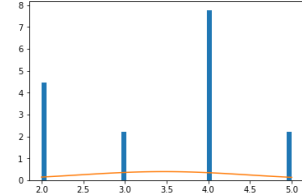


Figure 5.20: Gaussian plot for List 3 surprising ratings

After testing for gaussian plot distribution, we plotted a quantile plot. Figure 5.21 and Figure 5.22 also indicate Non-Normality of the data.

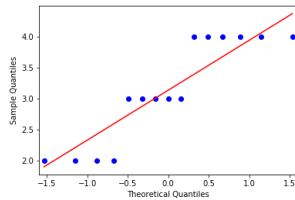


Figure 5.21: Q-Q Plot for List 2 enjoyment ratings

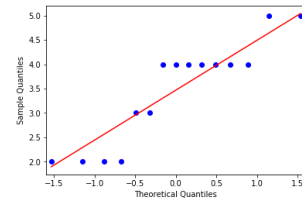


Figure 5.22: Q-Q plot for List 3 enjoyment ratings

Statistical Tests Additionally, we also performed a **shapiro wilk** test to test the normality of the data.

Recommendation list	Statistic Value	P value	Conclusion
List 2	0.799	0.004	Sample does not look gaussian
List 3	0.840	0.013	Sample does not look gaussian

Table 5.8: Statistical Test for List 2 and List 3 surprise ratings

From the above results, We conclude that both the recommendation lists follow Non-Gaussian distribution.

After obtaining the results, we further test for our Hypothesis **H3**. To achieve this, We perform a *Wilcoxon's signed-rank* test.

Table 5.9: Wilcoxon's signed-rank for List 2 and List 3 surprising ratings

		W	p	Rank-Biserial Correlation
List2	List3	34.500	0.450	-0.242

The above results indicate that the Null Hypothesis rejection is not statistically significant. That means that the mood of the user, in this case, happiness does not impact how the user rates the music recommendations.

Nevertheless, we checked for the mean surprisingness ratings given by the user for List 2 and List 3. This has been displayed in Table 5.10

The mean ratings indicate that the users found recommendation list 3 more surprising which is inline with our hypothesis but as the Wilcoxon's test shows that the

Table 5.10: Descriptive Statistics of List 2 and List 3 surprising ratings

	N	Mean	SD	SE
List2	15	3.133	0.834	0.215
List3	15	3.467	1.060	0.274

difference is not statistically significant and the effect size is small (-0.242), we **reject** our Hypothesis **H3**.

5.8.4 Unexpectedness 2: For Sad Users

H4: *Happy Users would find List 5 (High Danceability & Low Valence) recommendations more surprising than List 4 (Low Danceability & Low Valence) recommendations*

We follow the same procedure as we did for other tests here. First, we collect user rating for the question *How Surprising were the songs in the recommendation list for you?* for both recommendation lists 4 and 5.

Next, we plot a graph for ratings given for surprisingness and the frequency of the rating. This has been shown in Figure 5.23 and Figure 5.24.

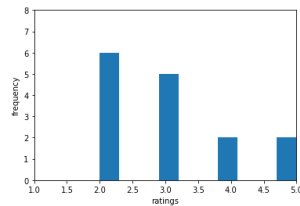


Figure 5.23: Surprising Ratings of Users for List 4 recommendations

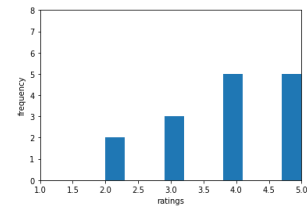


Figure 5.24: Surprising Ratings of Users for List 5 recommendations

As we are trying to compare the surprise ratings of list 4 and list 5 by the same population, we know that we need to perform a paired sample t-test. To check the normal or non-normal version of the test, we first need to check the normality of the data distribution. This is achieved by conducting Visual and Statistical normality tests for List 4 and List 5 surprising ratings.

Visual Tests

From Figure 5.25 and Figure 5.26, we can notice that the data follows a **Non-Normal distribution**

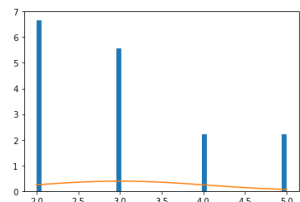


Figure 5.25: Gaussian Plot for List 4 surprising ratings

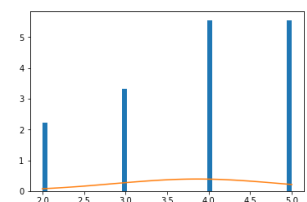


Figure 5.26: Gaussian plot for List 5 surprising ratings

After testing for gaussian plot distribution, we plotted a quantile plot. Figure 5.27 and Figure 5.28 also indicate **Non-Normality** of the data.

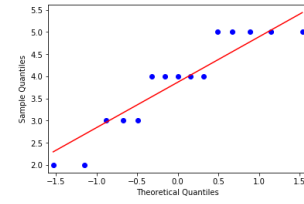
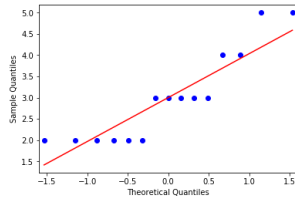


Figure 5.27: Q-Q Plot for List 4 enjoyment ratings Figure 5.28: Q-Q plot for List 5 enjoyment ratings

Statistical Tests

Additionally, we also performed a shapiro-wilk test to test the normality of the data.

Recommendation List	Statistic Value	P value	Conclusion
List 4	0.826	0.008	Sample does not look gaussian
List 5	0.862	0.026	Sample does not look gaussian

Table 5.11: Statistical Test for List 4 and List 5 surprise ratings

From the above results, we conclude that the data is **Non gaussian**. We further perform a *Wilcoxon's signed-rank* test on our data. The results of the test are as shown in Table 5.12

		W	p	Rank-Biserial Correlation
List4	List5	21.000	0.045	-0.600

Table 5.12: Wilcoxon's signed-rank test for List 4 and List5 surprise ratings

We can see from Table 5.12 that our p-value is statistically significant indicating that the rating distributions are different. We also notice the effect size is high(-0.6) indicating that the emotion of the user (here sadness) has an impact on the surprise ratings. We further look into the descriptive statistics of ratings which are shown in Table 5.13

	N	Mean	SD	SE
List4	15	3.000	1.069	0.276
List5	15	3.867	1.060	0.274

Table 5.13: Descriptive statistics of List 4 and List 5 surprise ratings

Results from Table 5.13 show that the users found List 5 more surprising than List 4 hence accepting our Hypothesis **H4**

5.8.5 Additional Analysis

After testing for our main hypothesis questions, we wanted to test for some additional things. One can recall that the participants were provided with a genre-based recommender list at the beginning before seeing mood specific recommendations. After testing for our Hypothesis for user satisfaction **H1** and **H2**, we wanted to compare user satisfaction for genre-based recommender system and mood specific recommender system. One can recall that the population for both the recommenders was the same. Hence, we have to perform a paired sample t-test. The recommendation lists for mood-based recommender has already been proven to Non-Normal (refer to Section 5.8.1 and 5.8.2). The genre enjoyment ratings can be seen below.

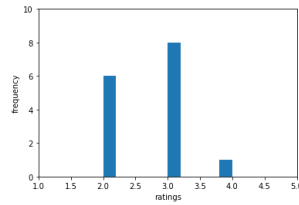


Figure 5.29: Enjoyment ratings for genre based recommender system

Further, Normality tests are performed for the genre ratings which can be seen below.

