NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Emotional Misalignment in Contemporary Music

Systematic Patterns of Audio-Lyrical Contrast Across Genres

Mohamed Taha Ben Attia

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Emotional Misalignment in Contemporary Music

Systematic Patterns of Audio-Lyrical Contrast Across Genres

by

Mohamed Taha Ben Attia

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Mijail Naranjo-Zolotov , PhD ,Nova IMS

Albert Acedo, PhD ,Nova IMS

June, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Mohamed Taha Ben Attia

Lisbon, 23rd of June 2025

ABSTRACT

This thesis investigates emotional alignment between audio features and lyrical sentiment in contemporary popular music, addressing a gap in music emotion research through a novel dual-modality approach. The study analyzed 1,080 songs across 12 genres using K-means clustering of Spotify audio features and GPT-4 lyrical emotion classification, identifying three emotional clusters: Aggressive/Intense, Sad/Calm, and Happy/Upbeat. Key findings reveal that emotional misalignment is the norm rather than the exception: only 46% of songs showed audio-lyrical alignment, while 54% exhibited systematic emotional contrast. Statistical analysis confirmed significant relationships between modalities ($\chi 2 = 99.45$, $p < 0.001$) but slight practical agreement ($\kappa = 0.164$), indicating systematic yet divergent patterns. Genre analysis revealed distinct strategies: hip-hop featured angry lyrics with happy audio (55% mismatches), jazz combined happy lyrics with sad audio (92% mismatches), and metal paired sad lyrics with angry audio (66% mismatches). The research demonstrates that emotional misalignment represents intentional artistic choices creating irony, amplification, or accessibility rather than anomalies. These findings have significant implications for music recommendation systems, therapeutic applications, and AI-generated music, suggesting need for nuanced approaches accounting for multi-modal complexity.

KEYWORDS

Music emotion analysis; Audio features; Lyrical sentiment; Multi-modal analysis; Emotional alignment; Machine learning

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

Statement of Integrity i
Abstract ii
List of Figures v
List of Tables vi
List of Abbreviations and Acronyms vii
1. Introduction 1
2. Literature Review 3
2.1. Emotion from Audio Features 3
2.2. Emotion from Lyrics 9
2.3. Emotions from Lyrics vs. Emotions from Audio Features 14
3.Methodology and Empirical Study 18
3.1. Research Design 18
3.2. Data Collection 20
3.3. Data Preprocessing 21
3.4. Audio-Based Emotion Assigning 23
3.5. Lyrics-Based Emotion Assigning 25
3.6. Alignment Analysis Methodology 26
4.Results and Discussion 28
4.1.Exploratory Data Analysis Results 28
4.2. Cluster Interpretation Framework 32
4.3.Emotion Assigning 34
4.4.Emotion Alignment Analysis 36
4.5.Emotional Alignment Patterns and Statistical Relationships 39
4.6.Genre-Specific Patterns of Emotional Misalignment 40
4.7.Common Emotional Mismatch Patterns and Their Implications 42
4.8. Limitations and Considerations 42
5.Conclusions and Future Works 45
5.1. Summary of Key Findings 45
5.2. Future Research Directions 46
5.3. Conclusions 48
Bibliographical References 50

LIST OF FIGURES

Figure 3.1 -- Feature correlation heatmap for audio features 30
Figure 3.2 -- t-SNE visualization of clustered audio features.....32
Figure 3.3 -- UMAP projection of clustered audio features 33
Figure 3.4 -- Emotion alignment heatmap between audio and lyrics.....36
Figure 3.5 -- Agreement vs. disagreement distribution pie chart.....37
Figure 3.6 -- Genre-specific mismatch patterns heatmap.....38

LIST OF TABLES

Table 2.1 -- Comparison of Key Audio Features and Their Emotional Roles 8
Table 3.1 -- Descriptive statistics for audio features..... 28

LIST OF ABBREVIATIONS AND ACRONYMS

AI Artificial Intelligence - Computer systems able to perform tasks that typically require human intelligence

API Application Programming Interface - A set of protocols and tools for building software applications

BPM Beats Per Minute - A unit of measurement for musical tempo

CNN Convolutional Neural Network - A deep learning algorithm particularly effective for analyzing visual imagery and patterns

dB Decibel - A unit used to measure the intensity of sound or the power level of an electrical signal

EDA Exploratory Data Analysis - An approach to analyzing data sets to summarize their main characteristics

EDM Electronic Dance Music - A broad range of percussive electronic music genres made largely for nightclubs and festivals

GPT Generative Pre-trained Transformer - A type of large language model developed by OpenAI

IQR Interquartile Range - A measure of statistical dispersion, equal to the difference between 75th and 25th percentiles

LDA Latent Dirichlet Allocation - A generative statistical model that allows sets of observations to be explained by unobserved groups

LLM Large Language Model - AI models trained on vast amounts of text data to understand and generate human-like text

LSTM Long Short-Term Memory - A type of recurrent neural network capable of

learning long-term dependencies

MIR Music Information Retrieval - An interdisciplinary science dealing with retrieval of information from music

NLP Natural Language Processing - A branch of artificial intelligence that helps computers understand, interpret and manipulate human language

NMF Non-negative Matrix Factorization - A group of algorithms where a matrix is factorized into two matrices with non-negative elements

POS Part-of-Speech - A category of words that have similar grammatical properties

R&B Rhythm and Blues - A genre of popular music that originated in African American communities in the 1940s

SDG Sustainable Development Goals - A collection of 17 interlinked global goals designed to be a blueprint for a better future

t-SNE t-Distributed Stochastic Neighbor Embedding - A machine learning algorithm for visualization of high-dimensional data

UMAP Uniform Manifold Approximation and Projection - A dimension reduction technique used for visualization and general non-linear dimension reduction

VADER Valence Aware Dictionary and Sentiment Reasoner - A lexicon and rule-based sentiment analysis tool

1. INTRODUCTION

Music represents a fundamental channel through which humans convey and perceive emotion - it can make us feel things deeply, no matter what language we speak or where we're from. That is due to two distinct channels combined: the emotional nature of the sounds themselves, and the literal sense of the words. It is not uncommon to encounter an upbeat, cheerful melody paired with lyrics expressing profound sorrow, or a melancholic, somber tune accompanying words of hope and resilience. These apparent contradictions illustrate a critical truth: musical feeling is not simple - it emerges from the complex interaction between what we know and what we hear.

Researchers have tended to explore these two points in separate ways. On the one hand, music scientists have quantified the way musical features like tempo, acousticness, and valence lead to emotional experiences .On the other hand, methods have been built for analyzing emotion in lyrics, from positive/negative word frequency counts to contextual understanding AI. But surprisingly few studies have attempted to identify how these two channels actually collaborate in real songs that listeners hear.

This gap in our knowledge has real-world implications. Today, when music streaming services recommend songs, they rely mainly on similar sounds or keywords and ignore the emotional link between music and lyrics. In music therapy, where songs are used to allow individuals to process through emotions, therapists would be more effective if they understood the link better. Even as artificial intelligence begins generating music, we need to understand these human factors of emotional expression.

We can find examples of this tension throughout the history of music. Outkast's 'Hey Ya!' features upbeat dance music contrasting with lyrics about a failing relationship. Modern hip-hop generally combines intense, violent lyrics with cheerful, danceable beats. While these seem like deliberate artistic choices, in reality we have very little systematic data about how common these mismatches are within genres, or how much effect they have on listeners.

This study aims to quantify the relationship between the emotional tone of a song’s audio and the emotional content of its lyrics. Using thousands of tracks across genres, the analysis investigates how often these emotions align, diverge, and what patterns emerge. Unsupervised learning techniques were applied to group songs based on Spotify audio features, while a large language model was used to classify the emotional sentiment of lyrics. This combination enables a detailed exploration of the emotional alignment between auditory and lyrical content.

This thesis seeks to answer the following research questions:

- How frequently do audio features and lyrics express the same emotions in contemporary popular music?

- How do emotional alignment patterns differ across musical genres?

- What are the most common patterns of emotional mismatch between audio and lyrics?

These findings have real-world utility. For streaming music services, our research may help enable wiser recommendation tools that are better attuned to emotional nuance. For therapists, it might be a script for selecting more accurately targeted song choices for the therapeutic goals they seek to fulfill. For composers, human or AI, it offers insight into why certain combinations of music and lyrics create profound emotional impact. And for listeners everywhere, it describes those magical moments when a song conveys emotions we're not sure how to express.

2. LITERATURE REVIEW

The emotional impact of music has been widely studied, with research investigating how both lyrical and audio features contribute to constructing the emotional response. Audio features such as energy, valence, loudness, etc … are well established to evoke specific emotions, and the lyrics of a song provide additional information and story. Traditional studies have actually focused on one or the other of these two—a breakdown of emotional content delivered through audio or analyzing sentiment through lyrics. However, newer research now is beginning to look into how these two facets are interconnected and inform each other to establish a song's total emotional impact. This literature review will treat the most critical studies of emotional expression in both lyrics and audio, the approach of analyzing them, and evolving research on the combined effect they have on the emotion of the listener.

2.1. EMOTION FROM AUDIO FEATURES

Introduction to Audio Features and Emotion

Music has a unique power to move us to feel something—whether happiness, sadness, excitement, or nostalgia—even when there are no words involved. This effect is derived largely from the meaning of different elements of sound. Researchers have, over the years, tried to understand the exact way music achieves this.

For instance, (Juslin & Västfjäll, 2008) proposed that emotional reactions to music can be achieved through several pathways— from automatic, physiological reactions to learned associations and personal memories. Thanks to streaming services and advances in audio analysis, it's now possible to break down music into measurable components. These are often called audio features, and they consist of measurements of tempo, energy, valence (musical positivity) ...

Services like Spotify provide these features for millions of songs, making it easier for researchers to examine how sound and emotion are connected on a large scale. In recent years, these features have been used by researchers to examine which musical patterns are most likely to cause specific emotions.

While early studies were often reliant on subjective listener reports, current studies are moving towards examining such patterns using more precise computational methods. The next section discusses audio features in greater depth and how they've been linked to different emotional responses in the literature.

Key Audio Features in Music and Their Scales

In music analysis, audio features are quantitative descriptions of a song that capture different perceptual attributes of the song. They provide an objective means of quantifying the subjective process of listening to music. The following are significant audio features in music analysis, many of which are accessible via streaming platforms like Spotify. Spotify provides fine-grained details of these features via their API, though they do not make publicly available the precise way in which they are computed. Despite this, these features have been widely used for exploring emotional expression in music.

Valence

Valence is the musical positivity of the song. The range is 0.0 to 1.0, where 0.0 is more negative or sad in emotional tone (e.g., sad, depressed, angry), and 1.0 is more positive, upbeat in emotional tone (e.g., happy, cheerful, euphoric). Valence is a significant aspect in determining how listeners perceive the overall mood of a song.

Danceability

Danceability is the degree to which a song is danceable, as per an average of musical attributes such as tempo, rhythmic stability, beat strength, and overall regularity. The scale of measurement is between 0.0 and 1.0, where 0.0 is the least danceable song and 1.0 is the most danceable song. This feature is particularly important in accounting for the role of rhythm and tempo in the physical engagement of the listener with music.

Energy

Energy measures the activity level and intensity of a track, based on perceptual features such as dynamic range, loudness, timbre, onset rate, and general entropy. The scale is 0.0 to 1.0, where 0.0 represents a very quiet or mellow track, and 1.0 represents a very energetic, fast, and loud track. Energy would normally be associated with genres like rock or electronic dance music that are energetic, while classical or ambient music would score low on this dimension.

Loudness

Loudness refers to the overall volume of a track, measured in decibels (dB). The

common values lie within the range of -60 dB to 0 dB, with -60 dB being the quietest and 0 dB being the loudest. Loudness is crucial in the emotional perception of a track, as louder tracks stimulate stronger emotions.

Acousticness

Acousticness is a confidence measurement of whether a track is acoustic or not. The range is from 0.0 to 1.0, with higher numbers indicating greater confidence that the track is acoustic . This feature is handy in distinguishing between acoustic music, which tends to bring about a more intimate, raw emotional response, and electronic music, which has a potential to evoke a different kind of emotional atmosphere.

