Final Year Project
Demo
Music Sentiment Analysis

Yuchen Zhu

Structure

- Background
- Significance
- Research
- Methodology
- Evaluation and Discussion

Background

Music's Historical and Cultural Impact

An ancient, universal art, integral to human history and cultural.
Most people listen to music throughout their lives.

Music and Emotional Expression

A key medium for expressing human emotions.

Music in Cultures and Societies

Plays a central role across diffrent cultural and social contexts.

Significance

Music's Emotional Influence

Music has a profound impact on human emotions.

Project Motivation

Focused on enhancing emotional well-being through music's influence on emotions and psychology.

Aim

To gain a deeper understanding of music's emotional classification by combining song lyrics and audio features.

Research

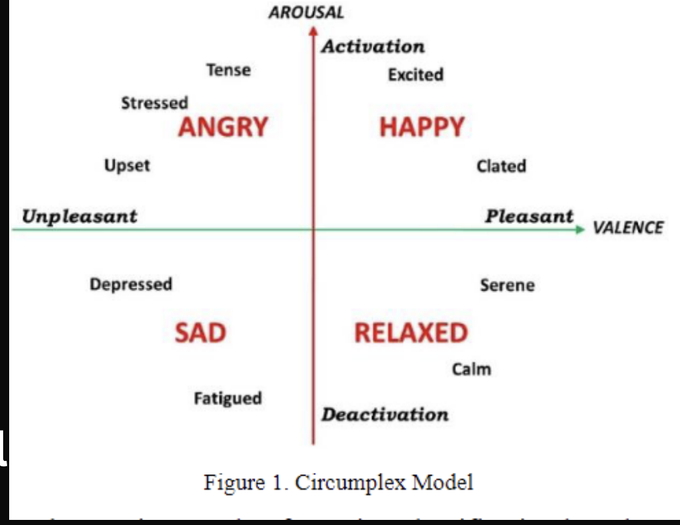


Figure 1. Circumplex Model

MER in Music Information Retrieval

Music emotion recognition (**MER**) is a research field in **music information retrieval**

Inspired by Russell's Model

Detailed emotion classification in MER, inspired by the **Russell emotion model**.

Comprehensive MER Model Development

Develop a comprehensive MER model that focuses on **lyrics and audio analysis**.

Benchmark: Jiddy Abdillah et al.'s Study

Study by Jiddy Abdillah et al. achieved **91.08%** accuracy using Bi-LSTM and GloVe.

Goal: Surpassing the Benchmark

Aimimg to outperform this benchmark in **generalization and accuracy of sentiment classification using my models.**

Methodology

Dataset

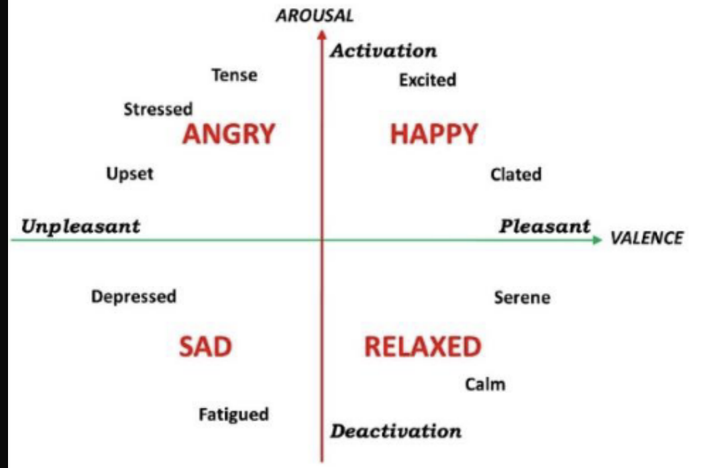


Figure 1. Circumplex Model

***Dataset1*: MoodyLyrics**: Contains 2595 songs labeled with one of the **4 categories of Russell's model** based on text(Labels **only from Lyrics**)

***Dataset2*: MoodyLyrics4Q**: Contains 2000 songs labeled with one of the **4 categories of Russell's model** based on Last.fm tags(Labels from **overall music tags (Lyrcis and Audio))**

Methodology

Dataset

Lyrics

LyricsData Acquisition and Optimization

- Initial Attempt: Genius API**(Lyrcis Collection Tools)**
  - Using lyricsgenius to obtain lyrics based on song names and artists.
  - Issue: Relies on exact match of song titles and artist names, which can result in obtaining **incorrect lyrics**.

Improved Method: Custom Web Scraper

- Using Google to parse HTML from the Genius website.
- Method: Locating HTML class names storing song titles and artist names.

Audio Feature

Audio Feature Extraction

- Using Spotify API**(Audio Feature Collection Tools)**.
  - Locating specific songs based on song names and artists.
  - Collecting audio features of songs.

Data Cleaning and Standardization:

Tool: Custom regular expressions.

Goal: Remove non-essential information (like “[Verse1]” tags) and non-English lyrics and error audio feature.

Methodology
Dataset

The **Dataset1** contains 2123 records

Happy:642

Relaxed:532

Angry:501

Sad:448

Dataset Structure

ML_Index	Artist	Title	Mood	Lyrics	Danceability	Energy	Key	Loudness	Mo
ML1	Usher	There Goes My Baby	Relaxed	"There goes my baby (Oooh, girl,	[Sample Value]	[Sample Value]	[Sample Value]	[Sample Value]	[Sa Va

***Downsampling for Balance*:** Applied downsampling techniques to Dataset 1 by randomly removing 90 "Happy" songs, using a specific random state to ensure the process is reproducible.