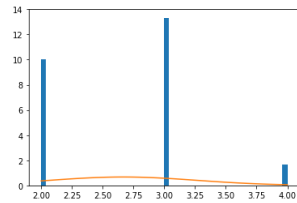


Figure 5.30: Distribution plot for Genre ratings

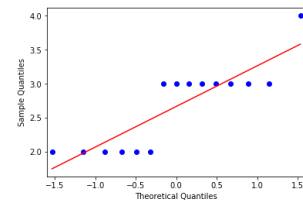


Figure 5.31: Q-Q plot for genre ratings

As we can clearly see that the genre ratings follow a non-normal distribution we perform non-parametric tests to compare the user satisfaction.

Comparing Genre and Happy ratings

We compared ratings given by the participants for genre based music recommendations and recommendations of list 2 and list 3 with features high valence & danceability and high valence & low danceability respectively. We wanted to find out the recommendation list preferred by the participants.

From Table 5.14, we can see that p value for genre and list 2 ratings is insignificant (0.696). The descriptive statistics also show a small difference between the mean ratings of two. But, participants clearly liked list 3 recommendations over genre based recommendations (p value 0.006).

Table 5.14: Wilcoxon's signed-rank test to compare genre ratings and ratings for list 2 and list 3(happy phase)

		W	p	Rank-Biserial Correlation
genre_enjoyment	list2_enjoyment	46.000	0.696	-0.124
genre_enjoyment	list3_enjoyment	4.000	0.006	-0.897

	N	Mean	SD	SE
genre_enjoyment	15	2.667	0.617	0.159
list2_enjoyment	15	2.867	1.187	0.307
list3_enjoyment	15	3.800	0.862	0.223

Table 5.15: Descriptive Statistics of enjoyment ratings for list 2 and list 3(happy phase)

Comparing Genre and Happy ratings

We further compared ratings given by the participants for genre-based music recommendations and recommendations of list 4(low valence & danceability) and list 5 (low

valence & high danceability). From Table 5.16, we can see that p-value for genre and list 4 ratings is significant (0.006). The descriptive statistics show a significant difference between their mean ratings (Table 5.17). We can also see that there is an insignificant difference in ratings for genre-based and recommendation list 5 with mean ratings for recommendation list 5 less than that of a genre-based recommender system.

		W	p	Rank-Biserial Correlation
genre_enjoyment	list4_enjoyment	4.500	0.006	-0.885
genre_enjoyment	list5_enjoyment	29.000	0.915	0.055

Table 5.16: Wilcoxon's signed-rank test to compare genre ratings and ratings for list 4 and list 5(sad phase)

	N	Mean	SD	SE
genre_enjoyment	15	2.667	0.617	0.159
list4_enjoyment	15	3.600	0.828	0.214
list5_enjoyment	15	2.600	0.910	0.235

Table 5.17: Descriptive statistics to compare genre ratings and ratings for list 4 and list 5(sad phase)

The results indicate that our sad songs were preferred more than genre-based recommendations by the participants. Moreover, the same participants preferred music with low danceability and high valence over genre-based recommendations.

5.9 Conclusion and Discussion

In Chapter 4, we discussed our system Moodify in detail. We discussed different phases of our system. In this chapter, we have discussed our dependent and independent variables and discussed the statistical tests to test our Hypothesis. The results from the tests are noted in the previous section.

In this section, we discuss our results and limitations of our evaluation. Our initial results from our evaluation suggest that we were successful in deriving a relationship between the mood of the user and their preference for latent musical features. Before diving into details, below we list our results from the Hypothesis tests conducted:

1. Our results suggest that a Happy user would prefer music with *Low Danceability* and *High Valence* values over music with *High Danceability* and *High Valence* values. Our Hypothesis **H1** was rejected in this case.
2. Our results also suggest that a Sad user would prefer music with *Low Danceability* and *Low Valence* values over music with *High Danceability* and *Low Valence* values. The Hypothesis **H2** was accepted in this case.
3. We also looked into user mood and the recommendations the user would find surprising. Our results suggest that Happy Users find music with *Low Danceability* and *High Valence* more surprising than music with *High Danceability* and *High Valence* but our result was not significant, rejecting our hypothesis **H3**.
4. We looked the same for sad users and found that they found music with *High Danceability* and *Low Valence* more surprising than music with *Low Danceability* and *Low Valence* which was in line with our hypothesis and the results were significant. Hence, our hypothesis **H4** was accepted.
5. Additionally, we also compared the user satisfaction for our genre-based and mood specific recommender system. Our insights were that the participants preferred sad recommendations(in a sad mood) over genre-based recommendations(in a neutral state). The same participants preferred recommendations with high valence & low danceability(in a happy state) over genre-based recommendations(in a neutral state)

We were able to answer our Hypothesis questions and find some insights about user mood and latent musical features. Though we have answers to our questions, we would like to bring into light some limitations of our study which might have affected user-centric evaluations. These **limitations** are listed below:

1. One limitation of our study is the effectiveness of the **Mood Induction** process. The Mood Induction process isn't 100% successful on all the participants. Some participants were not **Happy** or **Sad** after the Mood Induction process. Our System displays mood specific recommendations to the participant despite them not being in that specific mood. Out of n=15 participants, 7 felt happy after seeing the happy clip and 12 felt sad after seeing the sad clip. It could be hence said that the happy induction was not as successful as the sad induction.

2. For our study, we chose to use a within-subjects experimental design. This study has both its benefits and limitations. Many might argue that by using a between-subjects design, we might be able to capture user preferences better. Our motivation to use a within-subjects design was to have a good sample size and reduce the bias induced by personality and other user characteristics but we agree that it might have created fatigue in some participants. With a huge number of participants, we could perform the same experiment using a within-subjects design along with some external measures to reduce bias created by personality and other user characteristics.

We also drew some additional insights from our study which we would like to discuss.

- We realised there might be some issues with our songs which led to people answering very low to the recommendation list. Example as seen in Figure 5.5, a lot of participants rated the recommendation list low pulling down the mean user satisfaction. This could be because a few songs recommended to them were disliked them by so much that they provided a very low rating to the entire recommendation list.
- Our dataset had songs which were song excerpts and not very popular. This might have affected user satisfaction. Some participants might have not preferred listening to songs they have never heard of before.
- Additionally, Participants liked songs which had Low Danceability. This might be because the pleasant songs in our dataset had Low Danceability
- External factors like the sound quality of the song, state of the user (stressed, relaxed) while using the system might have affected their ratings.

Post Hoc Analysis

As you can recall, we provided users with a free text box where they could provide some additional comments about the recommendation list. Kindly refer to Figures 4.33 for this.

We analysed user comments to get some extra insights on user satisfaction and unexpectedness.

First, the comments were extracted via manual parsing and then we looked at the comments which were most consistent with the recommendation lists. These were some insights for users in the Happy and Sad phase.

Happy Phase	Sad Phase
Boring Recommendations	Enjoyable recommendations
Users misinterpreted country or folk songs	Users found songs in sad phase enjoyable because mostly they turned out to be classical

Table 5.18: Insights for Happy and Sad Phase recommendations

We also dig deeper for some insights for each recommendation lists. We looked at user comments for each recommendation lists and below we list the most consistent comments made by users for our Mood based recommendations.

Recommendation List 2	Recommendation List 3	Recommendation List 4	Recommendation List 5
Pleasant and unique than Baseline Enjoyed rock and pop songs Enjoyable	More enjoyable than List 2 Classical songs - nice Soothing	Not Surprising Enjoyable Goes with the movie clip shown	More unique than List 4 Diverse and Weird Liked Classical & Hip hop recommendations

Table 5.19: Insights for Happy and Sad Phase recommendations

From our Insights and Post hoc analysis, we can see that in general *Happy users* found recommendation list 3 more enjoyable than recommendation list 2. Additionally, *Sad Users* were found to enjoy recommendation list 4 more than recommendation list 5.