These features provide a systematic way to quantify and analyze emotional and perceptual properties of music. While Spotify does not disclose exact calculations that they employ to arrive at these features, access to these measurable properties allows researchers the opportunity to study how combinations of these features interact with specific emotional experiences for listeners

The Relationship Between Audio Features and Emotion

The correspondence between audio features and emotional reactions has been long explored in computational musicology and music psychology. Audio features, some of which have been mentioned earlier, have been found to strongly correlate with the emotions experienced in reaction to music. The audio features don't work in isolation of each other; instead, they combine to produce rich emotional reactions. Below, we outline how each of the main features—valence, danceability, energy, loudness, and acousticness—corresponds with listeners' emotional experiences.

Valence and Emotion

Valence is one of the greatest predictors of emotional music response. Studies have shown that valence profoundly influences how individuals categorize songs as happy or sad, and valence has a vital role in systems designed to generate mood-based recommendations (Eerola & Vuoskoski, 2013; Juslin & Västfjäll, 2008)For instance, well-known and upbeat electronic music is high in valence and deployed across a range of commercial and therapeutic uses to stimulate positive affect in users (Saarikallio, 2011). Valence is a key metric in many Music Information Retrieval (MIR) systems and affective computing models that attempt to

forecast the user's preference or emotional reactions (Yang & Chen, 2012). But note that emotional perception is variable in terms of listener's cultural background, personal associations, or current mood—so valence, effective at a general level, won't necessarily capture all the emotional nuance of music for all people (Hunter et al., 2010; Balkwill & Thompson, 1999).

Danceability and Emotion

Danceability is closely tied to emotions associated with physical movement and social interaction. Highly danceable music is commonly found in active, energetic environments—like parties or gym classes—where its effect is to bring about sensations of happiness, excitement, or even euphoria (Trost et al., 2012; Janata et al., 2012). This connection is based on the strong rhythmic and temporal cues that provoke body movement, which is often accompanied by positive affective responses (Leman, 2008; Phillips-Silver & Trainor, 2007). Embodied music cognition studies suggest that the body's interaction with rhythm plays a central role in shaping emotional experience (Maes, 2016). Conversely, low danceable songs can promote stillness and reflection and be in line with more introverted or contemplative emotional states (Saarikallio & Erkkilä, 2007). Danceability is not an emotion encoder, yet it strongly shapes the listener's affective reaction through engagement with rhythm and body motion.

Energy and Emotion

Energy exerts a strong influence on the arousal dimension of emotion in music. High-energy songs—often fast, loud, and rhythmically active—usually induce feelings of excitement, anger, or enthusiasm by stimulating heightened physiological responses like faster heart rate or galvanic skin response (Blood & Zatorre, 2001). Low-energy songs are likely to induce feelings of calmness, serenity, or sadness (Sloboda et al., 2001). While energy by itself does not determine the valence (positive or negative) of a track, it has a significant impact on the degree of the emotional experience and is therefore a key component of mood control and emotional engagement via music (Laukka, 2007). Research suggests that energy's impact on arousal can have have implications for emotional regulation in therapeutic settings (Koelsch et al., 2006).

Loudness and Emotion

Loudness is a key contributor to the intensity of musical emotional responses. Tracks with higher loudness induce stronger emotional states such as excitement, anger, or intensity by enhancing the arousal response of the listener (Laukka, 2007). This supports findings in psychophysiology that show that louder sound pressure levels lead to heightened physiological activation, such as heart rate and skin conductance (Krumhansl, 1997). Conversely, slower songs with decreased loudness are more likely to be associated with weaker emotional responses, such as relaxation or calmness. The emotional impact of loudness lies in its ability to elicit a stronger physiological response—louder music demands more energy and attention, so it is more likely to result in intense emotional responses. In electronic dance music or rock music genres, high loudness reinforces an energetic and arousing emotional atmosphere, while classical or ambient lower-loudness music would bring about a more contemplative or relaxing emotional atmosphere

Acousticness and Emotion

Acousticness is linked closely with affective reactions pointing to intimacy, authenticity, and openness. Pieces with higher values of acousticness—typically composed of natural instruments such as guitar, piano, or strings—are likely to induce emotions such as nostalgia, thoughtfulness, or calmness (Leman et al., 2013; Juslin & Västfjäll, 2008). Acoustic timbres are rated as having higher emotional content due to timbral complexity as well as relating to human playing (Gabrielsson, 2001). The directness and simplicity of acoustic music contribute to its emotional depth, producing a stronger feel of closeness between the listener and the performer. Electronic or synthesized music, lower in the acousticness scale, tends to produce emotional atmospheres more distant, stylized, or artificial (Zentner et al., 2008). The acoustic sound of a track allows for a more grounded emotional experience, and this is especially well-suited to genres like folk, singer-songwriter, or acoustic pop, where emotional storytelling and personal reflection are key.

Comparative Overview of Audio Features and Emotion

The table below provides a summary of important emotional roles and contrasts to help you better understand how each audio feature impacts your emotional perception.

Feature	Scale	Emotions Associated	Effect Type
Valence	0.0 (negative) – 1.0 (positive)	Distinguishes happy vs sad emotional tone	Emotional valence (positive–negative)
Danceability	0.0 – 1.0	Promotes joy, movement, and physical engagement	Physical engagement, mood boost
Energy	0.0 – 1.0	Influences arousal / feeling excited or relaxed	Arousal level (high–low)
Loudness	-60 dB – 0 dB	Heightens emotional intensity and arousal	Intensity and physiological response
Acousticness	0.0 – 1.0	Enhances emotional intimacy and natural feel	Emotional closeness, introspection

Table 2.1 -- Comparison of Key Audio Features and Their Emotional Roles

Emotions Arising from Feature Combinations

Musical emotion emerges from the complex interaction of acoustic and structural attributes rather than isolated audio features (Eerola & Vuoskoski, 2013). Three common emotional profiles demonstrate how feature combinations create distinct affective experiences based on empirical evidence from music psychology and affective computing research.

Aggressive and angry music typically exhibits low valence aligned with negative emotional content, combined with high energy and loudness that convey urgency and aggression through intensive, forceful sounds (Eerola & Vuoskoski, 2013; Ilie & Thompson, 2006; Zentner et al., 2008). These tracks often feature low acousticness with distorted or synthesized sounds dominating the mix, as organic instrumentation tends to reduce perceived hostility, while maintaining moderate-to-low danceability since aggressive intensity tends to suppress groove sensations and bodily engagement (Yang et al., 2008; Lartillot et al., 2008). This configuration appears typically in genres such as hard rock, metal, or aggressive electronic music, where low valence and high energy combine to create listener perceptions of aggression, tension, or anger independent of lyrical content (Juslin & Laukka, 2004).

Sad and calm music presents an inverse profile characterized by low valence conveying emotional negativity, paired with low energy and loudness creating serene, reflective soundscapes through soft, slow, restricted dynamics with light attack and sparse texture (Grewe et al., 2007). High acousticness distinguishes this category, with acoustic instruments like piano, strings, or unprocessed vocals optimizing emotional exposure, while moderate danceability provides contemplative sway rather than active movement or dance (Krumhansl, 1997). The reduction in arousal features—energy, loudness, and tempo—alongside low valence creates emotional reflection and calmness, consistently rated as melancholic, calm, or sad in emotion-tagging studies (Zentner et al., 2008; Eerola & Vuoskoski, 2013).

Happy and upbeat music emerges through high positive emotional valence creating cheerful, joyful experiences, combined with high danceability featuring evident rhythmic patterns and salient beats that encourage physical movement (Yang et al., 2008; Soleymani et al., 2013). Moderate-to-high energy transmits brightness and excitement through energetic delivery, quick tempo, and high dynamics, while low acousticness incorporates artificial instruments and polished production resulting in modern, light emotional environments (Juslin & Laukka, 2004). Unlike aggressive music, happy tracks balance moderate loudness to avoid listener fatigue while remaining engaging, with this feature intersection commonly found in dance pop, funk, or tropical house music consistently eliciting happiness, playfulness, and optimism in listeners (Eerola & Vuoskoski, 2013; Zentner et al., 2008).

2.2. EMOTION FROM LYRICS

While audio features provide sound context to emotional expression, lyrics provide direct access to the semantic content of a song. They play a controlling role in expressing emotion in narrative form, lexical choice, and figurative language. Studies on emotions in lyrics gained momentum with the introduction of Natural Language Processing (NLP) tools enabling an understanding of how text induces affective responses (Mihalcea & Strapparava, 2005; Mohammad, 2020).

Introduction to Lyrics and Emotion

Lyrics are an important aspect of musical storytelling, allowing musicians to express internal experiences and connect with listeners on a personal level (Juslin & Laukka, 2004). In contrast

with audio features—typically processed involuntarily—lyrics are cognitively processed, engaging the listener's linguistic and affective capacities.

There are several studies that have demonstrated how strongly lyrics can influence listeners' emotional experience of music. For instance, Eerola and Vuoskoski (2011) found that lyrical content strongly influences perceived emotion, especially when there is correspondence between lyrics and musical attributes.

Key Linguistic Methods in Lyrical Emotion Analysis

Emotion in lyrics is typically analyzed with NLP models that quantify textual features and associate them with emotional states. The conventional approaches include lexicon-based analysis, machine learning classification, and deep learning techniques. Some of the key features utilized in lyric emotion detection are:

Sentiment Polarity

Simple sentiment analysis labels lyrics as positive, negative, or neutral with a polarity score through lexicons such as VADER (Hutto & Gilbert, 2014) or TextBlob. Whereas polarity detection provides only a rough estimate of emotional tone, it is nonetheless a simple tool for lyric analysis (Yang & Lee, 2009).

Emotion Lexicons

Lexicons like the NRC Emotion Lexicon (Mohammad & Turney, 2013) provide words to specific emotions (e.g., anger, fear, joy, sadness). They are generally used in multi-label classification models to determine the presence of a set of emotions in song lyrics (Mayer et al., 2008; Trohidis et al., 2008). They have been successfully applied in large-scale music emotion datasets research (Delbouys et al., 2018).

POS Patterns and Grammatical Structures

Research has shown that first-person pronouns, affective adjectives, and emotion verbs (e.g., "love," "hate") are used more frequently in affectively expressive songs (Pennebaker & King, 1999; Tausczik & Pennebaker, 2010). This information can be scraped using Part-of-Speech (POS) tagging and used in machine learning pipelines for classification (Yang et al., 2008).

Topic Modeling and Thematic Clustering

Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) have been employed to uncover recurring emotional and thematic topics in lyrics datasets (Hu & Downie, 2010; Schedl et al., 2014). These methods help identify dominant themes like heartbreak, nostalgia, or empowerment that correlate with emotional tags.

Word Embeddings and Deep Learning

Current approaches utilize word embeddings like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and contextual representations like BERT (Devlin et al., 2019) to better capture semantic relationships in lyrics. Deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown good performance in multi-label emotion recognition (Delbouys et al., 2018).

Metaphor and Figurative Language

Metaphorical language is necessary when explaining complicated or ambiguous emotional circumstances. Work such as Shutova et al. (2010) and Veale et al. (2016) focuses on metaphor identification in text and presents issues with automatic interpretation of figurative language in song lyrics. As important as it is, metaphor identification in music is an unexamined area.

Emotions Arising from Lyrics

Emotions conveyed in song lyrics emerge from the convergence of linguistic markers, affect, and semantic meaning rather than individual words alone (Mohammad & Turney, 2013; Pennebaker & King, 1999). Three common emotional profiles demonstrate how linguistic patterns create distinct affective experiences in lyrical content based on empirical studies in computational linguistics and music psychology.

Aggressive and angry lyrics typically exhibit very negative sentiment polarity, with lexicons identifying words that connotatively induce hostility, conflict, and frustration while featuring high frequencies of emotion verbs like "hate" and "fight" alongside emotion adjectives such as "angry" and "furious" as common linguistic markers of aggression (Mohammad, 2012; Tausczik & Pennebaker, 2010). These lyrics employ direct and confrontational language through first-person pronouns combined with assertive or imperative mood, indicating personal and affective intensity, while demonstrating low use of soothing or reflective terms, with the absence of conciliatory words attesting to the violent tone characteristic of genres

like punk, metal, and hardcore hip-hop where emotional outburst serves as a fundamental element (Pennebaker & King, 1999; Yang et al., 2008; Mayer et al., 2008).

Sad and calm lyrics convey subdued negative sentiment with less overt aggression than angry lyrics, featuring frequent use of reflective and melancholic words related to loss, solitude, and introspection such as "tear," "alone," and "cry" that recur regularly throughout the content (Yang & Lee, 2009; Mohammad & Turney, 2013). These lyrics demonstrate extensive presence of figurative language through metaphors and similes used to express complex emotions implicitly, combined with emphatic first-person singular use where "I" and "me" enhance personal emotional expression (Shutova et al., 2010; Pennebaker & King, 1999). The overall tone remains subdued and resigned rather than confrontational, with lessened emotional arousal consistent with genres like ballads, folk, and soft rock music where introspective content dominates (Schedl et al., 2014).