***Random Shuffling for Unbiased Training*:** Implemented random shuffling of the entire dataset before the training process to avoid the model learning any potential order in the data, using ramdon state.

Methodology

Dataset

The *Final Dataset1* contains 2033 records

Happy:554（27.2%）

Relaxed:532（26.2%）

Angry:501（24.6%）

Sad:448（22.2%）

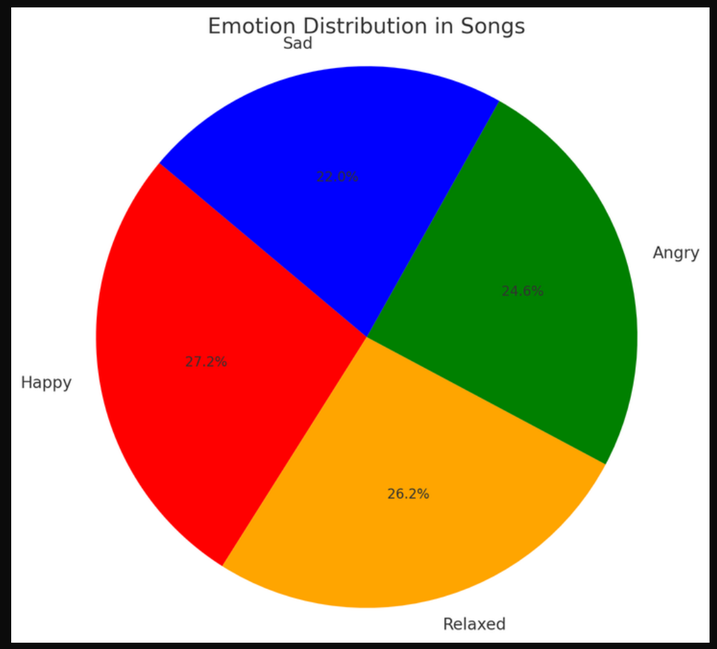
The *Final Dataset2* contains 1576 records

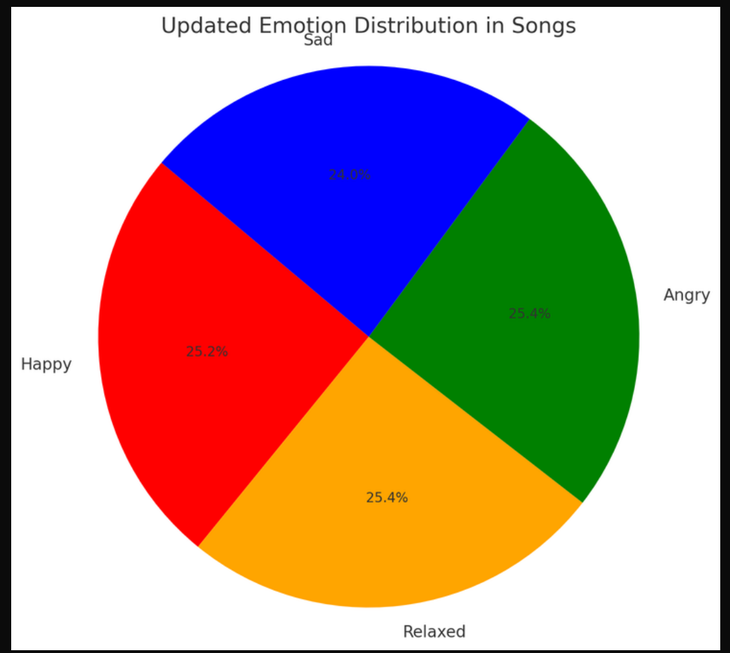
Happy:394（25.2%）

Relaxed:396（25.4%）

Angry:396（25.4%）

Sad:375（24%）





Methodology

Reproducing the paper(Bi-LSTM and GloVe)

1. ***Reproducing the Paper: Purpose and Process***
   - Objective: To verify the original research's **reliability** and **effectiveness**.
   - Benefits: Confirmed reproducibility, deepened understanding of methods and logic, **identified areas for improvement.**
2. ***Replication Methodology: Hyperparameters and Structure***
   - Used the **hyperparameters and structure specified by the paper**.
   - Used pretrained **GloVe 100**-dimensional vectors**(Global)** for word embedding.
3. ***Adjustments for Model Compatibility***
   - For Naive Bayes (NB), which doesn't accept negative values, used TF-IDF for word embedding.
   - Continued using GloVe for other models.
4. ***Successful Replication of Models***
   - Replicated various models: Naive Bayes: **(Probabilistic)**, K-Nearest Neighbors**(Distance-Based),** Support Vector Machine**(Optimal Boundary)**, Convolutional Neural Network**(Feature Extraction)**, Long Short-Term Memory Network**(Sequence Data)**, and Bidirectional Long Short-Term Memory Network**(Enhanced Contextual Understanding)**.
   - Achieved accuracy **similar** to the original paper for each model.

Paper Result

Table 4. Comparisons of the Different Methods

Method	Precision	Recall	F1-Score	Accuracy
Naïve Bayes	87%	81%	82%	83%
KNN	75%	74%	74%	76%
SVM	69%	68%	68%	71%
CNN	89%	89%	89%	90%
LSTM	90%	91%	90%	90%
Bi-LSTM	92%	90%	91%	91%

Reproducing Paper Result

Model	MARKAC	MINEAC	MARKF1	MINEF1
NB + tfidf	83	**82**	82	**82**
KNN + glove	76	**71**	74	**70**
SVM + glove	71	**78**	68	**78**
CNN + glove	90	**85**	89	**84**
LSTM + glove	90	**88**	90	**88**
BILSTM + glove	91	**88**	91	**88**

Methodology

Experiment Design

Main objective of experimental design

1.Word Embedding
Explore and apply various word embedding methods to improve model performance.