This analysis made us realise that all users in both the mood found music with **Low Danceability** more enjoyable than music with **High Danceability**.

This left us curious and wonder if there is an issue with our Dataset itself. Users, in general, did not enjoy music with **High Danceability**. This made us wonder if our dataset has rather bizarre music with **High Danceability** feature. Additionally, we tested for **High Valence** for **Happy users** and **Low Valence** for **Sad users**.

Hence, we wanted to test users need for Valence in different moods. Considering the limitation of the dataset, we realised our participants only liked music with **Low Danceability**. We conducted another User Experiment with Danceability constant to Low and varying Valence to High and Low for user moods Happy and Sad. This has been discussed in detail in the next chapter.

Before we proceed to the next chapter, we would like to discuss some possible evaluations that we could have performed but were not able due to time constraints and our experimental design.

Limitations and Future Work(Evaluations)

In this chapter, we discussed the necessary statistical tests performed to answer our research question. The procedure explained in the chapter is just one of the many ways to test a hypothesis. Here, we would like to discuss other possible approaches.

- We try to compare the user satisfaction and unexpectedness of the list in both happy and sad users. We could have chosen a different set of participants for each task, but with limited sample size($n=15$) we couldn't do the same. We could also have shown the participants one recommendation list for both happy/sad phase but population size was an issue again. Though these approaches have their benefits, we believe we have avoided the inconsistency that might have generated due to age, gender and personality by choosing the same participants for both the task.
- Many people might argue that we could have asked the participants the recommendation list preferred by them. Eg We could have asked them *In your current emotional state, Which recommendation list do you find most enjoyable (List 2 or List 3/ List 4 or List 5)?*. Instead, we just ask them to rate the recommendation list and compared the ratings on our own. This design choice was made to avoid any cognitive overload and difficulty in decision making the participants would have gone through if asked to compare the tasks.
- The participants always went through a happy phase first and then through a sad phase which could be an issue with the experimental design. If users randomly went through a happy/sad phase, we could have avoided the bias that might have been generated in some participants.
- It is also a point of concern that we ignore the self-reported emotion and instead show the participants sad/happy specific recommendations. This was again done

due to limited sample size and our evaluations are done for a perfect emotion induction process.

These are some of the limitations of our evaluation and experimental design. If asked to perform the study again, we would want to rectify these mistakes. Due to time constraints, we were unable to perform a perfect experiment. Our suggestion to researchers would be to keep these points in mind and have some time to recruit a decent number of participants ($n > 30$) to get valid results.

Chapter 6

Evaluation 2 - To study the relationship between user mood and valence

6.1 Introduction

As discussed in the previous chapter, we found some patterns from user responses which suggests that users in both moods liked music with *Low Danceability* values. Hence, we wanted to check the user's need for valence.

We conducted another User Experiment after our main Experiment which was discussed in the previous Chapter 5. We will discuss the results of our second experiment in this chapter. We try to answer the below Research question which was formulated after obtaining results of Experiment 1.

RQ3: *How does mood affect user's need for Valence in Music Recommender Systems?*

6.2 Design

In this section, we discuss our design for the second experiment conducted. For this task, We wanted to check if all music recommendations had danceability constant as **Low**, then which kind of recommendations would be liked by Sad and Happy users. We used **within subjects** design for our experiment. In this case, all participants are shown all recommendation lists. The participants are asked a few questions for the recommendation list. Here, we wanted to measure user satisfaction hence we asked them *How much did you enjoy the recommendations?*.

6.3 Independent Variable

In this section, We discuss the Independent variables for our second experiment. As discussed, We are measuring **User Satisfaction** when recommendations are presented to a *Happy User* and *Sad User*. In this task, our focus is to find a relationship between latent musical feature *Valence* and user mood *Happiness* and *Sadness*. Thus, Valence is the Independent feature in our experiment.

6.4 Dependent Variable

In this study, we try to measure user satisfaction when the users are provided with recommendation lists. We can recall the definition of User Satisfaction from Section 5.4. **User Satisfaction** is our dependent variable

6.5 Hypothesis

To answer our research question, We formulated two Hypothesis questions. These questions have been listed below:

- **Happy** users would prefer music with **High Valence**
- **Sad** users would prefer music with **Low Valence**

6.6 Approach

The procedure is the same as our main experiment. Each participant follows the same steps as before. These steps are restated below:

1. The first step was to take consent from the participants for their free will to participate in the experiment.
2. In the second step, we collect some basic demographic information about the participant by asking a few questions. These questions can be seen in Table 6.1

Index	Question
1	Gender
2	Age Group
3	Which Nationality do you identify yourself most with?
4	How many hours do you use a Music Recommender System in a week?

Table 6.1: Demographic questions

3. The third step was to form a user profile. This was achieved by showing the participants with an initial set of items and asking them to rate the songs on a scale from 1 to 5.
4. The fourth step involves a user's interaction with the genre-based recommender system.
5. The fifth step involves user seeing a video clip meant for emotion induction.
6. The sixth step involves user's interaction with the Mood specific recommender system which recommends songs based on latent features **Valence** with a constant range of values **Danceability** as low

In this system, each participant goes through the Happy and Sad phase. The properties of the recommendation lists are displayed below in Table 6.2

	Recommendation List	Recommendation List
Happy User	List 2- Low Danceability & Low Valence	List 3- Low Danceability & High Valence
Sad User	List 4- Low Danceability & Low Valence	List 5- Low Danceability & High Valence

Table 6.2: Characteristics of Recommendation Lists for Experiment 2

6.7 Participants

To conduct this experiment, we recruited new participants who were not familiar to our system. All the participants are students at the *Delft University of Technology*. For this task, we had an equal number of male($n=7$) and female($n=7$) participants as seen in Figure 6.2. All the participants belonged to the age group 18-34. Out of which 50% of the participants belonged to the age group 18-24 and 50% of the participants belonged to age group 25-34. Additionally, our participants belonged to different nationalities. We had 10 different nationalities in this study.

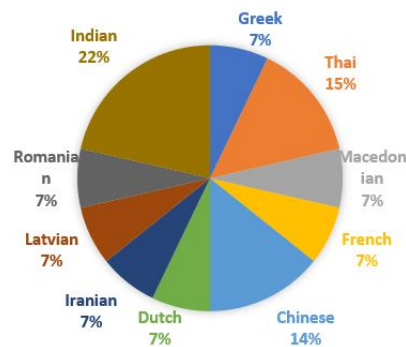


Figure 6.1: Participant Demographics by Nationality

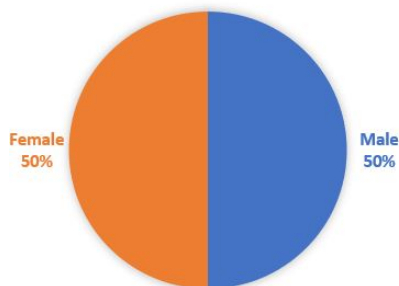


Figure 6.2: Participant Demographics by Gender



Figure 6.3: Participant Demographics by Age

We also looked into user's consumption behaviour of music recommender system. The results are displayed in Figure 6.4

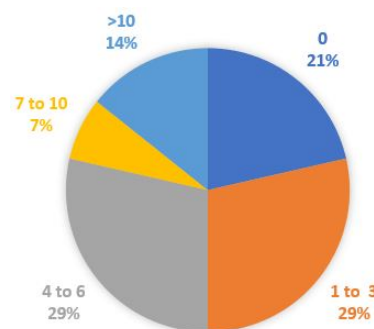


Figure 6.4: Participant Demographics by number of hours they consumed a Music Recommender System in a week

6.8 Statistical Tests and Results

In this section, we discuss the results of our user-centric evaluation. Our evaluation is aimed at understanding the relationship between need for latent feature *Valence* and user mood.