Happy and upbeat lyrics score high in positive sentiment polarity, consisting of words expressing happiness, celebration, and optimism while featuring high frequencies of positive affect words such as "love," "smile," "dance," and "sunshine" that depict lightness and happiness (Mohammad & Turney, 2013). These lyrics employ repetitive and melodic vocabulary with simple-to-remember words that reinforce positive mood and memorability, often emphasizing community and togetherness themes through shared experience, friendship, and party atmosphere content (Yang et al., 2008; Delbouys et al., 2018). These characteristics appear typically in fast-paced and positive pop, dance, and country music that enhance energetic and positive emotional states through their lyrical content (Soleymani et al., 2013).

Large Language Models in Lyrical Emotion Analysis

GPT-4 operates as a transformer-based Large Language Model (LLM), in contrast to conventional methods of emotion classification that depend on rule-based systems or pre-defined emotion lexicons (Mohammad & Turney, 2013). Instead of using a predetermined "library of emotional words," GPT-4 uses the text's contextual and semantic structures to infer meaning and emotion. The deep neural architecture of the model, which converts linguistic patterns into high-dimensional vector spaces, accomplishes this (Vaswani et al., 2017; Brown et al., 2020).

Also the model uses its internalized understanding of emotion, which it has developed through training on extensive and varied text corpora, to interpret input when asked to categorize the emotion of song lyrics. As a result, even in situations without explicit emotional keywords, it can recognize subtle emotional cues like metaphor, irony, or tone and determine an overall affective label.

GPT-4 is particularly suited to emotion classification in music lyrics due to:

- **Contextual understanding:** It evaluates meaning in context, allowing more accurate emotion classification across diverse narrative styles.
- **Handling of figurative language:** It can parse metaphor, allegory, and poetic structure—common elements in lyrical writing.
- **Zero-shot generalization:** It does not require retraining or fine-tuning for new emotion categories or genres.

These characteristics align with the challenges of processing creative, emotional, and non-literal forms of language found in music (Floridi & Chiriatti, 2020).

The Role of Genre, Culture, and Context

Lyrical emotion is highly conditioned by genre conventions. For example, vulnerability and loss are the norms for country and blues, and empowerment or rebellion themes dominate the likes of hip-hop or rock (Kim et al., 2010). Variation needs to be treated in building generalized feeling recognition models.

Culture also determines how lyrics are written and interpreted. Emotions in music are varying cross-culturally, both by frequency and meaning. For instance, the metaphoric use of nature (e.g., "storms", "rain") will convey sorrow in Western cultures but can carry different meanings in non-Western cultures (Tagg, 2012). Contextual elements such as historical period, artist background, or audience inference affect the emotional meaning of lyrics as well.

2.3. Emotions from Lyrics vs. Emotions from Audio Features: A Comparative Perspective

In order to understand emotion in music, one must consider two inherently different but complementary sources of emotional transmission: audio features and lyrics. Audio features

provide us with the sonic information—melody, rhythm, harmony—that conveys affective information in terms of sound, habitually evoking effortless emotional responses. Lyrics provide us with explicit semantic content manifesting emotion in an overt language form, engaging higher-level cognitive processes. This section analyzes the differences, interaction, and the delicate interplay among these modalities grounded in cross-disciplinary research in cognitive neuroscience, music psychology, computational linguistics, and cultural studies.

Cognitive and Neural Processing Differences

Cognitive neuroscience verifies that the brain processes lyrical and musical emotion via distinct yet interconnected pathways. Musical parameters such as tempo, energy, and acousticness primarily influence the limbic system—affecting affect and reward—along with auditory and motor areas processing rhythm perception (Koelsch, 2014; Zatorre et al., 2007). These systems often produce pre-linguistic, reflex responses such as chills or rhythmic entrainment, which are likely to be cross-culturally universal.

Conversely, lyrics entail semantic comprehension and higher-order cognitive processing, activating language-related brain regions such as Broca's and Wernicke's areas (Friederici, 2011). Comprehension of lyrical emotion entails reading metaphors, narrative setting, and cultural allusion, allowing listeners to access subtle, context-dependent emotions.

This neurobiological distinction partially explains the propensity of instrumental music to invoke general, generalized emotion, while lyrics provide specific, situational emotional detail.

Patterns of Lyrical and Audio Emotional Relationships

The relationship between emotions conveyed by lyrics and audio features varies considerably across musical contexts, creating different patterns of emotional expression:

Emotional Congruence

The overall emotional impact of many songs is increased by the alignment of lyrical and audio-derived emotions. Adele's "Someone Like You," for instance, reinforces melancholy with its

slow, minor-key piano accompaniment and melancholic lyrics. Similar to this, Bruno Mars' "Uptown Funk" combines joyful, upbeat lyrics with lively, high-energy instrumentation to produce a cohesive expression of happiness. According to this correlation, clear emotional messaging can be created through mutual reinforcement (Yang & Lee, 2009).

Emotional Divergence

Other songs intentionally create irony or complexity by fusing disparate emotions in the lyrics and audio. Outkast's "Hey Ya!" creates a mixed or ironic emotional effect by combining lyrics about ending a relationship with lively, danceable music. In a similar vein, "Every Breath You Take" by The Police pairs calming, pleasant music with lyrics about stalkers. This dissonance demonstrates how audio and lyrics can independently alter how emotion is perceived (Hu & Downie, 2010).

Factors Affecting Relationship Strength

Several factors affect the strength of emotional relationships between lyrics and audio features across different musical contexts. Genre conventions play a significant role, with rap and electronic music frequently exhibiting greater independence between lyrical and musical emotions, whereas pop and ballads tend to demonstrate stronger emotion alignment through more integrated compositional approaches. Cultural expectations impact emotional coupling based on how various traditions emphasize lyrics versus music, with some cultures prioritizing textual content while others focus more heavily on instrumental expression. Production choices also influence these relationships, as artists may purposefully contrast sound and lyrics to create tension or subtext that enhances the overall artistic message and emotional complexity of the work (Moore, 2012).

Genre and Cultural Contexts in Emotional Expression

By genre, the proportion of lyrics to audio features varies greatly. Using semantic and rhetorical complexity, hip-hop and rap prioritize lyrical content for social commentary and emotional storytelling. For instance, rather than its musical qualities, Kendrick Lamar's "Alright" primarily draws its emotional impact from its lyrical message of resiliency and hope.

In contrast, ambient and electronic dance music (EDM) frequently rely solely on audio elements, emphasizing atmosphere, texture, and rhythm to elicit feelings without the use of lyrics. Instead of communicating verbally, artists such as Aphex Twin and Brian Eno manipulate audio parameters to create intricate emotional landscapes.

Emotional perception is further influenced by cultural context. Due to differences in tonalities and cultural signifiers, metaphors and emotional symbols that are frequently found in Western music may not translate directly across cultural boundaries (Tagg, 2012). For example, the minor scale may not have the same emotional meaning in East Asian or Middle Eastern musical traditions, but it frequently denotes melancholy in Western music.

Emotion Granularity and Multidimensionality

Both lyrics and audio contribute to the multidimensional nature of musical emotions, which extend beyond simple positive or negative valence. Models such as Valence-Arousal-Dominance (Russell, 1980; Plutchik, 1980) describe emotions along multiple axes. Valence represents the positivity or negativity of the emotion, while arousal indicates the intensity or energy level of the emotional experience. Dominance reflects the sense of control or power within the emotional experience. These three dimensions work together to create a comprehensive framework for understanding emotional complexity in music.

According to research, lyrics affect valence through semantic content and dominance through narrative perspective, whereas audio features frequently drive arousal through parameters like tempo, energy, and loudness (Yang & Lee, 2009). For instance, a song with upbeat instrumentals and depressing lyrics may elicit both negative valence and high arousal at the same time.

Listener Perception and Contextual Influences

In addition to song content, listener characteristics like language proficiency, cultural background, memories, and mood at the moment of listening all influence how an emotion is perceived. Emotions might be experienced differently by a listener who speaks the language of the lyrics well than by someone who reacts mainly to the musical elements.

Emotional interpretation is also influenced by contextual elements such as the artist's persona, historical period, and social setting . For instance, the emotional impact of protest songs can vary based on the listener's political beliefs or historical background.

Automatic emotion recognition is put to the test by this contextual sensitivity, but it also creates opportunities for contextually and individually-aware personalized affective computing.

3. METHODOLOGY

This chapter presents both the methodological frameworks employed in this study and the empirical findings from applying these methods. Each section first describes the methodological approach, followed by the empirical results obtained from that approach, clearly distinguishing between the procedural framework and the actual findings.

The approach used to examine the emotional connection and contrast between the lyrical content and the audio characteristics of music is presented in this chapter. With theoretical foundations, it offers a thorough explanation of the data sources, preprocessing steps, exploratory analysis, and the use of machine learning and natural language processing techniques.

3.1. RESEARCH DESIGN

This study aims to investigate and assess the relationship and possible contrast between the emotional meaning of lyrics and the emotional tone of music audio features. The study is based on a multidisciplinary methodology that combines natural language processing (NLP), data science, and music informatics concepts. Three phases make up the study's methodology, each of which is intended to address one of the main goals:

Phase 1: Data Extraction and Preprocessing

This phase is foundational and concerns the collection and preparation of data to be analyzed. The first component involves audio feature collection, retrieving quantitative musical features for each track using Spotify's Web API. Features such as valence, energy, and danceability are essential to represent the emotional character of music from a signal-based perspective. The second component focuses on lyrics collection, extracting the textual content of song lyrics using the Genius API to provide the semantic, language-based dimension of emotion. Finally, preprocessing ensures both datasets undergo cleaning and transformation. For audio features, this involves normalization and selection of relevant variables, while for lyrics, this includes cleaning, removing non-textual content (e.g., [chorus], [verse]), lowercasing, and removing stop words. This step ensures data compatibility and quality for modeling and transforms raw multimedia content into structured numerical and textual representations suitable for computational modeling.

Phase 2: Emotion Inference

Using two distinct approaches, the second phase seeks to infer the emotional label from the audio and lyrics data sources.

A. Clustering on Audio Features (Unsupervised Learning)

The audio features are handled by a clustering algorithm in order to categorize the emotional tone according to the structure of the music. This makes it possible to find organic groupings in the data without the need for labels beforehand.

- The dataset is divided into groups based on similar emotional profiles (e.g., happy, sad, angry) using K-Means clustering.
- Silhouette analysis is used to choose the optimal number of clusters, and the aggregate values of valence, energy, danceability, acousticness, and loudness are used to interpret the clusters.

This approach uses audio signals to implicitly interpret emotional tone.

B. Emotion Classification from Lyrics (Language Model-Based NLP)

To extract emotion from lyrics, the study uses a large language model (LLM) — specifically the ChatGPT API.

- In order to instruct the model to return a single emotion label—Happy, Sad or Angry—the lyrics of each song are run through a predefined prompt.
- This allows for cross-modal comparison by converting the textual data into a structured emotion label.

Because generative transformer-based models are better at comprehending context, emotion, and subtle meaning in natural language, they are preferred over traditional text classification.

Phase 3: Emotional Alignment Analysis

This phase compares the emotion labels inferred from the two modalities — audio and lyrics — to determine the level of alignment or contrast.

- Agreement Rate: Calculated as the percentage of tracks where audio and lyrics express the same emotion.
- Statistical Testing: A Chi-Square Test of Independence is used to assess whether emotion labels from both sources are significantly related.
- Mismatch Patterns: For tracks where emotion labels differ, further analysis is conducted to explore possible genre-based or stylistic explanations.

This final phase provides both quantitative evidence (agreement rates, chi-square scores) and qualitative insights into how emotion manifests differently in lyrics vs. musical composition.

3.2. DATA COLLECTION

Two main sources were used in the data collection process: song lyrics that were pulled from the Genius API and audio features of songs that were taken from the Spotify Web API. These two sources were selected to allow for a dual-modality approach to emotion analysis by capturing the lyrical content as well as the acoustic qualities of music.

Audio Features from Spotify

Audio features were retrieved using Spotify's Web API, which provides comprehensive metadata for millions of tracks including valence (musical positiveness), danceability, energy, acousticness, loudness, and other features widely validated in music emotion research (Kim et al., 2010).