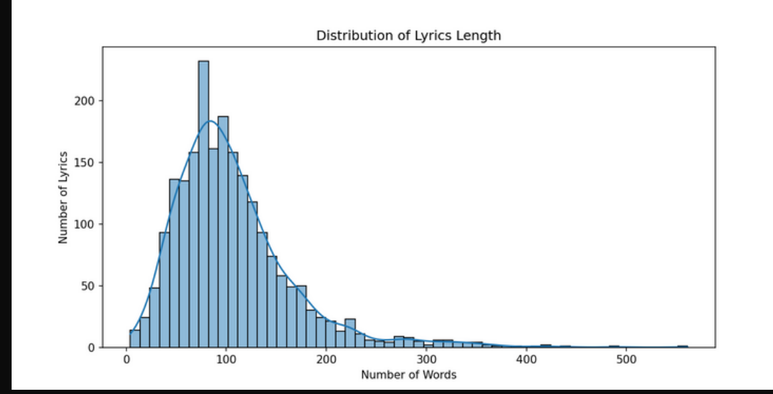
2.Preprocessing
Implement efficient data preprocessing strategies to optimize model inputs.

3.Audio Features
Integrate audio features to improve the model's ability to recognize emotions.

Methodology

Word Embedding

Sequence Length



1.Sequence Length Optimization

Reduced max sequence length from **1000(Paper) to 250** to minimize padding noise impact.
Most of the lyrics are **repetitive**, **rhythmic** texts**(Words and Local).**

2.*Word Embedding Approaches*

Applied Bag of Words, TFIDF**(Words)**, and Word2Vec300d(**Local**) with same **preprocessing** steps**(Lemma+LC+NR+SR)**.

3.*Model Tuning Strategies*

Fine-tuned parameters for Naive Bayes, SVM, and KNN; optimized deep learning models like Text-CNN, LSTM, Bi-LSTM for better performance.

4.*Embedding Technologies Performance*

BoW, TFIDF, and Word2Vec outperform GloVe and **exceed** the baseline accuracy.

Embedding Operation Result

Model + Method	Preprocessing Combination	Accuracy (ACC)	F1 Score
NB + BOW	Lemma + LC + NR + SR	92	92
NB + TFIDF	Lemma + LC + NR + SR	90	90
SVM + TFIDF	Lemma + LC + NR + SR	93	93
KNN + TFIDF	Lemma + LC + NR + SR	85	84
KNN + BOW	Lemma + LC + NR + SR	69	67
SVM + BOW	Lemma + LC + NR + SR	81	82
TEXTCNN + WORD2VEC	Lemma + LC + NR + SR	91	91
LSTM + WORD2VEC	Lemma + LC + NR + SR	89	89
BILSTM + WORD2VEC	Lemma + LC + NR + SR	89	89

Methodology

Preprocessing

1.<u>Four top-performing models in the embaddings:</u> Naive Bayes, SVM, Text-CNN, and BiLSTM.

2.<u>Preprocessing Strategies and Model Performance</u>
Evaluated the impact of Stemming(**Stem**), Lemmatization(**Lemma**), Noise Removal(**NR**), and Stopword(**SR**)

3.<u>Stability in Experimental Results</u>
Applied a **loop testing** method for ML and deep learning models to ensure result stability, accounting for r**andom weight initialization** for deep learning.

4.After embedding, tuning, and optimal preprocessing, my **SVM+TFIDF** achieves **94%** accuracy and F1 score, which **exceeds** the baseline accuracy

Preprocessing Operation

Operation
Stem
Stem + LC
Stem + NR
Stem + SR
Stem + LC + NR
Stem + LC + SR
Stem + NR + SR
Stem + LC + NR + SR

Operation
Lemma
Lemma + LC
Lemma + NR
Lemma + SR
Lemma + LC + NR
Lemma + LC + SR
Lemma + NR + SR
Lemma + LC + NR + SR