6.8.1 User Satisfaction for Happy Users

H1: Happy users would prefer recommendation list 2 (Low Danceability and High Valence) over recommendation list 3 (Low Danceability and Low Valence)

To test our Hypothesis, we collect user answers and perform statistical test on the data. Figures 6.5 and 6.6 show the user enjoyment ratings of users for recommendation lists 2 and 3.

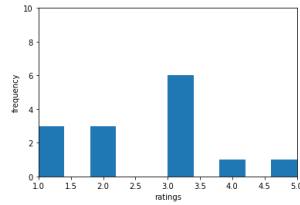


Figure 6.5: Enjoyment Ratings of Users for List 2 recommendations

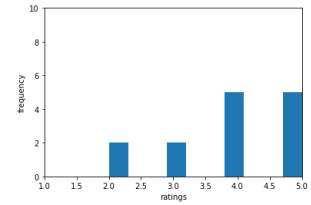


Figure 6.6: Enjoyment Ratings of Users for List 3 recommendations

We test for Normality of this data and find them Non-Normal via Visual and Statistical Tests. The visual tests can be formed in Figures 6.7 and 6.8

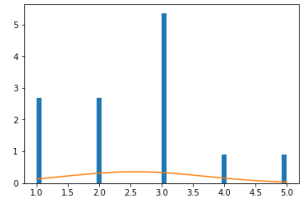


Figure 6.7: Gaussian Plot for List 2 enjoyment ratings

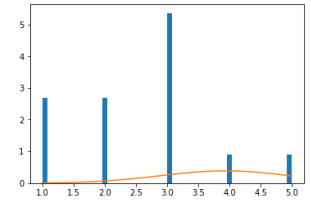


Figure 6.8: Gaussian plot for List 3 enjoyment ratings

After this, we perform a *Wilcoxon's signed-rank test* which answers if the data is from the same distribution.

The test results indicate that the recommendation List 2 and List 3 follow a different distribution as seen in Table 6.3

Table 6.3: Wilcoxon's signed-rank test for List 2 and List 3 enjoyment ratings

		W	p	Rank-Biserial Correlation
list2	list3	0.000	0.002	-1.000

The above results indicate a strong p value indicating that emotion of the user (here happiness) has an impact on their preference towards music with high and low valence. The effect value is very high indicating that the mood has a strong impact on the ratings given by the participants for list 2 and list 3. We further look into the descriptive statistics of the ratings as shown in Table 6.4

	N	Mean	SD	SE
List2	14	2.571	1.158	0.309
List3	14	3.929	1.072	0.286

Table 6.4: Descriptive statistics of List 2 and List 3 enjoyment ratings

The Mean ratings clearly suggest that the Happy Users enjoyed recommendation List 3 more than List 2. This indicates that our Hypothesis **H5** is accepted. Users liked music with **High Valence** even when we kept danceability constant.

6.8.2 User Satisfaction for Sad Users

H6: Sad users would prefer recommendation list 3 (Low Danceability and Low Valence) over recommendation list 4 (Low Danceability and High Valence)

To test our Hypothesis, we collect user answers and perform statistical test on the data. Figures 6.9 and 6.10 show the user enjoyment ratings of users for recommendation lists 2 and 3.

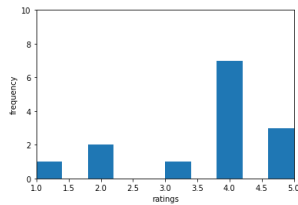


Figure 6.9: Enjoyment Ratings of Users for List 4 recommendations

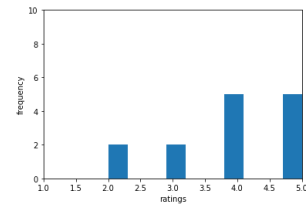


Figure 6.10: Enjoyment Ratings of Users for List 5 recommendations

We test for Normality of this data and find them Non-Normal via Visual and Statistical Tests. The visual tests can be formed in Figures 6.11 and 6.12

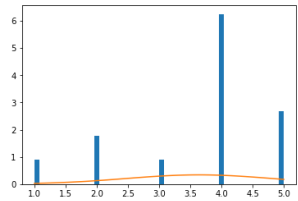


Figure 6.11: Gaussian Plot for List 4 enjoyment ratings

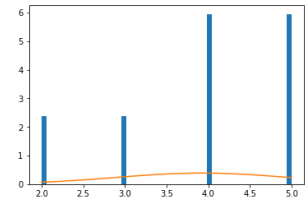


Figure 6.12: Gaussian plot for List 5 enjoyment ratings

After this, we perform a *Wilcoxon's signed-rank* test which answers if the data is from the same distribution.

The test results as seen in Table show a p value > 0.05 indicating that the lists follow the same distribution. The effect size is very low (0.167) indicating that emotion of the user doesn't have an impact on their ratings.

Table 6.5: Wilcoxon's signed-rank test for List4 and List5 enjoyment ratings

		W	p	Rank-Biserial Correlation
List4	List5	32.500	0.626	-0.167

We still look at the mean enjoyment ratings of the recommendation list and we can see that there is little difference between the mean of the ratings in list 4 and list 5.

Table 6.6: Descriptive statistics of List 4 and List 5 enjoyment ratings

	N	Mean	SD	SE
List4	14	3.643	1.216	0.325
List5	14	3.929	1.072	0.286

From the above results, we reject Hypothesis **H6**. The Wilcoxon's test indicates that the emotion (here sadness) doesn't have an impact on the ratings hence we can't accept the alternative hypothesis. The descriptive statistics do indicate their preference for music with High Valence but the mean difference is very low making it insignificant.

6.9 Conclusion and Discussion

After conducting the statistical tests, we found some interesting insights. The results from the statistical tests have been listed in the previous section. Our initial results from our evaluation suggest that we were successful in analysing a relationship between **Valence** need and **Mood** of the user. We also discuss the limitations of our study and some additional insights which were drawn from the user comments for the recommendation lists.

We can recall our Hypothesis questions **H5** and **H6** from Section 6.5. Our results indicate that:

- A Happy user would prefer music with **High Valence** over music with **Low Valence** when we keep danceability as **Low**. This goes the same as our Hypothesis **H5**
- A Sad user is also seen to prefer music with **High Valence** over music with **Low Valence**. The results are insignificant and the impact of emotion on their music preferences does not hold in this case.

We also conducted some posthoc analysis on the user comments for our recommendation lists. We explain this in detail below.

Post Hoc Analysis

We provided users with free text where they could provide their comments about the recommendation lists. This is similar to our first experiment. Kindly refer to Appendix for the snapshots of our interface for experiment 2. For this experiment, we kept danceability as **low** and changed **valence** to both **high** and **low** for both **happy** and **sad** users. It must be noted to avoid repetition of the songs for the same participant we chose Top N relevant songs and then presented the songs in a manner presented in Table.

	Recommendation List	Recommendation List
Happy User	List 2 - 5th to 10th relevant songs	List 3 - Top 5 relevant songs
Sad User	List 4 - Top 5 relevant songs	List 5 - 5th to 10th relevant songs

Table 6.7: Relevance of Recommendation Lists to User Profile

To get some deeper insights about the recommendation lists, we dig deeper into the user comments and present the most consistent comments in Table 6.8.