To build a musically diverse dataset capable of capturing a wide range of emotional expressions, 100 songs were initially collected for each of 12 main musical genres: Pop, Hip-hop, Rock, Jazz, Electronic, R&B, Country, Metal, Alternative, Blues, Soul, and Dance, totaling 1,200 songs. The tracks were selected randomly from various Spotify playlists, with care taken to ensure that each playlist reflected the intended genre. This random sampling from multiple curated sources helped guarantee genre authenticity and variety.

Lyrics from Genius

Lyrical content was collected using the Genius API, one of the largest platforms for crowdsourced lyrics. The retrieval process matched songs using artist name and track title information obtained from the Spotify dataset.

Approximately 95% of the lyrics were successfully retrieved using this method. However, a small portion of songs (around 5%) could not be matched or retrieved, primarily due to their absence from the Genius database or inconsistencies in naming formats (e.g., remixes, special characters, or rare songs).

This limitation led to incomplete lyrical data in some genres, creating an imbalance that could potentially affect downstream analysis. To maintain consistency and equal representation across genres, the dataset was adjusted by reducing the number of songs per genre to 90 songs each. This ensured that all songs included in the final dataset had both audio features and lyrics, allowing for robust multimodal analysis.

Dataset Overview

Both audio features and lyrics for each track were included in the final merged dataset, which was created by aligning the two sources according to song title and artist name. The dataset included metadata such as song name, artist, genre, audio features, and full lyrics, and there were 1080 total entries.

3.3. DATA PREPROCESSING

Data preprocessing is essential to clean and prepare the data for analysis and modeling. Different techniques were used for structured (audio features) and unstructured (lyrics) data.

Audio Features Preprocessing

Genre-Based Sampling and Initial Cleaning

The audio feature dataset was collected using the Spotify Web API, which provides a variety of features such as valence, danceability, energy , etc . To ensure a well-balanced and emotionally diverse dataset, 90 songs were selected from each of 12 different musical genres: alternative, blues, country, dance, electronic, hip-hop, metal, rock, soul, jazz, pop, and R&B. This genre-stratified sampling was essential to prevent the emotional bias that could result from an uneven distribution of genres, and to ensure that the subsequent clustering and

classification steps would be meaningful across various musical contexts. After selection, the dataset was cleaned by removing irrelevant attributes such as track IDs, album names, and external URLs. Only features like Artist Name, Track title, genre and numerical audio features were kept.

Handling Missing Values

No significant missing data were found due to the quality of Spotify's API. Still, a scan for null values was performed to ensure data integrity.

Feature Scaling

Feature scaling is crucial because clustering algorithms, such as k-means, are sensitive to feature magnitude. The RobustScaler was used to maintain robustness against skewed distributions and outliers. RobustScaler centers data using the median and scales it according to the interquartile range (IQR), in contrast to Min-Max scaling or Z-score normalization, which are significantly impacted by outliers (Pedregosa et al., 2011).

This technique ensured that audio features , which can vary greatly across musical genres, did not dominate distance calculations.

$$X_{\text{scaled}} = \frac{X - Q_2}{Q_3 - Q_1}$$

Where Q2 is the median, and Q1 , Q3 are the first and third quartiles respectively.

Preprocessing of Lyrics

Lyrics for each song were collected using the Genius API, a widely-used platform for retrieving annotated music lyrics. The raw lyrical content was first cleaned by converting all text to lowercase and removing punctuation, special characters, and bracketed metadata such as “[Chorus]”, “[Intro]”, and repetition indicators like “repeat x2”.

This step ensured the text retained only meaningful and interpretable lyric content. Importantly, no further natural language processing techniques such as tokenization or stopword removal were applied at this stage. This decision was intentional, as the cleaned but intact full lyrics were later used as input for the ChatGPT API to classify emotional content.

Removing stopwords or breaking the structure of the lyrics could have impaired the language model’s ability to accurately infer nuanced emotional tone. Thus, minimal but purposeful cleaning was conducted to preserve semantic integrity while eliminating non-lyrical noise.

3.4. AUDIO-BASED EMOTION ASSIGNING

This section explains the approach used to use unsupervised clustering techniques to assign emotional labels to audio segments. Feature correlation analysis, feature selection, cosine distance clustering, dimensionality reduction for visualization, and interpretative profiling of the resultant clusters were all steps in the process.

Feature Selection Process

Based on a literature review and preliminary correlation analysis, five key audio features were selected for clustering due to their established connections with emotional perception in music psychology. These features include valence, which measures musical positivity on a scale from 0.0 to 1.0; danceability, indicating how suitable a track is for dancing (0.0 to 1.0); energy, reflecting the perceptual intensity of a track (0.0 to 1.0); acousticness, representing the likelihood that a track is acoustic (0.0 to 1.0); and loudness, which measures the overall volume, typically ranging from -60 to 0 decibels.

Clustering Methodology and Optimal Configuration

K-Means clustering was selected based on its efficiency and interpretability, following prior work in music informatics. While alternative algorithms (e.g., DBSCAN, Agglomerative) could be explored in future work, K-Means provided robust results for the current dataset.

K-Means clustering was applied using cosine distance rather than standard Euclidean distance. Cosine distance evaluates the angular difference between vectors, emphasizing spectral shape similarity rather than magnitude—particularly beneficial in audio data where loudness may vary while spectral characteristics remain consistent.

Cosine distance, which is the measure used in clustering, is then derived as:

$$\text{cosine\_distance}(A, B) = 1 - \frac{A \cdot B}{\|A\|\|B\|}$$

Where:

- A·B is the dot product of vectors A and B
- ‖A‖ and ‖B‖ are the Euclidean norms (magnitudes) of the vectors

The optimal number of clusters was determined using silhouette analysis, which measures how similar an object is to its own cluster compared to other clusters. Multiple cluster configurations (k = 2 to 10) were tested, with k = 3 yielding the highest silhouette score.

Cluster Emotion Assignment Framework

The assignment of emotional labels to the three clusters was guided by a combination of analytical and theoretical approaches. This included analyzing the cluster centroids—the mean values of the selected audio features—alongside comparisons with established emotional profiles from music psychology literature. Additionally, the genre distribution within each cluster and the theoretical framework of Russell's (1980) circumplex model of affect were considered. Based on these criteria, the clusters were labeled as follows: tracks with high valence, high danceability, and moderate energy were categorized as Happy/Upbeat; those with low valence, high energy, and low acousticness were identified as Aggressive/Intense/Angry; and tracks with low valence, low energy, and high acousticness were classified as Sad/Calm/Melancholic.

3.5 LYRICS-BASED EMOTION ASSIGNING

In this study, GPT-4 was employed as the large language model (LLM) for classifying the emotional content of song lyrics. Its advanced natural language understanding capabilities made it suitable for analyzing and interpreting lyrical data across a wide range of musical genres.

Prompt Engineering Strategy

Prompt engineering was used to guide the model's task execution. The final prompt design was:

You are an emotion classification model. Analyze the following song lyrics and classify them into one of the following categories:

0 – Aggressive / Intense / Angry

1 – Sad / Calm / Melancholic

2 – Happy / Upbeat

This structured format can be viewed as a form of instruction-based learning, where the LLM receives task-specific directions. While this could be interpreted as "biasing" the model toward specific categories, the classification decision itself is still made autonomously by the model. The prompt does not embed rules or keywords but instead sets task boundaries for consistent labeling.

This method can be classified as zero-shot classification, where the model is not shown any prior labeled examples but infers the classification task solely from the natural language prompt (Brown et al., 2020). Although the use of numbered categories may resemble conventional machine learning labels, the actual classification is driven by GPT-4's interpretation of lyrical emotion, not by any fixed rule or logic embedded in the prompt.

Lyrics Classification Process

The emotion classification process involved submitting preprocessed song lyrics to GPT-4 alongside the engineered prompt. The model analyzed the lyrical content and returned a single integer (0, 1, or 2) representing the predicted emotional category. These classifications were systematically collected and structured for comparative analysis with audio-based emotions.

This approach enabled scalable emotion annotation across diverse musical styles without requiring human raters or supervised model training, while leveraging GPT-4's contextual understanding capabilities.

3.6 Alignment Analysis Methodology

Alignment Metrics

To assess the relationship between emotional content in audio and lyrics, three primary alignment metrics were applied: agreement rate, chi-square test of independence, and Cohen's Kappa coefficient. The agreement rate measured the percentage of songs in which both modalities conveyed the same emotion, calculated as (Number of matching songs / Total songs) × 100. To statistically evaluate whether a significant association exists between audio-based and lyrics-based emotion, a chi-square test of independence was performed. The null hypothesis (H_o) assumed that the two emotion sources were independent, while the alternative hypothesis (H_1) proposed that they were statistically related. This test provided an objective basis for determining emotional dependency between modalities. In addition, Cohen's Kappa coefficient (κ) was calculated to measure the level of agreement beyond chance, using the formula $\kappa = (P_o - P_e) / (1 - P_e)$, where P_o is the observed agreement and P_e the expected agreement by chance. The results were interpreted according to standard benchmarks: values below 0.20 indicate slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and values above 0.80 suggest almost perfect agreement.

Genre-Specific Analysis

Emotional mismatches between audio and lyrics were further examined across different musical genres to uncover underlying stylistic and cultural influences. This analysis aimed to identify genre-specific conventions in emotional expression, recurring patterns of contrast

between modalities, and contextual factors that may shape emotional alignment in distinct musical traditions.

4. RESULTS AND DISCUSSION

This chapter presents a comprehensive analysis and interpretation of the findings from the emotional alignment study between audio features and lyrical sentiment in music. The results reveal complex patterns of emotional expression that challenge conventional assumptions about how music communicates emotion across different modalities.

4.1. EXPLORATORY DATA ANALYSIS RESULTS

Exploratory Data Analysis (EDA) was conducted to gain an initial understanding of the structure and distribution of both the audio features and lyrical content of the dataset.

Audio Features Analysis

Feature	Mean	Std Dev	Min	25%	50%	75%	Max
danceability	0.594605	0.153693	0.132	0.492	0.6025	0.70675	0.951
energy	0.61871	0.222087	0.0264	0.4632	0.644	0.79475	0.997
loudness	-7.12178	3.194407	-21.29	-8.695	-6.376	-4.8585	-0.73
Speechiness	0.079899	0.077691	0.0	0.0348	0.0478	0.088	0.464
acousticness	0.28312	0.30662	0.0	0.022	0.13	0.50	0.984
instrumentalness	0.014885	0.066544	0.0	0.0	0.0	0.000226	0.556
liveness	0.176964	0.125441	0.0243	0.098125	0.126	0.22475	0.783
valence	0.475606	0.233528	0.0	0.2905	0.469	0.66	0.979
tempo	122.008	30.1	45.857	98.02	120.03	141.86	208

Table 3.1 -- Descriptive statistics for audio features

The analysis of audio features revealed distinct distribution patterns across the dataset. Central tendency analysis showed danceability with a mean of 0.595 ± 0.154, where the close alignment between mean and median (0.603) suggests a relatively symmetric distribution. Energy exhibited considerable variability (mean = 0.619 ± 0.222) with an extremely wide range from 0.026 to 0.997, indicating substantial diversity between low and high-energy tracks. Tempo averaged 122.0 ± 30.1 BPM with a broad range spanning from 46 to 208 BPM, while valence showed a mean of 0.476 ± 0.234 with balanced quartile distribution (Q1 = 0.291, Q3 = 0.660), indicating diverse emotional representation across the dataset.

Distribution asymmetries were evident in several features. Acousticness demonstrated right-skewed distribution with a median of 0.13 significantly below the mean of 0.283, and 75% of values below 0.50, indicating predominance of electronic over acoustic content. Instrumentalness showed extreme right-skew with a median of 0 and 75th percentile of only 0.000226, reflecting the vocal-dominant nature of popular music. Similarly, speechiness exhibited right-skew with a median of 0.048 well below the mean of 0.080. Loudness values clustered between -8.695 dB (Q1) and -4.859 dB (Q3), though some extreme outliers extended to -21.29 dB. These patterns reveal a dataset dominated by vocal, electronically-produced music with diverse energy levels and emotional content.