Operation
SR
SR + LC

Operation
NR
NR + LC
NR + SR

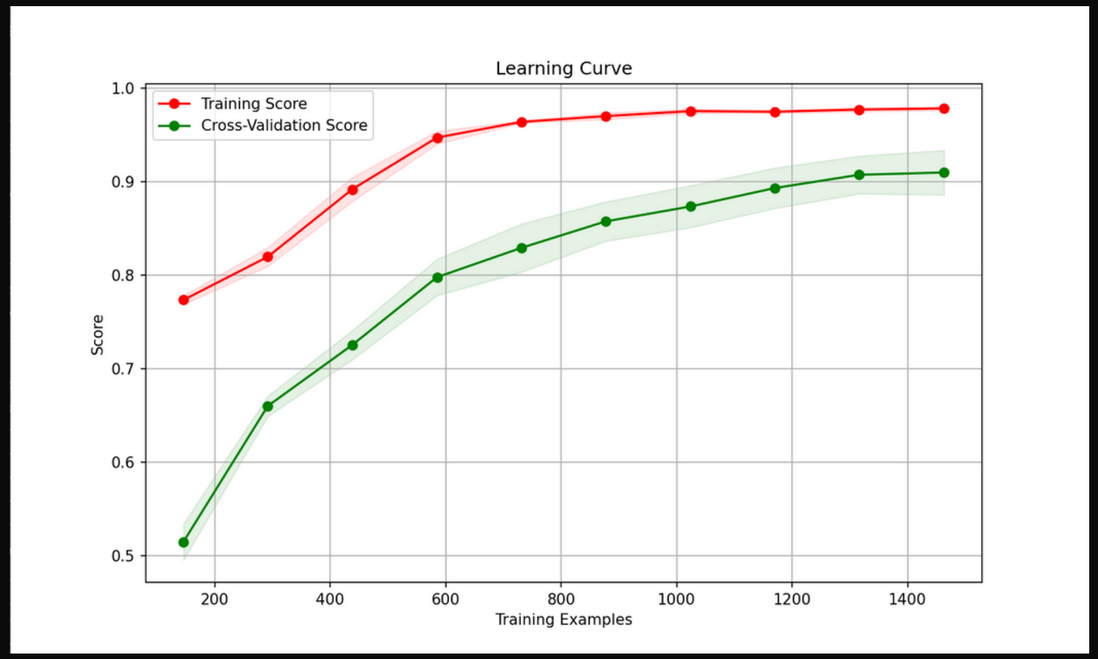
Best Preprocessing Model

Model + Method	Best Preprocessing Combination	Accuracy (ACC)	F1 Score
NB + BOW	Lemma + LC + NR + SR	92	92
SVM + TFIDF	LC + NR+ SR	94	94
TEXTCNN + WORD2VEC	LC + NR + SR	92	92
BILSTM + WORD2VEC	Lemma + NR+ SR	90	90

Methodology

Preprocessing

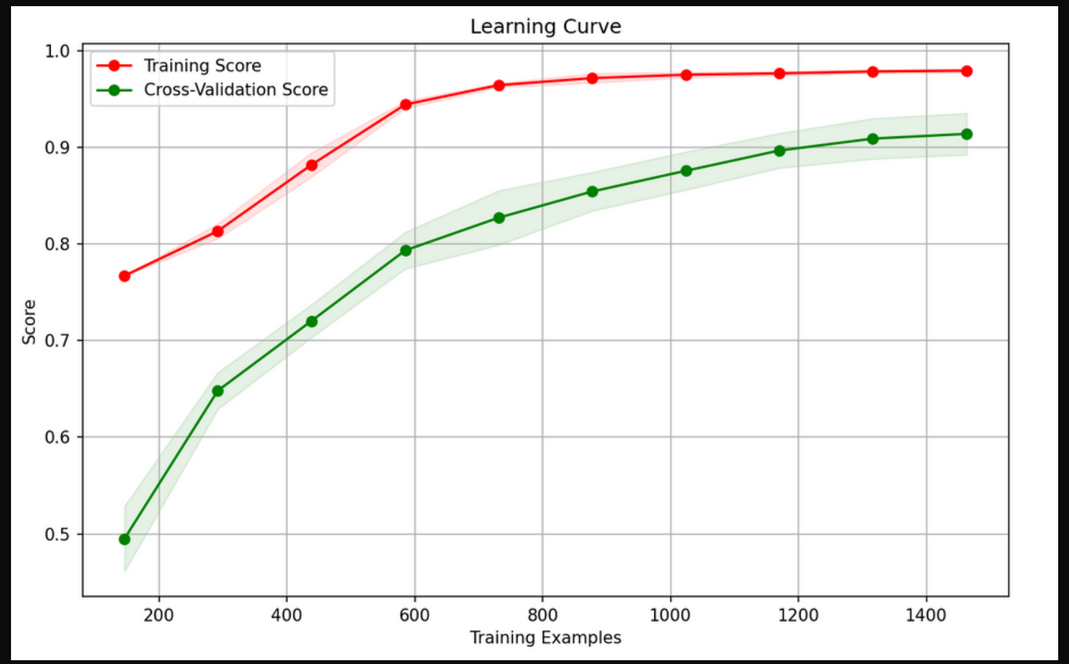
SVM **Before** best Preprocessing



F1:93.6%

CV:90.9%

SVM Best Preprocessing



F1:94.4%

CV:91.3%

Methodology

Audio Feature

1. ***Audio Feature Fusion and Model Performance***
   - Focused on the impact of **combine audio features** on model performance.
   - Normalized**(StandardScaler)** audio features for consistent data analysis and model training.
2. **A*udio Feature Analysis***
   - Utilized **Heat Maps, PCA, t-SNE**, and the **Hopkins statistic** to understand audio feature distribution and clustering.
   - Found mixed results in **clustering tendency** and **randomness overlap** in data points, indicating potential correlation loss in high-dimensional data.
3. ***Correlation Analysis of Audio Features***
   - **Heat maps** showed varied correlations with emotional categories, with some like **energy** and **loudness(Audio features)** in **"Angry"** showing strongest positive correlation, but others less significant.
4. ***Exclusion of Certain Audio Features***
   - Based on the **RF** and **heatmap** results, decided to exclude features such as **key**, **mode**, and **time signatures**, considering the limited contribution to sentiment classification and the **curse of dimensionality** problem.
5. ***Training with Audio-Only Features***
   - Only audio features are used for training and prediction, but the results are not satisfactory.