Recommendation List 2	Recommendation List 3	Recommendation List 4	Recommendation List 5
Not Enjoyable	More enjoyable than List 2	Decent recommendations	Better recommendations
Expected Cheerful songs after the video	Classical songs are enjoyable	Lot of classical recommendations	Enjoyable
Not Surprising	Amazing list of songs	-	-

Table 6.8: Insights for Happy and Sad Phase recommendations

From Post Hoc Analysis, we can notice that both users liked recommendations with **High Valence** but this was not a strong indication in case of Sad users. Additionally, we see that participants enjoyed **Classical** recommendations that were given to them. This was also the case in Experiment 1. This indicates that our dataset has really good set of classical songs.

The results of this experiment successfully solve the limitations of first Experiment in determining the relationship between mood of the user and their need for latent feature **valence** in music.

Chapter 7

Discussion

In this chapter, we discuss in detail the limitations of our study and possible ways to improve it.

To address the analyse the relationship between user emotion and their preference towards latent musical features, we conducted an in-depth study to understand the kind of music people like in a certain emotional state. We studied different techniques used for emotion induction and detection. Additionally, we closely looked at the musical features used in the music emotion recognition domain. We based our study on the research direction that analyses user satisfaction and surprisingness in a certain emotional state for the given musical feature values *danceability* & *valence*. In Chapter 1, we discussed the problems in the field:

- *User emotion and music preferences*: Limited research has been done to analyse the music preferences of people in a certain emotional state.
- *Musical features and emotion*: Most of the studies deal with modulating tempo and mode to express emotion in a song. Limited research has been done to analyse the same for other musical features.

We define our main research objective based on the problems defined above.

- *Analyse the relationship between latent musical features and user satisfaction for a given emotion*: We study the influence of high/low danceability valence provided to users based on their emotional state happiness/sadness on user satisfaction.

Additionally, studies show that the metric *serendipity* (unexpectedness) is influenced by the emotional state of the user. This brings us to our second research objective:

- *Analyse the relationship between latent musical features and unexpectedness for a given emotion* : We study the influence of high/low danceability valence provided to users based on their emotional state happiness/sadness on unexpectedness ratings.

7.1 Analyse the relationship between latent musical features and user satisfaction for a given emotion

Our first objective was to study the preferences for latent musical features for emotion happiness and sadness. To achieve this, we built an interactive interface that made it possible for the participants to navigate through our system easily. Revisiting our hypothesis, we propose that a happy person would enjoy listening to happy-sounding music and a sad person would prefer listening to sad-sounding music. To achieve this, we built a system which would first provide the participants with a list of songs and ask them to rate the songs to learn user preferences. Later, they are shown emotion-inducing movie clips for emotions happy and sad. After this step, they are shown two recommendations lists for each emotion. We conducted two experiments to learn user's preferences for features *danceability* and *valence*. One can recall the characteristics of these recommendation lists from Table 4.4 and 6.2. We wanted to compare for the following by conducting our experiments:

- We wanted to compare if a happy person would prefer happy music that has features *high danceability & high valence* over sad sounding music *low danceability & high valence*
- We wanted to compare if a sad person would prefer sad music that has features *low danceability & low valence* over happy-sounding music *high danceability & low valence*

After performing experiment 1, we were able to compare the above statements. From the obtained results, we concluded that a happy person liked sad sounding music with features *low danceability & high valence* which was contradictory to our assumption that they would prefer music with features *high danceability & high valence*. For a sad person, our assumption was true. They were seen to like sad music with features *low danceability & low valence* over music with *high danceability & low valence*. This was in line with our hypothesis. Additionally, we conducted a posthoc analysis which has been explained in detail in Section 5.9 and noticed that most of the participants preferred listening to music with low danceability. We thought it could be due to our dataset, which might have pleasant songs with low danceability values. Hence, we decided to conduct a second experiment where danceability was kept constant as low and valence values were changed to high/low. This has been shown in Table 6.2. The results from this experiment showed that happy people indeed liked listening to music with feature *high valence* but our results for sad people were not significant in this case.

From the experiments conducted we can say that a happy person preferred listening to music with feature values high valence over low valence. Sad people showed a clear preference towards music with *low danceability & low valence* over music with *high danceability & low valence* but failed to distinguish between music where *low danceability* was common in songs but valence was modulated from high to low.

7.2 Analyse the relationship between latent musical features and unexpectedness for a given emotion

For our second research objective, we used the same interface. One can recall from Chapter 4 that with each recommendation list the participants were asked two questions, one asking the user about their enjoyment ratings and the other asking how surprising were the recommendations for them. By performing the experiments, we wanted to compare the following:

- We wanted to compare if a happy person would find sad music that has features *low danceability & high valence* surprising over happy-sounding music *high danceability & high valence*
- We wanted to compare if a sad person would find happy music that has features *high danceability & low valence* surprising over sad sounding music *low danceability & low valence*

Our results indicate that a happy person doesn't find sad music surprising over happy music, but a sad person clearly finds happy music surprising over sad music.

7.3 Limitations

In this section, we discuss the limitations of our study. We identify the following main limitations of our approach:

- **Dataset:** Our dataset was small(1802 data points) to train a recommender system model. Moreover, the audio clips were short song excerpts and were not the entire songs.
- **Algorithm:** Due to no user ratings for the given songs, we decided to use a content-based algorithm for our study. Literature shows that collaborative filtering approaches are more sophisticated to find relevant items.
- **Sample size:** In our study, we had a limited number of participants (n=15). Though the sample size is decent for our study, we would get a strong indication of results with a larger sample size.
- **Limited features:** Due to time constraints, we based our study only on features danceability and valence. This could be extended to other features in the music recommender system domain.
- **Limited features for similarity:** Currently, the mood-based recommender systems learns user's preferences for features tempo and loudness and provides recommendations from subsets that match their preferences for these features. To extend this study, we could include artist and track popularity, and other musical features.
- **Gender, age and depression:** In this study, we ignore the effects of gender, age and personality towards their music preference. It has been seen that these

factors play a key role in one's preferences and hence it would be interesting to study them.

- **Experimental design:** For this study, the within-subjects design was used which might have created bias and fatigue in the participants towards the end of the study. With a larger sample size, we could have used a between-subjects design for the experiment.
- **Emotion induction and detection:** Due to a limited number of participants, we show mood specific recommendations to people who are not that in a particular emotional state. This is a huge drawback and could be solved with larger sample size. By this approach, we would consider only participants for whom the emotion induction process was successful.

Chapter 8

Conclusion

In this study, we research about the relationship between emotion (happiness & sadness) and musical features (danceability & valence). We try to understand a user's preference for these features in a certain emotional state. Additionally, we try to find the music they would find surprising in a certain emotional state. To answer our research questions, we first conducted a literature survey of emotion-based recommender systems and features used for emotion recognition in music. This provided us with an insight into different emotion recommender systems, different emotion induction and detection approaches and music preferences of people when they are induced with a certain emotion.

Previous research in the field of recommender systems provides music recommendations based on features genre, artist and popularity of the track/artist. More recent research has been done using other features to recommend items. Features like tempo and mode have been studied in music psychology to express emotions like happiness and sadness in music but the use of other musical features to express emotion in music has not been studied in the domain of recommender systems. Also, emotion induction techniques are still new in the recommender system domain.

To answer our research questions, we built an interactive recommender system-MooDify that combines emotion with a traditional content-based recommendation algorithm. This allowed us to incorporate an emotion induction technique and provide them with recommendations specific to the emotional phase. We conducted online evaluations for these recommendations.

We conclude that our system was able to draw some important insights. It analysed the preference of feature values for features danceability and valence by people for emotion happiness and sadness. Our study indicated some interesting results and has opened doors for research in this domain.