Feature Correlation Analysis

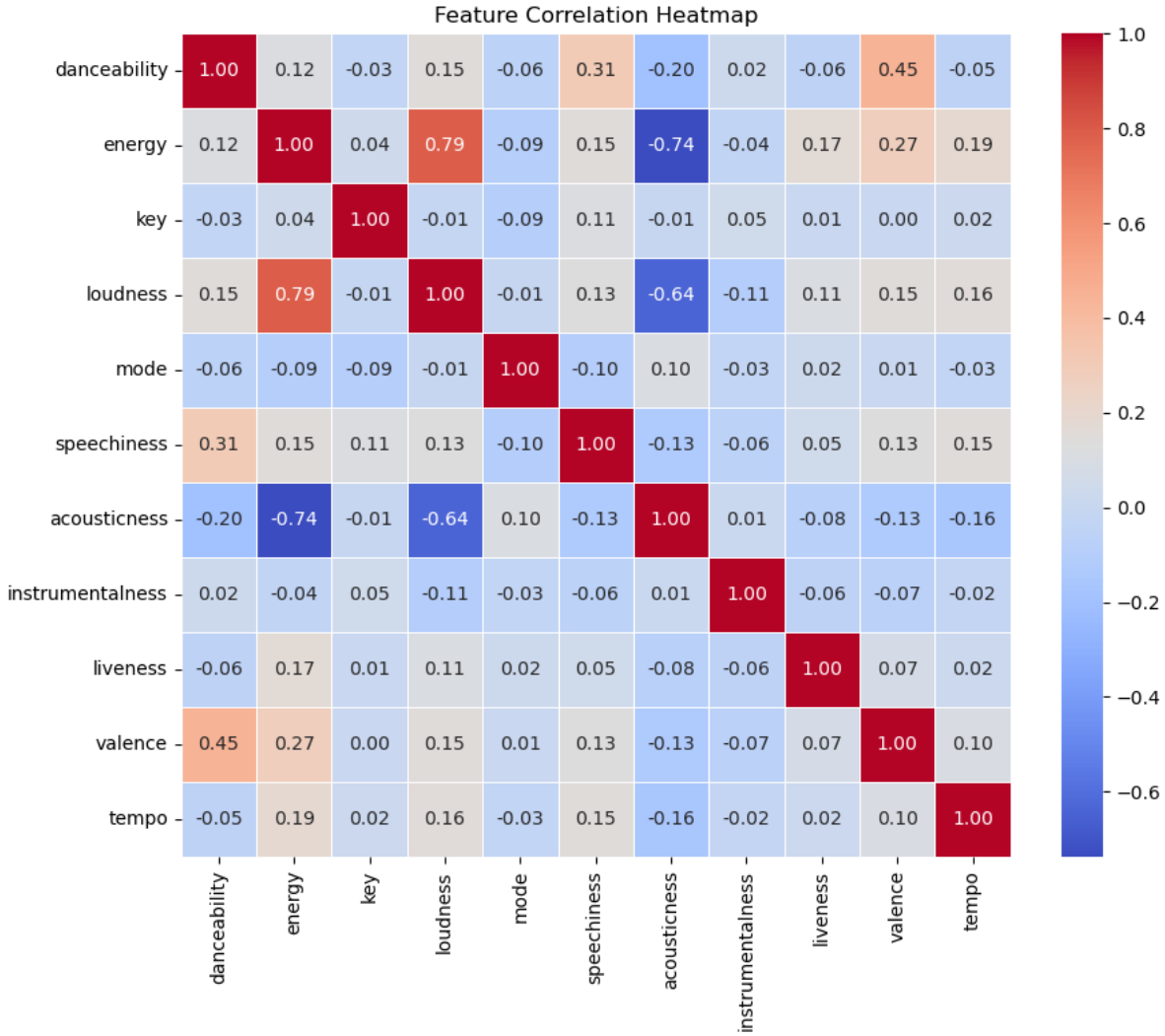

Figure 3.1 -- Feature correlation heatmap for audio features

The Figure 3.1 revealed distinct patterns across different strength levels. Strong correlations ($|r| \geq 0.7$) were observed between energy and loudness ($r = 0.79$), energy and acousticness ($r = -0.74$), and loudness and acousticness ($r = -0.64$), indicating that these features capture related aspects of musical intensity and production style. Moderate correlations ($0.3 \leq |r| < 0.7$) emerged between danceability and valence ($r = 0.45$) and danceability and speechiness ($r = 0.31$), suggesting connections between rhythmic engagement and emotional positivity. Weak or negligible correlations ($|r| < 0.3$) characterized key/mode relationships with all

features, tempo and valence (r = 0.10), and instrumentalness with other features, indicating these elements operate relatively independently in the dataset.

Lyrics Analysis

An exploratory overview of the lyrical data was done even though no advanced NLP preprocessing was used because the ChatGPT API would process the lyrics later. This involved comparing text length, word count, line count, and other fundamental statistics across genres. These metrics assisted in identifying anomalies that might affect the quality of emotion classification, such as lyrics that are abnormally long or too short. Additionally, to investigate whether verbosity might be associated with particular emotional tones, bar plots were used to visualize genre-level differences in average lyric length.

The exploratory analysis of lyrics revealed significant variation in verbosity across genres. Hip-hop demonstrated the highest average word count (1003 words) and character length (5166 characters), reflecting the genre's emphasis on storytelling and social commentary. In contrast, jazz showed the most concise lyrical content with only 328 average words and 1697 characters, aligning with the genre's focus on musical improvisation over verbal expression.

The distribution analysis showed that most songs cluster around the dataset mean of 665 words, with a normal distribution pattern. However, notable outliers were identified, including songs with extremely short lyrics (<50 words) and verbose tracks exceeding 1500 words. These outliers were primarily found in hip-hop and alternative genres, suggesting genre-specific approaches to lyrical density.

Genre-wise analysis revealed that electronic and dance music also showed relatively short lyrics (571 and 559 words respectively), which aligns with these genres' emphasis on rhythm and melody over narrative content. Country and rock genres showed moderate verbosity, while soul and R&B demonstrated higher lyrical content, reflecting their tradition of emotional storytelling.

4.2 Cluster Interpretation Framework

Visualization Methodology

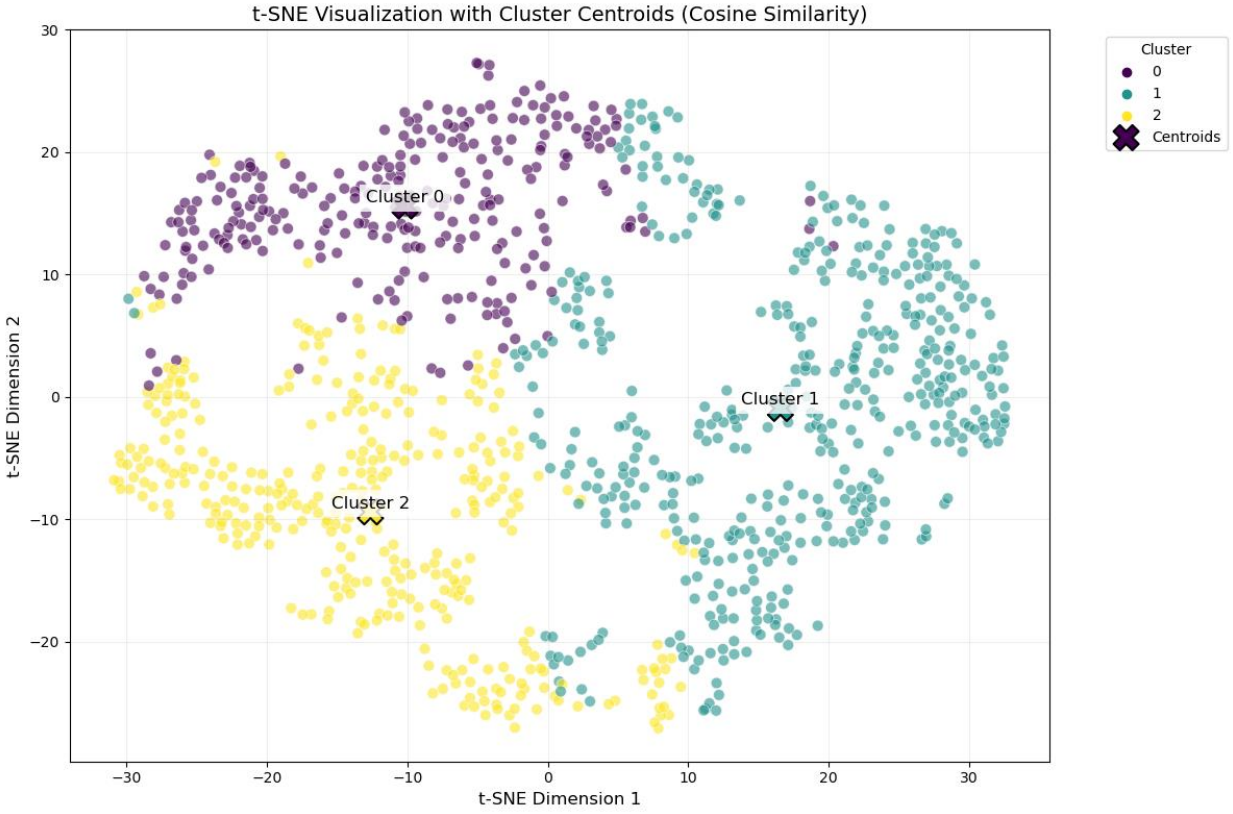

Figure 3.2 -- t-SNE visualization of clustered audio features

Figure 3.2 demonstrates clear separation between the three clusters in reduced dimensionality:

- **Cluster 0**: Concentrated in the lower-left quadrant
- **Cluster 1**: Dominates the upper-right region
- **Cluster 2**: Occupies a distinct central band

Cluster centroids are well-spaced, supporting the validity of the K-means partitioning. Minimal overlap between clusters suggests effective feature-based separation.

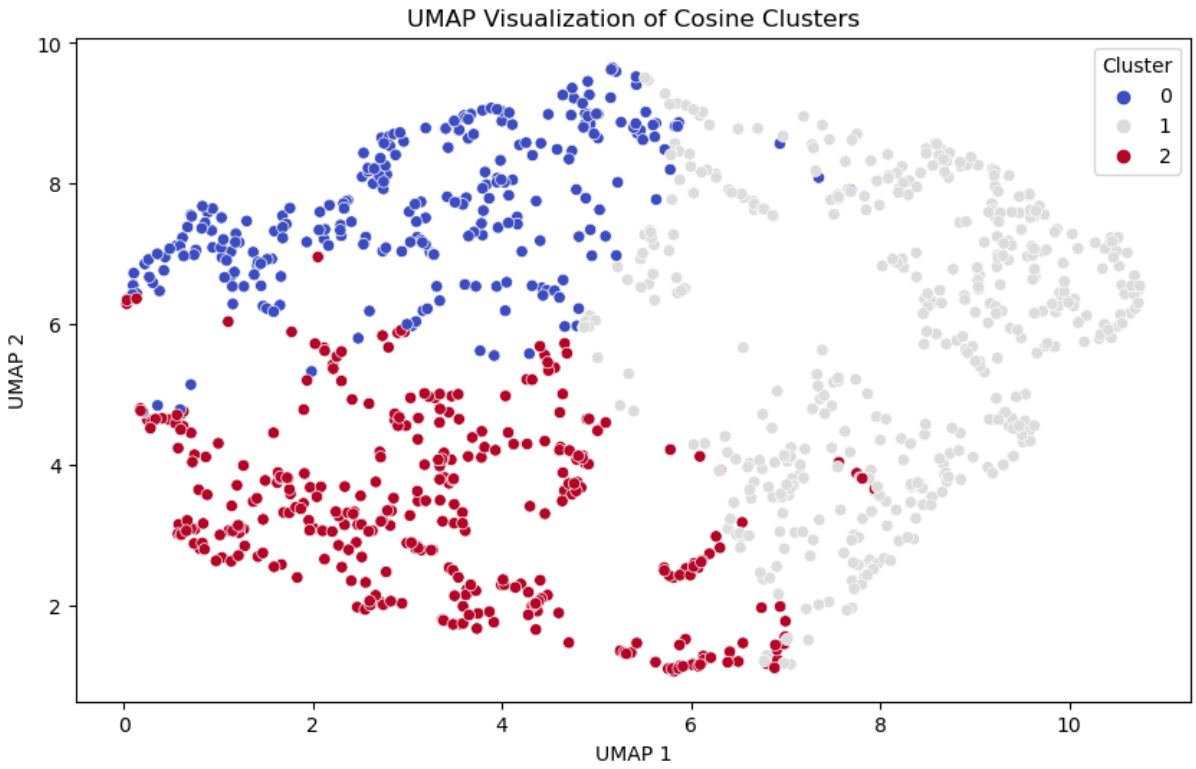

Figure 3.3 -- UMAP projection of clustered audio features

Figure 3.3 shows cluster distinctness with alternative dimensionality reduction:

- Cluster 0: Forms a tight grouping in the negative UMAP1 range
- Cluster 1: Dispersed along UMAP2
- Cluster 2: Separates cleanly along both UMAP axes

Both visualizations shows:

- Internal Cohesion: Points within clusters group tightly
- External Separation: Minimal inter-cluster overlap

4.3 EMOTION ASSIGNING

The assignment of emotional labels to the three identified clusters was conducted through a systematic process that integrated cluster feature profiles with established theoretical frameworks from music psychology research. This multi-step approach ensured that the emotional classifications were grounded in empirical evidence rather than subjective interpretation.