Feature	Description
Danceability	The degree of danceability of a song, indicating how suitable it is for dancing
Energy	The energy of a song, representing its activity level or intensity
Key	The key of a song, indicating its musical key or basic pitch
Loudness	The loudness of a song, representing its overall volume level in decibels (dB)
Mode	The mode of a song, indicating its scale type, usually Major or Minor
Speechiness	The presence of spoken words in a song, indicating the degree of speechiness
Acousticness	The acousticness of a song, indicating the presence of acoustic elements
Instrumentalness	The instrumentalness of a song, indicating the presence of instrumental elements
Liveness	The liveness of a song, indicating whether it is a live recording or a studio track
Valence	The valence of a song, representing its positive emotional intensity
Tempo	The tempo of a song, representing its speed in beats per minute (BPM)
Duration_ms	The duration of a song, representing its playback time in milliseconds
Time_signature	The time signature of a song, indicating the number of beats per bar and the beat type

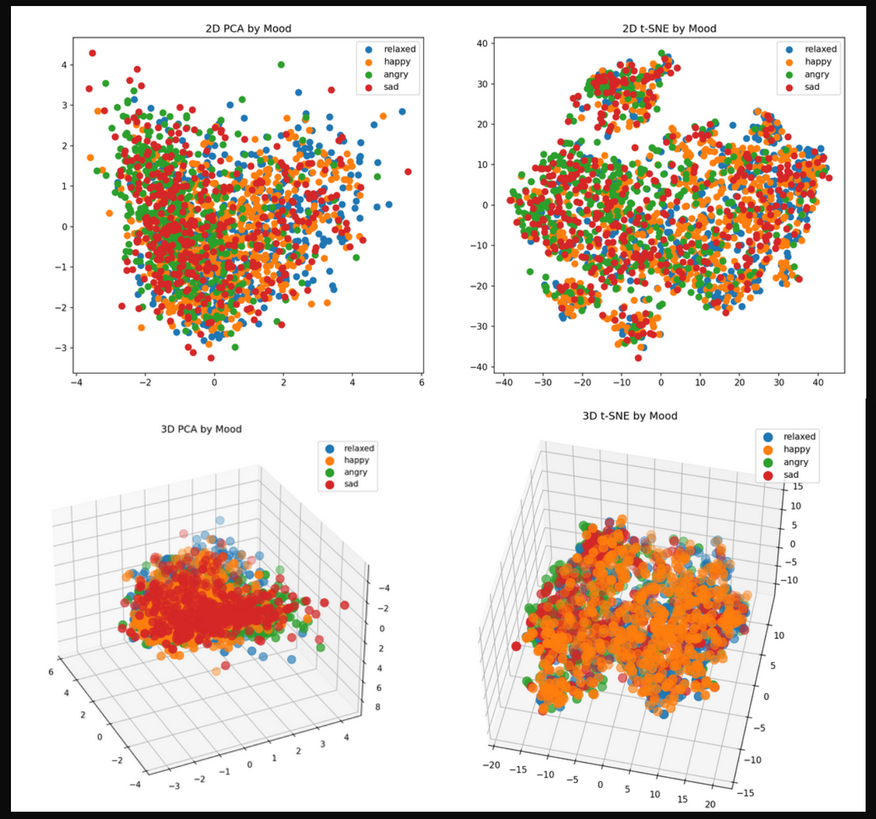
Audio Only Model Performance

Model	Accuracy (AC)	F1 Score
SVM	39%	36
Nb(MinMax)	40%	33 (sad F1=0)
XGBoost	40%	40
Desen	38%	37
RF	40%	40