Below we present some additional insights derived from our study. We believe these factors might have an impact on our study and would like to discuss them below:

- We believe our dataset had musical pieces which were unknown to the participants. We believe this led to unbiased results as they wouldn't just like popular songs. Hence, It is a strong point of our thesis

- Additionally, we believe the dataset had a lot of classical songs which were low in the feature danceability. Moreover, these songs were pleasant to hear to in both user moods. We also believe our dataset did not have many good songs with high danceability. This might have an impact on our results.
- Mood induction is not 100% effective and hence has a great impact on our results.
- We also saw that the participants enjoyed recommendations generated by our mood-based system (music with high valence & low danceability and low valence & low danceability) more than genre-based recommendations.

8.1 Future work

Our experiments show that we successfully answered our research questions. In this section, we discuss the possible research extension of our work.

- **Participants:** We believe with more number of participants, we could get more significant results.
- **Mood Induction and Detection:** A limitation of our research is we present the user mood specific recommendations irrespective of their mood after the mood induction process. We believe this might greatly affect our results. As a future work with a large sample size, this issue could be resolved by providing a user with recommendations for their **actual current mood**
- **Collaborative Filtering:** Currently, our system uses a content-based approach for generating recommendations. This decision was made due to lack of user ratings for our data. We believe if we could get user ratings for our dataset, We would be able to build a more sophisticated recommender system.
- **Comparing different algorithms:** Currently, our system uses a clustering approach to create subsets of data based on *high/low Valence and Danceability*. We could use a scoring re-ranking function to do the same. It would be interesting to see the results from different approaches.
- **Gender, Age and Depression:** Characteristics of a person such as their gender, age and if they suffer from depression is seen to have an impact on their preference towards sad sounding music [119]. As part of future work, we would like to analyse the relationship between these factors, the person's emotion and their preference towards latent musical features.
- **More Features:** Currently, our system provides recommendations based on features *Danceability* and *Valence*. It would be interesting to see users' needs for other features for a mood.

Bibliography

- [1] Yi-Hsuan Yang and Homer H Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):40, 2012.
- [2] Example of content based system. <https://johnlamendy.wordpress.com/2015/10/14/collaborative-filtering-in-apache-spark/>. Accessed: 2019-06-25.
- [3] K-means clustering process. <https://www.brandidea.com/kmeans.html>. Accessed: 2019-06-25.
- [4] Markus Schedl, Andreu Vall, and Katayoun Farrahi. User geospatial context for music recommendation in microblogs. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 987–990. ACM, 2014.
- [5] Yi-Hsuan Yang and Jen-Yu Liu. Quantitative study of music listening behavior in a social and affective context. *IEEE Transactions on Multimedia*, 15(6):1304–1315, 2013.
- [6] Gustavo Gonzalez, Josep Lluís de la Rosa, Miquel Montaner, and Sonia Delfin. Embedding emotional context in recommender systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop, ICDEW '07*, pages 845–852, Washington, DC, USA, 2007. IEEE Computer Society.
- [7] Yong Zheng, Robin Burke, and Bamshad Mobasher. The role of emotions in context-aware recommendation.
- [8] Chao Xue, Tian Li, Shufei Yin, Xinyi Zhu, and Yuxin Tan. The influence of induced mood on music preference. *Cognitive processing*, 19(4):517–525, 2018.
- [9] E Glenn Schellenberg, Ania M Krysciak, and R Jane Campbell. Perceiving emotion in melody: Interactive effects of pitch and rhythm. *Music Perception: An Interdisciplinary Journal*, 18(2):155–171, 2000.

- [10] Gregory D Webster and Catherine G Weir. Emotional responses to music: Interactive effects of mode, texture, and tempo. *Motivation and Emotion*, 29(1):19–39, 2005.
- [11] Irène Deliège, John A Sloboda, et al. *Perception and cognition of music*. Psychology Press, 2004.
- [12] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *Recommender Systems Handbook*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 2010.
- [13] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, January 2003.
- [14] Carlos A. Gomez-Urbe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4):13:1–13:19, December 2015.
- [15] Algorithmic music recommendations at spotify. https://www.slideshare.net/MrChrisJohnson/algorithmic-music-recommendations-at-spotify/6-Challenge_20_Million_songs_how. Accessed: 2018-11-28.
- [16] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.
- [17] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Know.-Based Syst.*, 46:109–132, July 2013.
- [18] Christine Bauer and Alexander Novotny. A consolidated view of context for intelligent systems. *Journal of Ambient Intelligence and Smart Environments*, 9(4):377–393, 2017.
- [19] Anind K Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.
- [20] Xinxi Wang, David Rosenblum, and Ye Wang. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 99–108. ACM, 2012.
- [21] Shuiguang Deng, Dongjing Wang, Xitong Li, and Guandong Xu. Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications*, 42(23):9284–9293, 2015.
- [22] Kyoungro Yoon, Jonghyung Lee, and Min-Uk Kim. Music recommendation system using emotion triggering low-level features. *IEEE Transactions on Consumer Electronics*, 58(2):612–618, 2012.
- [23] Byeong-Jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460, 2010.

- [24] Adrian North and David Hargreaves. *The social and applied psychology of music*. OUP Oxford, 2008.
- [25] Han-Saem Park, Ji-Oh Yoo, and Sung-Bae Cho. A context-aware music recommendation system using fuzzy bayesian networks with utility theory. In *International conference on Fuzzy systems and knowledge discovery*, pages 970–979. Springer, 2006.
- [26] Zhiyong Cheng and Jialie Shen. On effective location-aware music recommendation. *ACM Transactions on Information Systems (TOIS)*, 34(2):13, 2016.
- [27] Ricardo Dias and Manuel J Fonseca. Improving music recommendation in session-based collaborative filtering by using temporal context. In *2013 IEEE 25th international conference on tools with artificial intelligence*, pages 783–788. IEEE, 2013.
- [28] Antonio R. Damasio. Penguin Group (USA).
- [29] D Goleman. *Emotional intelligence*. Bantam Books, 1995.
- [30] Joseph Sam. *AI*. Japan Inc, 2001.
- [31] Marco Polignano. The influence of user’s emotions in recommender systems for decision making processes. In *DC@CHIItaly*, 2015.
- [32] Marko Tkalcic. *Affective recommender systems : the role of emotions in recommender systems*. 2011.
- [33] Gustavo González, Beatriz López, and Josep Lluís De La Rosa. The emotional factor: An innovative approach to user modelling for recommender systems. In *In Proceedings of AH2002 Workshop on Recommendation and Personalization in e-Commerce*, pages 90–99, 2002.
- [34] Ai Thanh Ho, Ilusca LL Menezes, and Yousra Tagmouti. E-mrs: Emotion-based movie recommender system. In *Proceedings of IADIS e-Commerce Conference. USA: University of Washington Bothell*, pages 1–8, 2006.
- [35] Yu Chen and Pearl Pu. Cofeel: Using emotions to enhance social interaction in group recommender systems. In *Alpine Rendez-Vous (ARV) 2013 Workshop on Tools and Technology for Emotion-Awareness in Computer Mediated Collaboration and Learning*, 2013.
- [36] Ioannis Arapakis, Yashar Moshfeghi, Hideo Joho, Reede Ren, David Hannah, and Joemon M Jose. Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1440–1443. IEEE, 2009.
- [37] Ivana Andjelkovic, Denis Parra, and John O’Donovan. Moodplay: Interactive mood-based music discovery and recommendation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 275–279. ACM, 2016.