Cluster-Specific Emotion Assignment

Cluster 0: Aggressive/Intense/Angry : Cluster 0 was assigned the "Aggressive/Intense/Angry" label based on its distinctive feature combination of low valence (0.345), high energy (0.801), low acousticness (0.057), and high loudness (-4.87 dB). This profile directly corresponds to the aggressive emotion pattern identified by Zentner et al. (2008), who characterized angry music as exhibiting high arousal (energy) combined with negative valence. The low acousticness value supports this interpretation, as aggressive music typically features synthetic or heavily processed sounds rather than organic instrumentation (Yang et al., 2008). This combination places Cluster 0 in the high-arousal, negative-valence quadrant of Russell's (1980) circumplex model of affect, which is consistently associated with anger and aggression in music perception studies.

Cluster 1: Sad/Calm/Melancholic: This label was determined for Cluster 1 by its characteristic low energy (0.422), high acousticness (0.520), low valence (0.403), and quiet loudness (-9.70 dB). This feature profile aligns precisely with acoustic markers of melancholic music and the sad music characteristics documented by Eerola & Vuoskoski (2013). The high acousticness value is particularly indicative, as research consistently shows that sad music relies heavily on organic instrumentation such as piano, strings, and unprocessed vocals to enhance emotional intimacy . The combination of low energy and moderate-low valence corresponds to the low-arousal, negative-valence region associated with sadness and contemplation in affective music research.

Cluster 2: Happy/Upbeat: Cluster 2 received the "Happy/Upbeat" designation based on its high valence (0.669), high danceability (0.713), moderate-high energy (0.741), and moderate loudness (-5.41 dB). This feature configuration matches the positive emotion profile established by Juslin & Laukka (2004) and the happy music characteristics identified in multiple cross-cultural studies (Eerola & Vuoskoski, 2013). The high danceability score is particularly significant, as it reflects the rhythmic engagement and motor activation associated

with positive emotional states in music (Soleymani et al., 2013). This cluster occupies the high-valence, moderate-to-high arousal space that consistently correlates with happiness, joy, and optimism in music emotion research.

Lyrics-Based Emotion Assignment

The classification methodology produces the following categorical distribution across the dataset:

- **Category 1 (Sad/Calm/Melancholic)**: 478 songs (44.4%)
- **Category 2 (Happy/Upbeat)**: 330 songs (30.7%)
- **Category 0 (Aggressive/Intense/Angry):** 268 songs (24.9%)

4.4 Emotion Alignment Analysis

A thorough comparative analysis was carried out using both statistical techniques and visual exploration methods in order to assess the degree of alignment between the emotional content of the audio and lyrics.

Alignment Assessment

A heatmap was generated to visualize the co-occurrence of predicted emotions from lyrics and those derived from audio analysis. This matrix representation enabled the identification of patterns of agreement or divergence between the two modalities, highlighting frequent matches or mismatches in emotional interpretation.

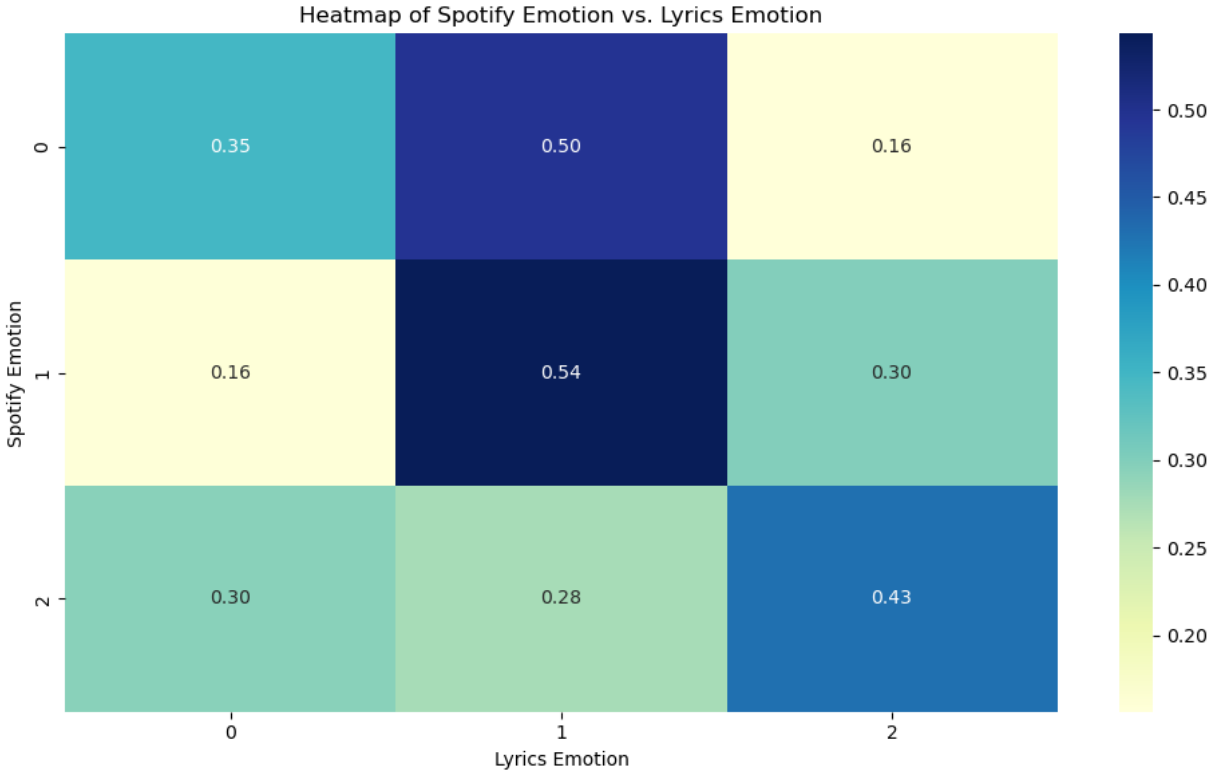

Figure 3.4 -- Emotion alignment heatmap between audio and lyrics

Figure 3.4 shows the emotion alignment heatmap analysis results, revealing distinct agreement patterns across emotional categories. The analysis demonstrates that Category 1 (Sad/Melancholic) achieved the highest perfect alignment at 54%, while Category 2 (Happy/Upbeat) showed 43% perfect alignment, and Category 0 (Angry/Aggressive) exhibited the lowest alignment at 35%. The cross-modal patterns indicate significant emotional divergence, with Audio Category 0 frequently paired with Lyrics Category 1 (50% co-occurrence), Audio Category 1 with Lyrics Category 2 (30% co-occurrence), and Audio Category 2 with Lyrics Category 1 (28% co-occurrence), suggesting complex emotional interplays between musical and lyrical content.

Agreement Rate Calculation

An agreement rate was computed by measuring how often the emotion detected from audio aligned with the one classified from lyrics. Each instance was marked as a "match" or "mismatch" and aggregated to provide a general agreement percentage across the dataset.

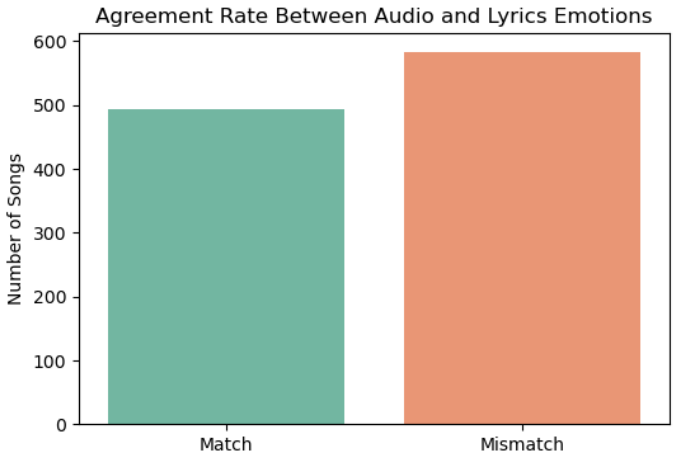

Figure 3.5 -- Agreement vs. disagreement distribution pie chart

Figure 3.5 reveals that mismatches outnumbered matches, with approximately 490 songs showing emotional agreement between audio and lyrics, while approximately 586 songs exhibited emotional disagreement between modalities.

Chi-Square Test Results

- Chi-Square statistic: 99.45
- p-value: 0.0000
- Decision: There IS a significant relationship (Reject H_o)

Cohen's Kappa Results

- Cohen's Kappa: 0.164
- Agreement Level: Slight agreement

Genre-Specific Mismatch Results

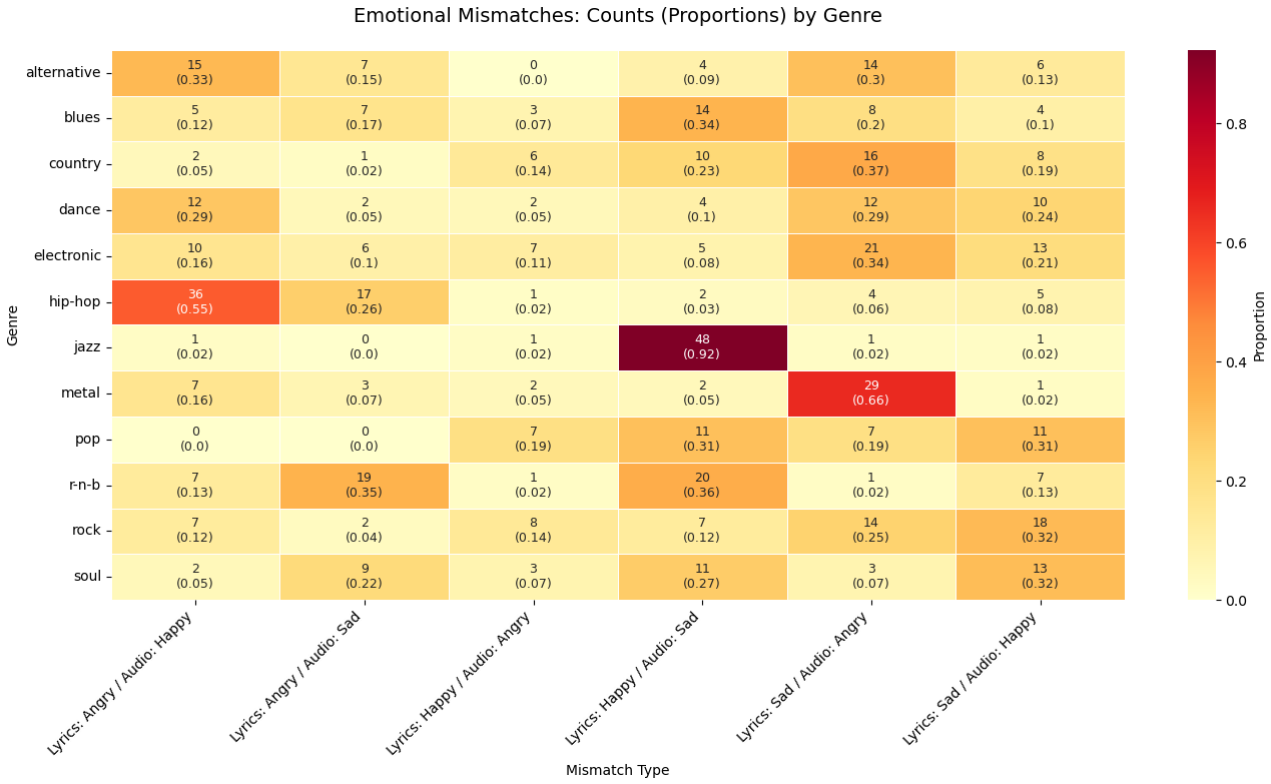

Figure 3.6 -- Genre-specific mismatch patterns heatmap

Figure 3.6 reveals distinct genre-specific patterns of emotional misalignment that reflect underlying artistic and cultural conventions. Hip-hop demonstrates the strongest preference for contrasting angry lyrics with happy audio (55% of mismatches), followed by sad lyrics with happy audio (26%), suggesting the genre's sophisticated use of upbeat production to deliver serious social commentary. Jazz exhibits the most consistent mismatch pattern, with an overwhelming 92% of misaligned songs featuring happy lyrics over sad audio, reflecting the genre's tradition of finding optimism within melancholic musical frameworks. Metal shows a pronounced tendency toward sad lyrics paired with angry audio (66% of mismatches), indicating the genre's use of aggressive instrumentation to amplify underlying emotional distress rather than contradict it.