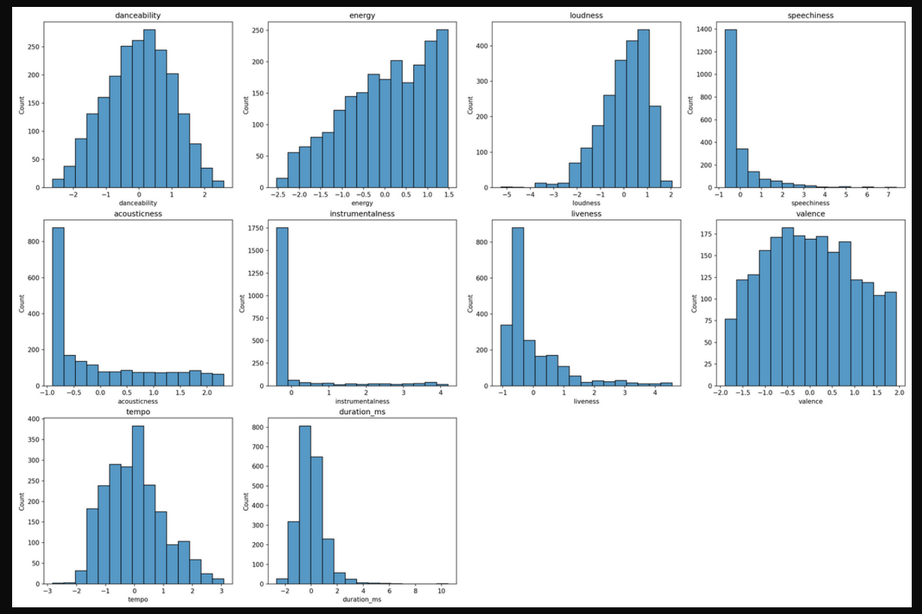
Methodology

Audio Feature

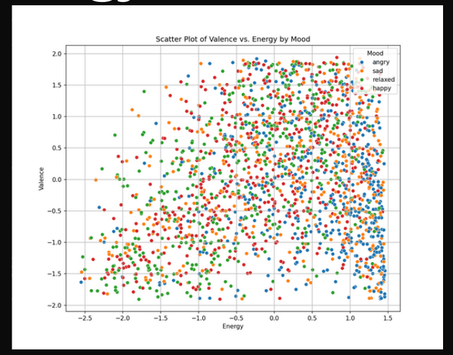
Dataset1 : PCA T-SNE



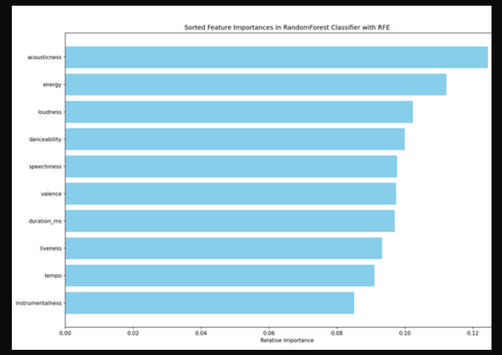
Dataset1 : Histogram data distribution



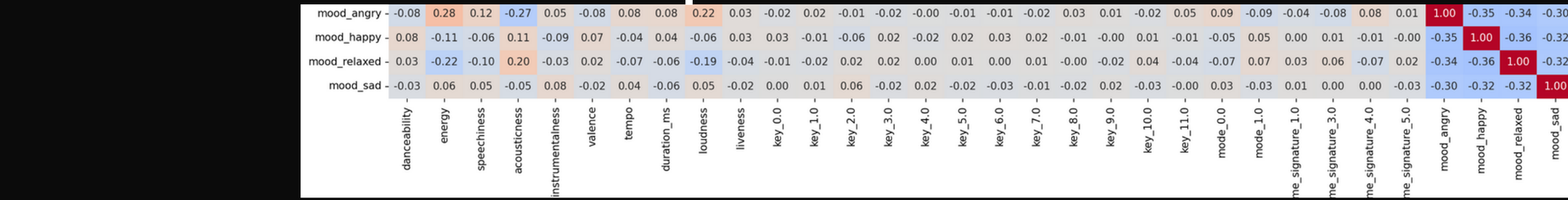
Dataset1 Energy VS Valence by mood



Dataset1 : RF



Dataset1 : Heat Map



Methodology

Audio Feature

6.***Integration of Audio Features into Models***

- Implemented feature **pre-fusion** with SVM and added **audio input layers** in Text-CNN and BiLSTM models.
- Explored **stacking ensemble learning(Robustness of the model)** with Random Forest for audio features, SVM for text features, and XGBoost as meta-classifier.

7.***Preliminary Findings on Audio Features*****(Dataset1 Train Test with text and audio feature)**

- Initial experiments showed limited improvement by adding audio features, possibly due to the fact that **Dataset1 only focused on the lyric sentiment dimension**.

8.***Comparative Performance Tests*****(Dataset2 Test)**

- Lower overall accuracy on the test set may be attributed to the presence of **7,775 unseen tokens**, reflecting potential vocabulary difference between the Dataset2 and Dataset 1.
- Testing with **Dataset2** The combine model outperforms the single model in terms of F1 score, especially the CNN, with an F1 score of **38%**.

9.***Benchmarking Against XL-NET Study*****(Dataset2 Train and Test)**

- Compared with a benchmark study using XL-NET**(59%)** and Lemmatization, combine models achieved higher F1 scores, with CNN reaching up to **68%.**
- Utilizing lyrics and audio features **separately**, the model did not outperform the combined model. lyrics only BEST**(57%)** audio only BEST **(61%)**

10. ***Visualization and Analysis Findings***

- Found strong correlations and clustering between audio features and emotional categories in **Dataset2,** which verifies the effectiveness of the fusion of lyrics and audio features for music sentiment classification.

Dataset1 Train and Test with lyrics and audio feature

Model	Accuracy (AC)	F1 Score
SVM + TFIDF (Pre-Fusion)	89	89
TEXT CNN + DESEN	92	92
BILSTM + DESEN	90	90
Ensemble Learning (Stacking) [SVM (Lyrics) + RF (Audio) + XGBoost (Final)]	91	91

Dataset2 Test

Method	Configuration	Overall F1 Score
SVM	Lyrics Only	32%
SVM	Combined	**34%**
TEXT CNN	Lyrics Only	36% (Angry F1: 49)
TEXT CNN	Combined	**38% (Angry F1: 54)**
BILSTM	Lyrics Only	35%
BILSTM	Combined	**37%**
Ensemble	Combined	**37%**
Naive Bayes (NB)	Lyrics Only	35%