- [38] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 13–22. ACM, 2012.
- [39] Feng Lu and Nava Tintarev. A diversity adjusting strategy with personality for music recommendation. In *Recsys workshop on Interfaces and Decision Making in Recommender Systems*, 2018.
- [40] Yucheng Jin, Nava Tintarev, and Katrien Verbert. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 13–21. ACM, 2018.
- [41] Bruce Ferwerda, Andreu Vall, Marko Tkalcić, and Markus Schedl. Exploring music diversity needs across countries. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 287–288. ACM, 2016.
- [42] Markus Schedl and David Hauger. Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*, pages 947–950. ACM, 2015.
- [43] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Emilia Gómez, and Perfecto Herrera. Content-based music recommendation based on user preference examples. *Copyright Information*, page 33, 2010.
- [44] Renato Panda, Bruno Rocha, and Rui Pedro Paiva. Music emotion recognition with standard and melodic audio features. *Applied Artificial Intelligence*, 29(4):313–334, 2015.
- [45] Jens Madsen, Bjørn Sand Jensen, and Jan Larsen. Learning combinations of multiple feature representations for music emotion prediction. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 3–8. ACM, 2015.
- [46] Satoru Fukuyama and Masataka Goto. Music emotion recognition with adaptive aggregation of gaussian process regressors. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 71–75. IEEE, 2016.
- [47] Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. Emotion analysis of songs based on lyrical and audio features. *arXiv preprint arXiv:1506.05012*, 2015.
- [48] Humberto Corona and Michael P OâMahony. An exploration of mood classification in the million songs dataset. In *12th Sound and Music Computing Conference, Maynooth University, Ireland, 26 July-1 August 2015*. Music Technology Research Group, Department of Computer Science Maynooth âŠ, 2015.

- [49] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas. Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011(1):4, 2011.
- [50] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, volume 86, pages 937–952. Citeseer, 2010.
- [51] Tuomas Eerola, Olivier Lartillot, and Petri Toivainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Ismir*, pages 621–626, 2009.
- [52] Yu-Ching Lin, Yi-Hsuan Yang, Homer H Chen, I-Bin Liao, and Yeh-Chin Ho. Exploiting genre for music emotion classification. In *2009 IEEE International Conference on Multimedia and Expo*, pages 618–621. IEEE, 2009.
- [53] Lise Gagnon and Isabelle Peretz. Mode and tempo relative contributions to "happy-sad" judgements in equitone. *Cognition and Emotion*, 2003.
- [54] Patrick G Hunter, E Glenn Schellenberg, and Stephanie M Stalinski. Liking and identifying emotionally expressive music: Age and gender differences. *Journal of Experimental Child Psychology*, 110(1):80–93, 2011.
- [55] Ronald S Friedman, Elana Gordis, and Jens Förster. Re-exploring the influence of sad mood on music preference. *Media Psychology*, 15(3):249–266, 2012.
- [56] Christa L Taylor and Ronald S Friedman. Sad mood and music choice: Does the self-relevance of the mood-eliciting stimulus moderate song preference? *Media Psychology*, 18(1):24–50, 2015.
- [57] E Glenn Schellenberg, Kathleen A Corrigan, Olivia Ladinig, and David Huron. Changing the tune: listeners like music that expresses a contrasting emotion. *Frontiers in psychology*, 3:574, 2012.
- [58] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.
- [59] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.
- [60] Swearingen Sinha, Kirsten Swearingen Rashmi, and Rashmi Sinha. Beyond algorithms: An hci perspective on recommender systems, 2001.
- [61] Keith Bradley and Barry Smyth. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, pages 85–94. Citeseer, 2001.

- [62] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.
- [63] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 109–116, New York, NY, USA, 2011. ACM.
- [64] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA, 2014. ACM.
- [65] Marco Tulio Ribeiro, Anisio Lacerda, Adriano Veloso, and Nivio Ziviani. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 19–26. ACM, 2012.
- [66] John Paul Kelly and Derek Bridge. Enhancing the diversity of conversational collaborative recommendations: A comparison. *Artif. Intell. Rev.*, 25(1-2):79–95, April 2006.
- [67] Yan Yang and Jian Zhong Li. Interest-based recommendation in digital library. *Journal of Computer Science*, 1(1):40–46, 2005.
- [68] Makoto Nakatsuji, Yasuhiro Fujiwara, Akimichi Tanaka, Toshio Uchiyama, Ko Fujimura, and Toru Ishida. Classical music for rock fans?: Novel recommendations for expanding user interests. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 949–958, New York, NY, USA, 2010. ACM.
- [69] Liang Zhang. The definition of novelty in recommendation system. *Journal of Engineering Science & Technology Review*, 6(3), 2013.
- [70] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. Challenges of serendipity in recommender systems. In *WEBIST 2016: Proceedings of the 12th International conference on web information systems and technologies. Volume 2, ISBN 978-989-758-186-1*. SCITEPRESS, 2016.
- [71] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. A survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111:180–192, 2016.
- [72] Andrii Maksai, Florent Garcin, and Boi Faltings. Predicting online performance of news recommender systems through richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 179–186. ACM, 2015.

- [73] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101. ACM, 2006.
- [74] Suboojitha Sridharan. Introducing serendipity in recommender systems through collaborative methods. 2014.
- [75] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10*, pages 257–260, New York, NY, USA, 2010. ACM.
- [76] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 230–237. ACM, 1999.
- [77] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, 2004.
- [78] Yoshinori Hijikata, Takuya Shimizu, and Shogo Nishida. Discovery-oriented collaborative filtering for improving user satisfaction. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 67–76. ACM, 2009.
- [79] Iman Avazpour, Teerat Pitakrat, Lars Grunske, and John Grundy. Dimensions and metrics for evaluating recommendation systems. In *Recommendation systems in software engineering*, pages 245–273. Springer, 2014.
- [80] Allen Foster and Nigel Ford. Serendipity and information seeking: an empirical study. *Journal of documentation*, 59(3):321–340, 2003.
- [81] Keith Oatley and Jennifer M Jenkins. *Understanding emotions*. Blackwell publishing, 1996.
- [82] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.
- [83] Alf Gabrielsson and Erik Lindström. The influence of musical structure on emotional expression. 2001.
- [84] Tuomas Eerola. Are the emotions expressed in music genre-specific? an audio-based evaluation of datasets spanning classical, film, pop and mixed genres. *Journal of New Music Research*, 40(4):349–366, 2011.
- [85] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

- [86] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [87] James A Russell, Anna Weiss, and Gerald A Mendelsohn. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology*, 57(3):493, 1989.
- [88] G Sutherland, B Newman, and S Rachman. Experimental investigations of the relations between mood and intrusive unwanted cognitions. *British Journal of Medical Psychology*, 55(2):127–138, 1982.
- [89] Peter J Lang. A bio-informational theory of emotional imagery. *Psychophysiology*, 16(6):495–512, 1979.
- [90] Gordon H Bower. Affect and cognition. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 302(1110):387–402, 1983.
- [91] Maryanne Martin. On the induction of mood. *Clinical Psychology Review*, 10(6):669–697, 1990.
- [92] Dirk Hagemann, Ewald Naumann, Stefanie Maier, Gabriele Becker, Alexander Lürken, and Dieter Bartussek. The assessment of affective reactivity using films: Validity, reliability and sex differences. *Personality and Individual differences*, 26(4):627–639, 1999.
- [93] Johannes Hewig, Dirk Hagemann, Jan Seifert, Mario Gollwitzer, Ewald Naumann, and Dieter Bartussek. Brief report. *Cognition & Emotion*, 19(7):1095–1109, 2005.
- [94] James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995.
- [95] Pierre Philippot. Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition and emotion*, 7(2):171–193, 1993.
- [96] Luz Fernández-Aguilar, José Miguel Latorre, Laura Ros, Juan Pedro Serrano, Jorge Ricarte, Arturo Martínez-Rodrigo, Roberto Zangróniz, José Manuel Pastor, María T López, and Antonio Fernández-Caballero. Emotional induction through films: A model for the regulation of emotions. In *International Conference on Innovation in Medicine and Healthcare*, pages 15–23. Springer, 2016.
- [97] Leo Pauly and Deepa Sankar. A novel online product recommendation system based on face recognition and emotion detection. In *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2015 International Conference on*, pages 329–334. IEEE, 2015.
- [98] Toon De Pessemier, Damien Verlee, and Luc Martens. Enhancing recommender systems for tv by face recognition. In *12th International Conference on Web Information Systems and Technologies (WEBIST 2016)*, volume 2, pages 243–250, 2016.