Among other genres, R&B displays relatively balanced emotional contrasts between angry lyrics with sad audio (35%) and happy lyrics with sad audio (36%), while country music favors sad lyrics with angry audio (37%) alongside happy lyrics with sad audio (23%). Electronic music shows the most distributed mismatch pattern, with sad lyrics and angry audio representing only 34% of misalignments and other combinations accounting for half of all cases. Similarly, dance, alternative, pop, rock, and soul genres demonstrate more varied emotional contrast strategies, with no single mismatch type exceeding 35% prevalence. Blues stands out for its relatively balanced approach, with happy lyrics over sad audio comprising just 34% of mismatches. These patterns suggest that while emotional misalignment is common across all genres, specific combinations reflect deliberate artistic choices rooted in genre conventions and cultural expectations.

4.5 Emotional Alignment Patterns and Statistical Relationships

The Prevalence of Emotional Misalignment

The most striking finding of this study is the predominance of emotional misalignment between audio and lyrical content. With only 46% of songs showing emotional agreement between modalities, the majority (54%) of tracks exhibit some form of emotional contrast or dissonance. This finding challenges the common assumption that music is emotionally unified across its constituent elements and suggests that emotional complexity—rather than simplicity—is the norm in contemporary music.

The heatmap analysis revealed that sad emotions showed the highest alignment rate at 54%, indicating that when music sounds melancholic, it is most likely to have correspondingly sad lyrics. This suggests that sadness may be the most "honest" emotion in music, where artists tend to align musical and lyrical expression rather than create contrast. Conversely, happy-sounding music showed the most emotional confusion, with lyrics distributed relatively evenly across all three emotional categories (30% angry, 28% sad, 43% happy). This pattern indicates that upbeat musical arrangements are frequently used as vehicles for non-positive lyrical content, creating what might be termed "deceptive happiness" in musical communication.

Statistical Significance and Practical Agreement

The statistical analysis revealed a nuanced relationship between audio and lyrical emotions that requires careful interpretation. The Chi-square test of independence yielded a statistic of 99.45 with $p < 0.001$, definitively establishing that audio and lyrical emotions are not

independent variables. This confirms that when songs exhibit certain emotional characteristics in their audio features, there are predictable patterns in their lyrical emotional content.

However, Cohen's Kappa coefficient of 0.164 indicates only "slight agreement" between the two modalities according to conventional interpretation scales. This creates an apparent paradox between statistical significance and practical agreement that illuminates the complex nature of emotional expression in music.

Resolving the Statistical Paradox

These seemingly contradictory results reflect different aspects of the same phenomenon. The Chi-square test demonstrates that audio and lyrical emotions are systematically related—their co-occurrence patterns are not random. Cohen's Kappa reveals that this systematic relationship manifests primarily through predictable disagreement rather than alignment.

In practical terms, this means that artists make deliberate choices about emotional combinations between audio and lyrics, but these choices predominantly favor emotional contrast over consonance. For example, hip-hop consistently pairs aggressive lyrics with upbeat audio (systematic relationship), while rarely matching angry lyrics with angry audio (low agreement).

4.6 Genre-Specific Patterns of Emotional Misalignment

Hip-Hop: The Master of Emotional Contrast

Hip-hop emerged as the genre with the highest proportion of emotional mismatches, particularly exhibiting a strong tendency toward "angry lyrics with happy audio" (55% of mismatches). This pattern reflects the genre's sophisticated approach to emotional expression, where aggressive or confrontational lyrical content is often delivered over rhythmically engaging, sometimes celebratory instrumental tracks. This contrast serves multiple functions: it makes potentially harsh social commentary more palatable to mainstream audiences, creates ironic tension that amplifies the message's impact, and maintains the genre's fundamental connection to dance and rhythm despite serious subject matter.

The prevalence of this pattern in hip-hop aligns with the genre's historical role as both entertainment and social commentary. Artists like Eminem exemplify this approach in tracks

like "The Real Slim Shady," where aggressive critiques of fame and society are delivered over cheerful, radio-friendly production. This emotional dissonance amplifies the satirical critique while ensuring commercial viability.

Jazz: Bittersweet Sophistication

Jazz demonstrated an overwhelming tendency toward "happy lyrics with sad audio" (92% of mismatches), representing the most consistent mismatch pattern observed across any genre. This pattern reflects jazz's sophisticated emotional palette and its cultural tradition of finding hope or optimism within melancholic musical frameworks. The genre's emphasis on improvisation and emotional nuance allows for complex layering of sentiment that can simultaneously express joy and sorrow.

Louis Armstrong's "What a Wonderful World" exemplifies this pattern perfectly, combining overtly optimistic lyrics celebrating nature's beauty and human connection with melancholic musical arrangements featuring slow tempos and wistful delivery. This creates a bittersweet emotional experience that transforms apparent joy into something more complex—a nostalgic lament for an idealized world that may exist only in memory or aspiration.

Metal: Intensity Through Emotional Convergence

Metal showed a strong preference for "sad lyrics with angry audio" (66% of mismatches), suggesting that the genre uses aggressive musical intensity to express or channel underlying sadness and despair. This pattern reflects metal's therapeutic function for many listeners, where aggressive sound provides an outlet for internalized pain. Unlike other genres that use emotional contrast for artistic effect, metal appears to use sonic aggression as an amplification mechanism for emotional distress rather than a contradiction of it.

The Prevalence of Emotional Misalignment

Pop music demonstrated remarkable emotional versatility, with roughly equal proportions of "happy lyrics with sad audio" (31%) and "sad lyrics with happy audio" (31%). This balanced distribution of emotional contrasts reflects pop's commercial imperative to appeal to diverse audiences and emotional states simultaneously. Pop artists often use bittersweet combinations to create songs that can function in multiple contexts—upbeat enough for radio play yet emotionally resonant enough for personal connection.

4.7 Common Emotional Mismatch Patterns and Their Implications

The "Trojan Horse" Effect: Happy Audio with Non-Happy Lyrics

The prevalence of upbeat musical arrangements carrying sad or angry lyrical content (235 songs combined) represents what could be termed the "Trojan Horse" effect in popular music. This pattern allows artists to deliver potentially challenging emotional content within accessible, commercially viable packaging. The catchy, dance-friendly musical surface acts as a delivery mechanism for deeper, more complex emotional messages that might otherwise be rejected by mainstream audiences.

This technique has proven particularly effective in addressing serious social issues, personal trauma, or existential themes while maintaining broad appeal. The contrast between musical and lyrical content creates a form of cognitive dissonance that can make the emotional message more memorable and impactful once listeners engage with the lyrics more deeply.

Amplification Through Opposition: Sad Lyrics with Angry Audio

The combination of sad lyrics with aggressive musical backing (130 songs) represents an amplification strategy where intense musical energy serves to heighten the emotional impact of melancholic content. This pattern is particularly common in genres like emo, post-hardcore, and certain subgenres of rap, where the aggressive musical framework provides an outlet for emotional pain that might otherwise feel too subdued or internal.

Ironic Commentary: Angry Lyrics with Happy Audio

The pairing of aggressive lyrical content with upbeat musical arrangements (104 songs) often serves an ironic or satirical function, using the contrast to highlight societal contradictions or personal hypocrisies. This approach can make critical commentary more palatable while simultaneously emphasizing the disconnect between surface appearances and underlying realities.

4.8 Limitations and Considerations

Methodological Limitations

The absence of systematic validation for GPT-4 emotion classifications represents the most significant methodological limitation. While the model demonstrates sophisticated language understanding capabilities, the accuracy of lyrical emotion assignments remains unverified against human annotators or established emotion lexicons, affecting the credibility of comparative analysis. The prompt engineering approach may have subtly influenced classification decisions by framing a limited emotional spectrum, and GPT-4's lack of explicit reasoning makes it difficult to trace decision paths or understand classification factors.

The reliance on Spotify's proprietary audio features, while providing access to sophisticated analysis, limits transparency in understanding how emotional indicators are calculated. These algorithmic estimates may differ from psychological conceptions of emotion and represent computational rather than perceptual measures. Additionally, the K-means clustering approach assumes spherical cluster shapes and may not capture complex emotional boundaries that exist in musical expression.

Cultural and Linguistic Considerations

This study focused primarily on English-language popular music distributed through major streaming platforms, limiting generalizability to other cultural contexts, languages, or musical traditions. Different cultures may have distinct conventions for emotional expression in music, and non-Western musical systems may exhibit different patterns of emotional alignment or use different conventions for expressing emotional contrast. The dataset's emphasis on contemporary Western popular music may not represent traditional, folk, or culturally-specific musical forms that could reveal alternative patterns of emotional expression.

Temporal and Contextual Factors

The research provides a snapshot analysis that does not account for how emotional perception of music changes over time, across different listening contexts, or as cultural meanings evolve. Songs may be reinterpreted by listeners as societal contexts shift or personal circumstances change, potentially altering perceived emotional alignment or contrast. The study's focus on studio recordings excludes live performance dynamics, audience interaction effects, and temporal changes in artistic expression that might influence emotional interpretation.

Individual Differences in Emotion Perception

While this study employed objective classification methods, the subjective nature of emotional experience means individual listeners may perceive different emotions based on personal history, cultural background, current mood, or other contextual factors. The classifications represent general tendencies rather than universal emotional truths about musical content. Musical training, age, cultural exposure, and psychological factors could significantly influence how individuals perceive emotional alignment or contrast between audio and lyrical modalities.

Technical Limitations

The genre classification system, while comprehensive, relies on streaming platform categorizations that may not reflect artistic intent or musical complexity. Some songs could belong to multiple genres or represent hybrid forms that don't fit traditional categorical boundaries. The three-emotion classification framework, while theoretically grounded, may oversimplify the emotional complexity present in musical expression. Additionally, the study's focus on complete songs doesn't account for temporal changes in emotion throughout individual tracks or the effect of song structure on emotional perception.

5. CONCLUSIONS AND FUTURE WORKS

5.1 SUMMARY OF KEY FINDINGS

This research investigated the emotional alignment between audio features and lyrical sentiment in contemporary popular music, analyzing 1,080 songs across 12 musical genres. The study employed a novel dual-modality approach, combining unsupervised clustering of Spotify audio features with GPT-4-based lyrical emotion classification to examine how emotions are expressed and potentially contrasted across different channels of musical communication.

Principal Discoveries

The most significant finding of this research is that emotional misalignment between audio and lyrics is not the exception but the norm in contemporary popular music. With only 46% of songs showing emotional agreement between modalities, the majority of tracks (54%) exhibit some form of emotional contrast or complexity. This challenges conventional assumptions about unified emotional expression in music and suggests that emotional sophistication, rather than simplicity, characterizes modern musical composition.

The statistical analysis revealed a paradoxical relationship: while the Chi-square test confirmed a significant relationship between audio and lyrical emotions ($p < 0.001$), Cohen's Kappa coefficient indicated only slight practical agreement ($\kappa = 0.164$). This apparent contradiction illuminates that emotions in music are systematically related yet frequently divergent in specific implementation, reflecting intentional artistic choices rather than random variation.

Genre-specific analysis uncovered distinct patterns of emotional expression. Hip-hop demonstrated the highest propensity for emotional contrast (particularly angry lyrics with happy audio), jazz overwhelmingly favored happy lyrics with sad audio, metal predominantly combined sad lyrics with angry audio, and pop showed balanced versatility in emotional combinations. These patterns suggest that genres have developed specific conventions for managing emotional complexity that serve both artistic and commercial functions.

Methodological Contributions

This study demonstrated the effectiveness of combining traditional machine learning approaches with modern large language models for multi-modal emotion analysis. The K-means clustering of audio features achieved robust emotional categorization (silhouette score = 0.4703), while GPT-4 proved capable of nuanced lyrical emotion interpretation that captured metaphorical, cultural, and contextual elements beyond simple sentiment analysis. The research also validated the use of cosine distance in audio feature clustering, which proved more effective than Euclidean distance for capturing spectral similarity patterns in musical emotion. The successful integration of Spotify's proprietary audio features with natural language processing techniques provides a replicable framework for future multi-modal music analysis.