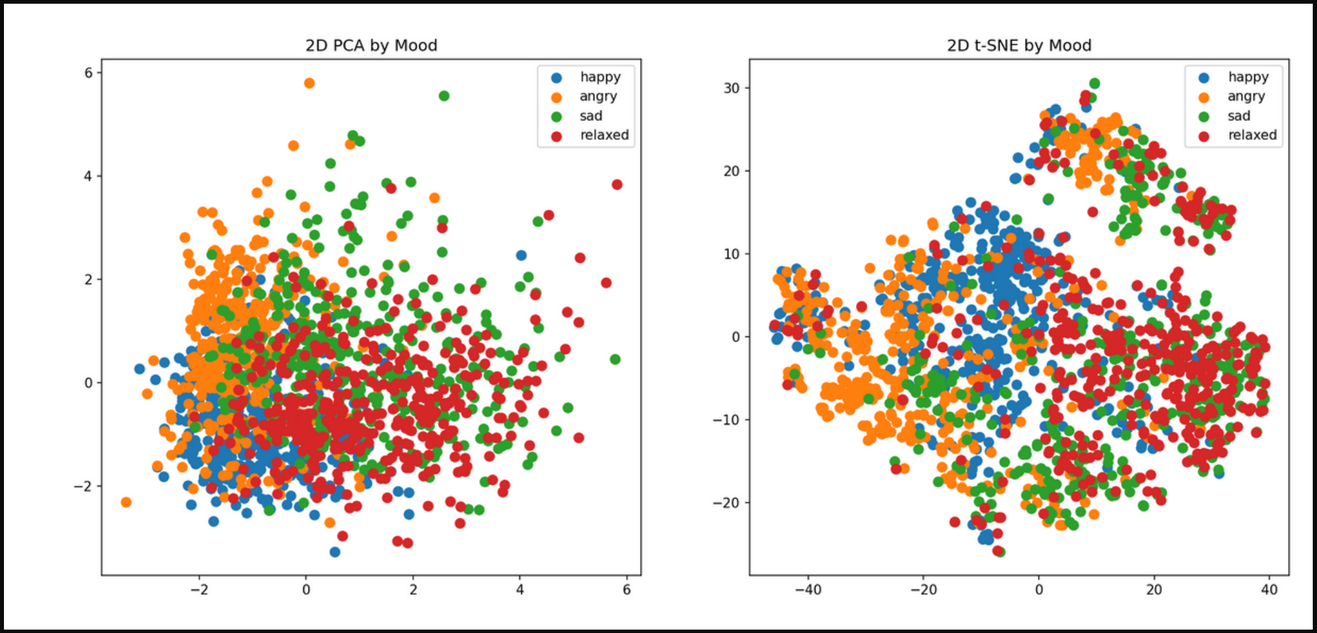
Dataset2 Train and Test

Model	Configuration	F1 Score
XL-NET + Lemma Benchmark	Lyrics Only	59%
SVM	Lyrics Only	54%
SVM	**Combined**	**62%**
CNN	Lyrics Only	57%
CNN	**Combined**	**68%**
Ensemble	**Combined**	**64%**
Naive Bayes - Bag of Words (NB-BOW)	Lyrics Only	52%
BILSTM	Lyrics Only	53%
BILSTM	**Combined**	**64%**
SVM	Audio Only	61%

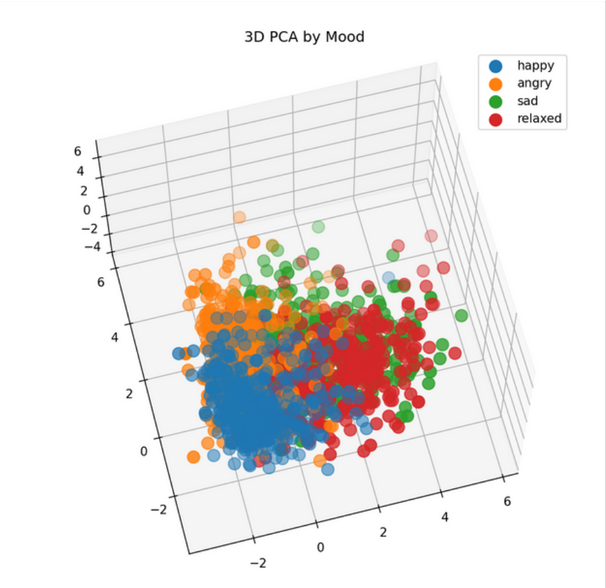
Methodology

Audio Feature

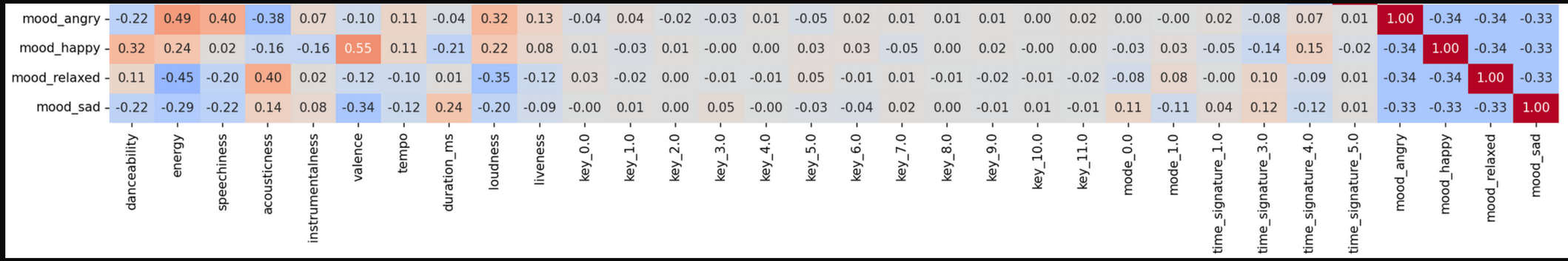
Dataset2: PCA T-SNE



Dataset2: 3D PCA

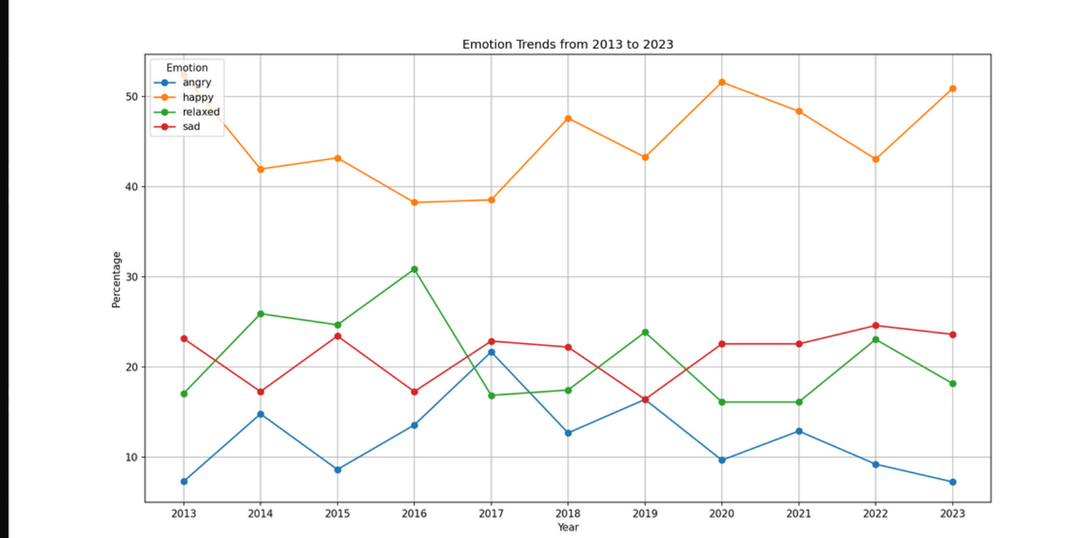


Dataset2 : Heat map



Methodology

Use case



1.Use Case: Real-World Application

To test the model's potential and accuracy with real-world data.
Applied the developed model to analyze emotions in Spotify's **Top 100 songs** from **2013 to 2023.**

2.*Model Selection Process*

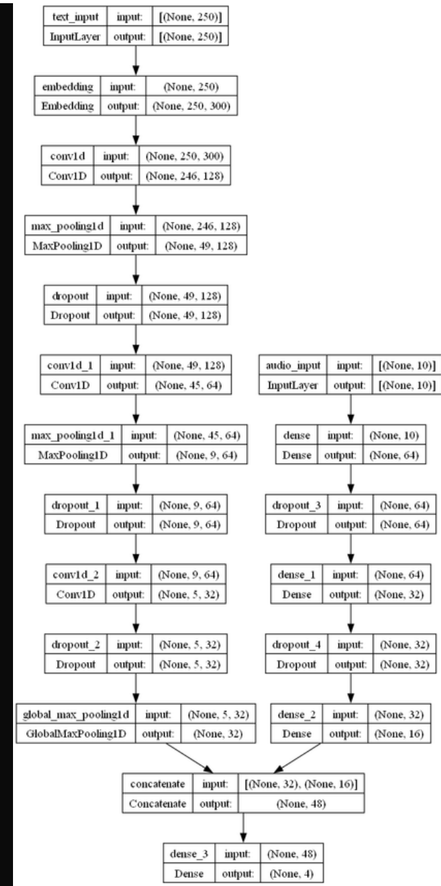
Chose the model based on **cross-validation results and performance on two datasets**.
Selected the CNN model trained on Dataset 2, because of the **best performance** and **generalization ability**.

3.*Analysis of Trends in Music Emotion*

Between **2020 and 2022**, two parallel trends emerged. On the one hand, the number of **"sad"** songs **continues to increase**, reaching its **highest level in a decade by 2022**. On the other hand, the number of **"happy"** songs **continued to decrease** over the same period, hitting a **five-year low by 2022**. These changes may be related to the impact of the global **COVID-19** pandemic and the **Russia-Ukraine conflict.**

These events generated a wide range of **negative sentiments**, which may have mapped the profound impact of global events on **listeners' musical tastes.**

Combine model (Text-CNN With Word2vec and Dense Architecture)



Evaluation and Discussion:

1. ***Ensuring Dataset Quality***
   - Adopted a two-stage method for lyric collection, initially using Genius API, then switching to **custom web scraping** for **better accuracy**.
   - Achieved **dataset balance** through downsampling and ensured **unbiased training** by applying consistent random shuffling.
2. ***Paper Replication and Insights***
   - **Replicated key research** to confirm the original study's **reliability** and deepen my understanding of the **methodologies**, **successfully identifying potential issues**.
3. ***Experiment Design: Focus on Word Embedding, Preprocessing and Audio feature***
   - **Reduced** the **maximum sequence length** in accordance with the **real dataset distribution** to improve performance.
   - Discovered BoW, TFIDF**(Words)**, and Word2Vec**(Local)** outperformed GloVe**(Global)** in analyzing **rhythmic and repetitive texts(Like lyrics)**, highlighting the need for **task-specific embedding selection.**
4. ***Model Tuning and Evaluation***
   - Deeply understood and tuned models using **global searches** and analysis of **learning and loss curves.**
   - Found **appropriate preprocessing** improved model accuracy and performance.
5. ***Audio Feature Integration Assessment***
   - Audio features significantly **enhanced model generalization and performance,** in extensive testing across datasets, highlighting the **importance of audio features** in music sentiment analysis.
6. ***Practical Application and Real-World Testing***
   - Applied model to Spotify's Top 100 songs over ten years, confirming model's **potential and generalizability.**
7. ***Conclusion: Project's Design and Insights***
   - Meticulous design at each phase provided key insights, continuously optimizing the model.
   - Experiments highlighted the potential of combining **lyrics** with **audio features** to advance the field of music sentiment analysis.

Next step

- Conduct a detailed emotional analysis of Spotify's Top 100 songs using **trusted event data**.

- Complete a final report on the findings.

Thank You :)