- [99] Guibing Guo and Mohamed Elgendi. A new recommender system for 3d e-commerce: an eeg based approach. *Journal of Advanced Management Science*, 1(1):61–65, 2013.
- [100] Deger Ayata, Yusuf Yaslan, and Mustafa E Kamasak. Emotion based music recommendation system using wearable physiological sensors. *IEEE Transactions on Consumer Electronics*, 2018.
- [101] Thierry Bertin-mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.
- [102] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392, 2017.
- [103] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.
- [104] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. Essentia: An open-source library for sound and music analysis. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 855–858, New York, NY, USA, 2013. ACM.
- [105] Sebastian Streich and Perfecto Herrera. Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization. In *In Proceedings of the AES 118th Convention*. Citeseer, 2005.
- [106] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. Music emotion classification: A regression approach. In *2007 IEEE International Conference on Multimedia and Expo*, pages 208–211. IEEE, 2007.
- [107] Dan Su, Pascale Fung, and Nicolas Auguin. Multimodal music emotion classification using adaboost with decision stumps. pages 3447–3451, 10 2013.
- [108] Qi Lu, Chen Xiaoou, Deshun Yang, and Jun Wang. Boosting for multi-modal music emotion classification. pages 105–110, 01 2010.
- [109] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM.
- [110] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M Mc-Nee, Joseph A Konstan, and John Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134. ACM, 2002.
- [111] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, December 2015.

- [112] FO Isinkaye, YO Folajimi, and BA Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261–273, 2015.
- [113] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. Personality & emotional states: Understanding users’ music listening needs. CEUR-WS. org, 2015.
- [114] Definition of flat affect. <https://www.medicinenet.com/script/main/art.asp?articlekey=26293/>. Accessed: 2019-06-25.
- [115] Kirsten Smith, Matt Dennis, Judith Masthoff, and Nava Tintarev. A methodology for creating and validating psychological stories for conveying and measuring psychological traits. *User Modeling and User-Adapted Interaction*, pages 1–46, March 2019.
- [116] Dietmar Jannach, Iman Kamehkhosh, and Geoffray Bonnin. Analyzing the characteristics of shared playlists for music recommendation. In *RSWeb@ Rec-Sys*, 2014.
- [117] Iman Kamehkhosh, Dietmar Jannach, and Geoffray Bonnin. How automated recommendations affect the playlist creation behavior of users. In *IUI Workshops*, 2018.
- [118] Carlos M Jarque and Anil K Bera. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172, 1987.
- [119] John D Hogue, Andrea M Crimmins, and Jeffrey H Kahn. âso sad and slow, so why canât i turn off the radioâ: The effects of gender, depression, and absorption on liking music that induces sadness and music that induces happiness. *Psychology of Music*, 44(4):816–829, 2016.

Appendix A

System for Experiment 2

MooDify

Recommendation List 2

Listen to all the songs from the Recommendation List 2 and Kindly form an opinion on how much you like the entire list. You will be asked few questions about it at the end of this page.

The River

Genre: Folk

00:00 00:00

Sirens

Genre: Soulb-Blues-Folk

00:00 00:00

Benedictus

Genre: Classical

00:00 00:00

Moonlight and Roses

Genre: Electronic

00:00 00:45

I

Genre: Jazz-Experimental

00:00 00:00

Kindly answer few questions on recommendation list 2 in your present mood.

1. On a scale of 1 to 5, How much did you enjoy listening to songs in Recommendation List 2?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

2. On a scale of 1 to 5, How surprising were the songs in Recommendation List 2 for you?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

3. Kindly give your opinion on Recommendation List 2. If you do not have any opinion, type NaN in the textbox below.

Proceed

7/14

Figure A.1: Recommendations generated for List 2

MooDify

Recommendation List 3

Listen to all the songs from the Recommendation List 3 and Kindly form an opinion on how much you like the entire list. You will be asked few questions about it at the end of this page.

Nickles And Dimes

Genre: Hip-Hop

00:00

If

Genre: Classical

00:00

Ghost Dance

Genre: Classical

00:00

00:45

Soleil nocturne

Genre: International-Folk-Experimental

00:00

Bester Jungling

Genre: Classical

00:00

Kindly answer few questions on recommendation list 3 in your present mood.

1. On a scale of 1 to 5, How much did you enjoy listening to songs in Recommendation List 3?

☐

1 ☐

2 ☐

3 ☒

4 ☐

5

2. On a scale of 1 to 5, How surprising were the songs in Recommendation List 3 for you?

☐

1 ☐

2 ☐

3. Kindly give your opinion on Recommendation List 3. If you do not have any opinion, type NaN in the textbox below.

Proceed

B/14

Figure A.2: Recommendations generated for List 3

MooDify

Recommendation List 4

Listen to all the songs from the Recommendation List 4 and Kindly form an opinion on how much you like the entire list. You will be asked few questions about it at the end of this page.

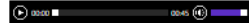
The Window

Genre: Blues-Soulrb



Serenade for Strings Op22 in E Major larghetto

Genre: Classical



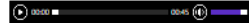
5 00 AM

Genre: Classical



Virtutes Instrumenti

Genre: Classical



Impressions of Saturn

Genre: Classical



Kindly answer few questions on recommendation list 4 in your present mood.

1. On a scale of 1 to 5, How much did you enjoy listening to songs in Recommendation List 4?

☐ 1 ☒ 2 ☐ 3 ☐ 4 ☐ 5

2. On a scale of 1 to 5, How surprising were the songs in Recommendation List 4 for you?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☒ 5

3. Kindly give your opinion on Recommendation List 4. If you do not have any opinion, type NaN in the textbox below.

Proceed

12/14

Figure A.3: Recommendations generated for List 4

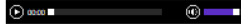
MooDify

Recommendation List 5

Listen to all the songs from the Recommendation List 5 and Kindly form an opinion on how much you like the entire list. You will be asked few questions about it at the end of this page.

Bethlehem Down

Genre: Classical



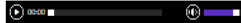
I Need Love

Genre: Rock



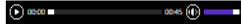
Excessive Resistance to Change

Genre: Classical



The Entertainer (1902 piano roll)

Genre: Country



Bad Trip

Genre: Blues-Soulrb-Rock



Kindly answer few questions on recommendation list 5 in your present mood.

1. On a scale of 1 to 5, How much did you enjoy listening to songs in Recommendation List 5?

☐ 1 ☐ 2 ☒ 3 ☐ 4 ☐ 5

2. On a scale of 1 to 5, How surprising were the songs in Recommendation List 5 for you?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☒ 5

3. Kindly give your opinion on Recommendation List 5. If you do not have any opinion, type NaN in the textbox below.

Proceed

13/14

Figure A.4: Recommendations generated for List 5