5.2 Future Research Directions

Cross-Cultural and Multilingual Studies

Future research should extend this methodology to diverse cultural and linguistic contexts to understand whether patterns of emotional alignment and contrast are universal or culturally specific. Comparative studies across different musical traditions could reveal fundamental versus culture-specific aspects of emotional expression in music.

Specific research questions include:

- Do non-Western musical traditions exhibit similar rates of emotional contrast?
- How do different languages and poetic traditions affect lyrical emotion classification?
- Are there universal patterns of emotional expression that transcend cultural boundaries?

Longitudinal and Temporal Analysis

Research examining how emotional alignment patterns change over time could offer valuable insights into the evolution of musical expression and shifts in cultural emotional communication. This could involve a historical analysis of emotional patterns across different decades, revealing how societal values and artistic trends have shaped music's emotional

tone. Additionally, real-time analysis of listener emotional perception throughout a song could uncover dynamic changes in engagement and sentiment. Furthermore, investigating how significant cultural events influence emotional expression in contemporary music may highlight the interplay between collective experiences and artistic responses.

Individual Differences and Personalization

Future studies should explore how individual characteristics influence the perception of emotional alignment in music, potentially paving the way for personalized emotion models. This research could include analyzing how personality traits affect emotional perception, examining the impact of musical training on sensitivity to emotional contrast, and developing individualized emotion recognition systems tailored to specific listener profiles. Such investigations would deepen our understanding of how personal factors shape emotional experiences in music.

Enhanced Multi-Modal Analysis

Expanding the analysis of musical emotion beyond audio and lyrics to include additional modalities could offer a more comprehensive understanding of emotional expression in music. This could involve the integration of music video visual content for tri-modal emotion analysis, allowing researchers to examine how visuals reinforce or contrast with audio and lyrical emotions. Additionally, analyzing live performance elements such as stage presence and visual aesthetics can reveal how performers convey emotion through physical expression. Investigating the influence of album artwork and song titles on emotional interpretation may further uncover the subtle ways in which visual and textual elements shape the listener's emotional experience.

Real-Time and Interactive Applications

Developing systems capable of analyzing emotional alignment in real time could open the door to a range of innovative applications. For instance, live music analysis could assist DJs in mixing tracks or generating playlists that maintain a consistent emotional flow. In therapeutic settings, real-time emotional assessment based on biometric feedback could guide music selection to support emotional regulation or mental health goals. Additionally, interactive music generation systems that respond to users' emotional states could create dynamic and personalized musical experiences, enhancing engagement and emotional resonance.

Deeper Genre and Subgenre Analysis

A more granular analysis of musical genres could uncover nuanced emotional patterns that might be overlooked in broader categorizations. This could involve investigating emotional trends within specific subgenres, such as various styles of hip-hop or electronic music, to identify distinct emotional signatures. Additionally, examining how genre evolution influences emotional expression can shed light on the shifting emotional landscapes within musical communities over time. Analyzing cross-genre fusion may further reveal how emotional conventions blend, offering insight into the dynamic interplay between tradition and innovation in musical emotion.

5.3 CONCLUSIONS

This research demonstrates that emotional alignment in popular music is far more complex than previously understood, with emotional misalignment being the norm rather than the exception. Contemporary music is characterized by sophisticated patterns of emotional contrast that vary systematically across genres and serve specific artistic and commercial functions, challenging existing approaches to music emotion recognition and recommendation.

As artificial intelligence increasingly intersects with creative and therapeutic applications, understanding these human patterns of emotional expression becomes crucial for developing systems that can meaningfully engage with the full complexity of musical emotion. This research provides both methodological frameworks and empirical insights that can inform more sophisticated approaches to music analysis, recommendation, and generation, ensuring that digital technologies can appreciate the full richness of human emotional expression rather than reducing it to simplified categories.

BIBLIOGRAPHICAL REFERENCES

Avdeeff, M. (2014). Song Means: Analysing and Interpreting Recorded Popular Song. *IASPM@Journal*, *4*(2), 117–119. https://doi.org/10.5429/ij.v4i2.659

Balkwill, L.-L., & Thompson, W. F. (1999). A Cross-Cultural Investigation of the Perception of Emotion in Music: Psychophysical and Cultural Cues. *Music Perception*, *17*(1), 43–64. https://doi.org/10.2307/40285811

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11818–11823. https://doi.org/10.1073/pnas.191355898

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2005.14165

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., & Moussallam, M. (2018). *Music Mood Detection Based On Audio And Lyrics With Deep Neural Net* (No. arXiv:1809.07276). arXiv. https://doi.org/10.48550/arXiv.1809.07276

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Eerola, T., Himberg, T., Toiviainen, P., & Louhivuori, J. (2006). Perceived complexity of western and African folk melodies by western and African listeners. *Psychology of Music*, *34*(3), 337–371. https://doi.org/10.1177/0305735606064842

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49. https://doi.org/10.1177/0305735610362821

Eerola, T., & Vuoskoski, J. K. (2013). A Review of Music and Emotion Studies: Approaches, Emotion Models, and Stimuli. *Music Perception*, *30*(3), 307–340. https://doi.org/10.1525/mp.2012.30.3.307

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, *5*(1_suppl), 123–147. https://doi.org/10.1177/10298649020050S105

Grewe, O., Nagel, F., Kopiez, R., & Altenmüller, E. (2007). Listening To Music As A Re-Creative Process: Physiological, Psychological, And Psychoacoustical Correlates Of Chills And Strong Emotions. *Music Perception*, *24*(3), 297–314. https://doi.org/10.1525/mp.2007.24.3.297

Hu, X., & Downie, J. S. (2010). Improving mood classification in music digital libraries by combining lyrics and audio: 10th Annual Joint Conference on Digital Libraries, JCDL 2010. *JCDL'10 - Digital Libraries - 10 Years Past, 10 Years Forward, a 2020 Vision*, 159–168. https://doi.org/10.1145/1816123.1816146

Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 mirex audio mood classification task: 9th International Conference on Music Information Retrieval, ISMIR 2008. *ISMIR 2008 - 9th International Conference on Music Information Retrieval*, 462–467.

Hunter, P. G., Schellenberg, E. G., & Schimmack, U. (2010). Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed

emotions. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(1), 47–56. https://doi.org/10.1037/a0016873

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Ilie, G., & Thompson, W. F. (2006). A Comparison of Acoustic Cues in Music and Speech for Three Dimensions of Affect. *Music Perception*, *23*(4), 319–330. https://doi.org/10.1525/mp.2006.23.4.319

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*, *5*(5), 171274. https://doi.org/10.1098/rsos.171274

Janata, P., Tomic, S. T., & Haberman, J. M. (2012). Sensorimotor coupling in music and the psychology of the groove. *Journal of Experimental Psychology. General*, *141*(1), 54–75. https://doi.org/10.1037/a0024208

Juslin, P. N., & Laukka, P. (2004). Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research*. https://doi.org/10.1080/0929821042000317813

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(5), 559–575. https://doi.org/10.1017/s0140525x08005293

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., & Turnbull, D. (2010). *MUSIC EMOTION RECOGNITION: A STATE OF THE ART REVIEW*.

Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nature Reviews Neuroscience*, *15*(3), 170–180. https://doi.org/10.1038/nrn3666

Koelsch, S., Fritz, T., V Cramon, D. Y., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: An fMRI study. *Human Brain Mapping*, *27*(3), 239–250. https://doi.org/10.1002/hbm.20180

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale*, *51*(4), 336–353. https://doi.org/10.1037/1196-1961.51.4.336

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab Toolbox for Music Information Retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 261–268). Springer. https://doi.org/10.1007/978-3-540-78246-9_31

Laukka, P. (2007). Uses of music and psychological well-being among the elderly. *Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being*, *8*(2), 215–241. https://doi.org/10.1007/s10902-006-9024-3

Leman, M. (2007). *Embodied Music Cognition and Mediation Technology*. The MIT Press. https://doi.org/10.7551/mitpress/7476.001.0001

Leman, M., Moelants, D., Varewyck, M., Styns, F., van Noorden, L., & Martens, J.-P. (2013). Activating and relaxing music entrains the speed of beat synchronized walking. *PloS One*, *8*(7), e67932. https://doi.org/10.1371/journal.pone.0067932

Maes, P.-J. (2016). Sensorimotor Grounding of Musical Embodiment and the Role of Prediction: A Review. *Frontiers in Psychology*, *7*, 308. https://doi.org/10.3389/fpsyg.2016.00308

Mayer, R., Neumayer, R., & Rauber, A. (2008). Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In Proceedings of the 9th International Conference on Music Information Retrieval (pp. 337–342). http://hdl.handle.net/20.500.12708/52254

Mihalcea, R., & Strapparava, C. (2005). Making computers laugh: Investigations in automatic humor recognition. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 531–538. https://doi.org/10.3115/1220575.1220642

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mohammad, S. (2012). #Emotional Tweets. In E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, & D. Yuret (Eds.), SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 246–255). Association for Computational Linguistics. https://aclanthology.org/S12-1033/

Mohammad, S. M., & Mohammad, S. (n.d.). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *Sentiment Analysis*.

Mohammad, S. M., & Turney, P. D. (2013). CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Mullen, J. (2014). Philip Tagg, Music's Meanings: a modern musicology for non-musos, New York & Huddersfield: Mass Media Music Scholars' Press, 2012, 691 pages. Quaderna, 2, 281–284.

Pedregosa, F., Pedregosa, F., Varoquaux, G., Varoquaux, G., Org, N., Gramfort, A., Gramfort, A., Michel, V., Michel, V., Fr, L., Thirion, B., Thirion, B., Grisel, O., Grisel,

O., Blondel, M., Prettenhofer, P., Prettenhofer, P., Weiss, R., Dubourg, V., … Cournapeau, D. (n.d.). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037//0022-3514.77.6.1296

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. https://doi.org/10.3115/v1/d14-1162

Phillips-Silver, J., & Trainor, L. J. (2007). Hearing what the body feels: Auditory encoding of rhythmic movement. *Cognition*, *105*(3), 533–546. https://doi.org/10.1016/j.cognition.2006.11.006

Plutchik, R. (1980). Chapter 1—A general psychoevolutionary theory of emotion. In R.Plutchik & H. Kellerman (Eds.), Theories of Emotion (pp. 3–33). Academic Press. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Saarikallio, S. (2011). Music as emotional self-regulation throughout adulthood. *Psychology of Music*, *39*(3), 307–327. https://doi.org/10.1177/0305735610374894

Saarikallio, S., & Erkkilä, J. (2007). The role of music in adolescents' mood regulation. *Psychology of Music*, *35*(1), 88–109. https://doi.org/10.1177/0305735607068889

Schedl, M., Gómez, E., & Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. Foundations and Trends in Information Retrieval, 8(2–3), 127–261. https://doi.org/10.1561/1500000042

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., & Elahi, M. (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, *7*(2), 95–116. https://doi.org/10.1007/s13735-018-0154-2

Shutova, E., Sun, L., & Korhonen, A. (n.d.). *Metaphor Identification Using Verb and Noun Clustering*.

Sloboda, J. A., O'Neill, S. A., & Ivaldi, A. (2001). Functions of Music in Everyday Life: An Exploratory Study Using the Experience Sampling Method. *Musicae Scientiae*. https://doi.org/10.1177/102986490100500102

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., & Yang, Y.-H. (2013). 1000 songs for emotional analysis of music. *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, 1–6. https://doi.org/10.1145/2506364.2506365

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927x09351676

Toward Multi-modal Music Emotion Classification. (2008). In Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, & H. H. Chen, *Lecture Notes in Computer Science* (pp. 70–79). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-89796-5_8

Trost, W., Ethofer, T., Zentner, M., & Vuilleumier, P. (2012). Mapping Aesthetic Musical Emotions in the Brain. *Cerebral Cortex*, *22*(12), 2769–2783. https://doi.org/10.1093/cercor/bhr353

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Yang, Y.-H., & Chen, H. H. (2012). Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology*, *3*(3), 1–30. https://doi.org/10.1145/2168752.2168754

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2008). A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 448–457. https://doi.org/10.1109/tasl.2007.911513

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory-motor interactions in music perception and production. *Nature Reviews. Neuroscience*, *8*(7), 547–558. https://doi.org/10.1038/nrn2152

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion (Washington, D.C.)*, *8*(4), 494–521. https://doi.org/10.1037/1528-3542.8.4.494

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa