Final Year Project Demo

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 heritage.

Music and Emotional Expression

A key medium for expressing human emotions.

Music in Cultures and Societies

Plays a central role across diverse 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.

Objective

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

Research

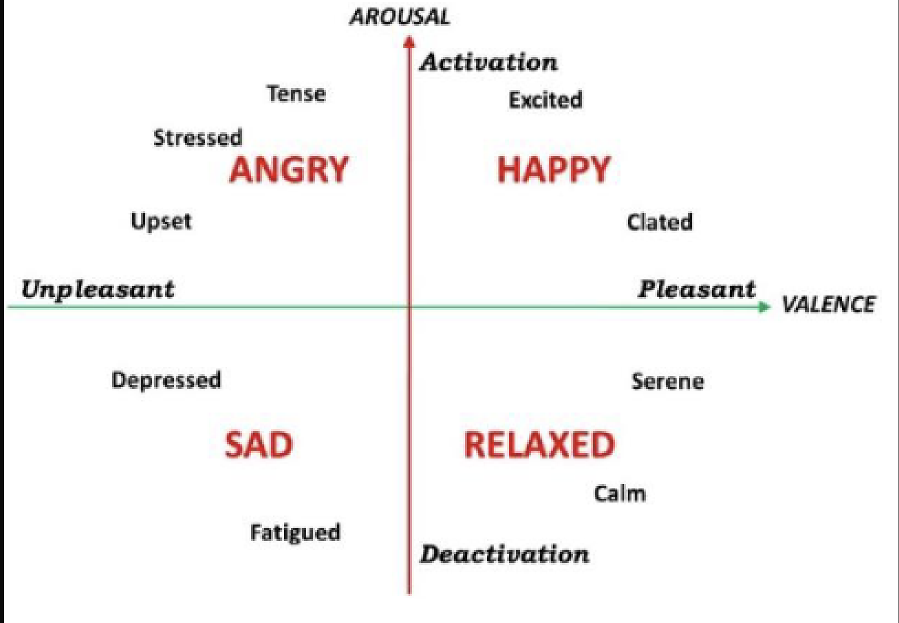


Figure 1. Circumplex Model

MER in Music Information Retrieval
Music Emotion Recognition uses lyrics and audio features for predicting musical emotions.

Inspired by Russell's Model
Detailed emotion classification in MER, inspired by the Russell emotion model.

Comprehensive MER Model Development
Developing an all-encompassing MER model, focusing on lyrics and audio analysis.

Benchmark: Jiddy Abdillah et al.'s Study
Benchmarking against research by Jiddy Abdillah et al., achieving 91.08% accuracy with Bi-LSTM and GloVe.

Goal: Surpassing the Benchmark
Aiming to exceed this benchmark in emotion classification's granularity and accuracy with a composite model.

Methodology

Dataset

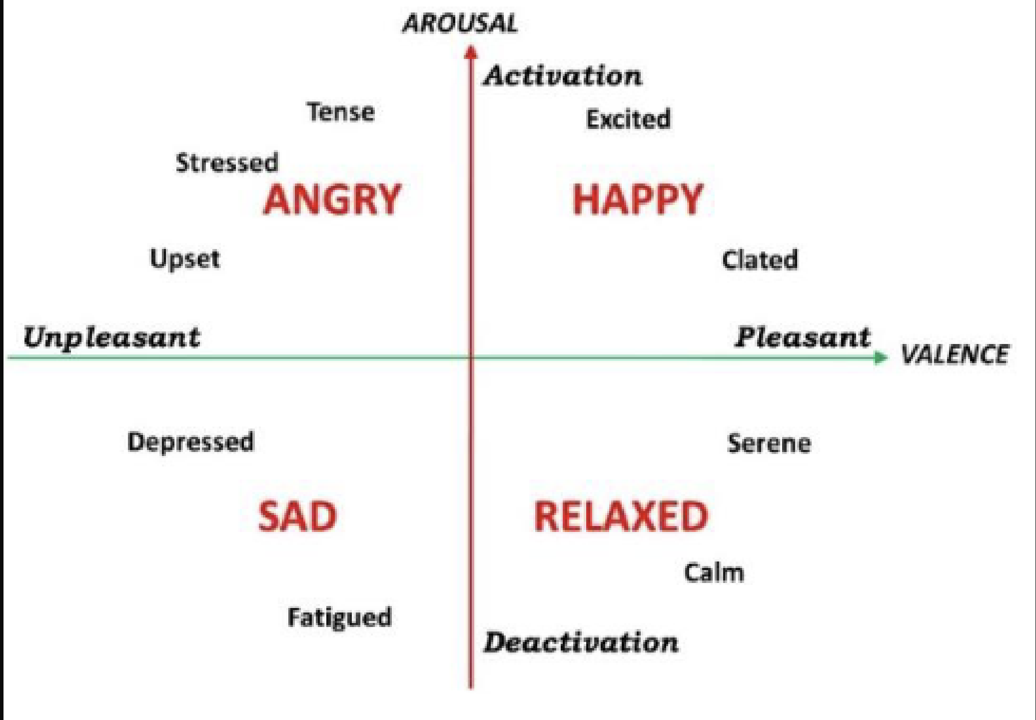


Figure 1. Circumplex Model

__Dataset1__: **MoodyLyrics**: Contains 2595 songs annotated in 4 quadrants 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**).

Methodology

Dataset

Lyric

Lyric Data Acquisition and Optimization

- Initial Attempt: Genius API
  - Using lyricsgenius to obtain lyrics based on song names and artists.
  - Issue: Relies on exact match of song titles and artist names, prone to errors.

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

Audio Feature Extraction

- Using Spotify API.
  - Locating specific songs based on song names and artists.
  - Acquiring 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** is 2123

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

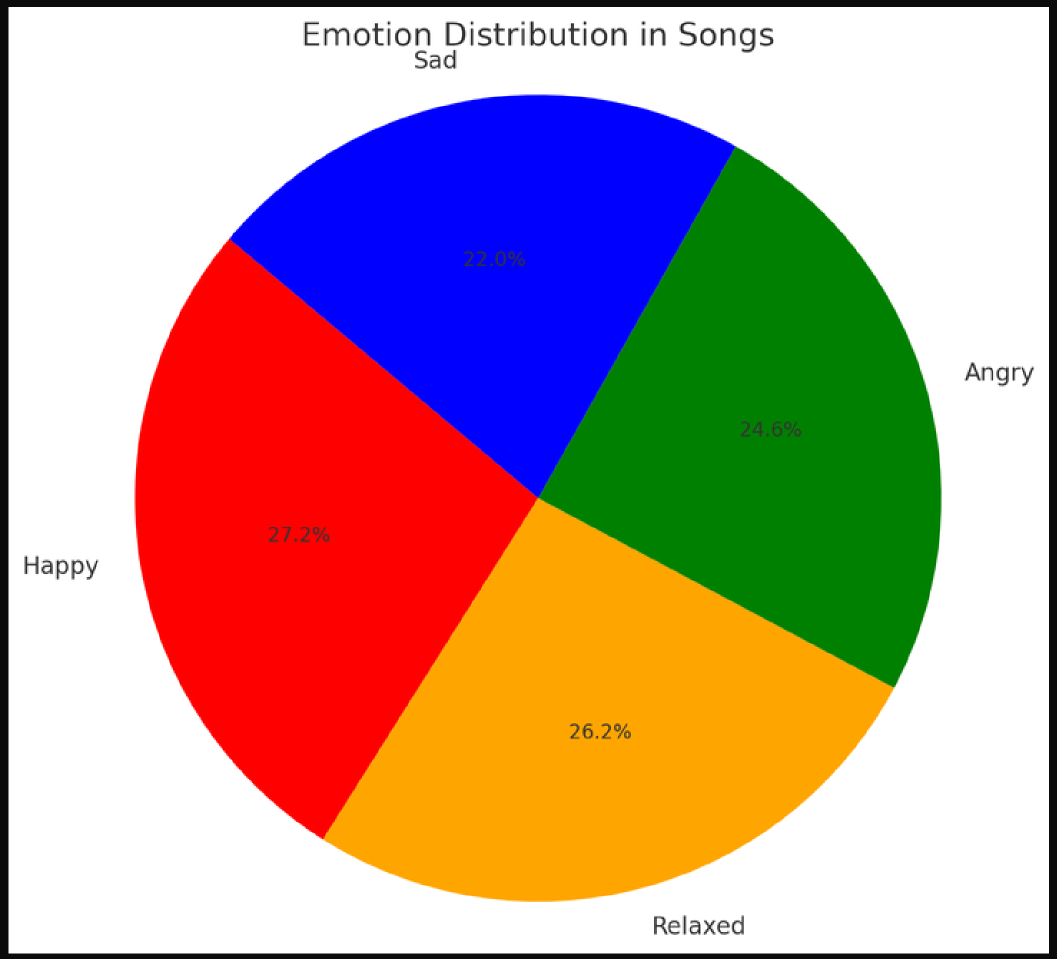
The ***Final Dataset1*** is 2033

Happy:554（27.2%）

Relaxed:532（26.2%）

Angry:501（24.6%）

Sad:448（22.2%）



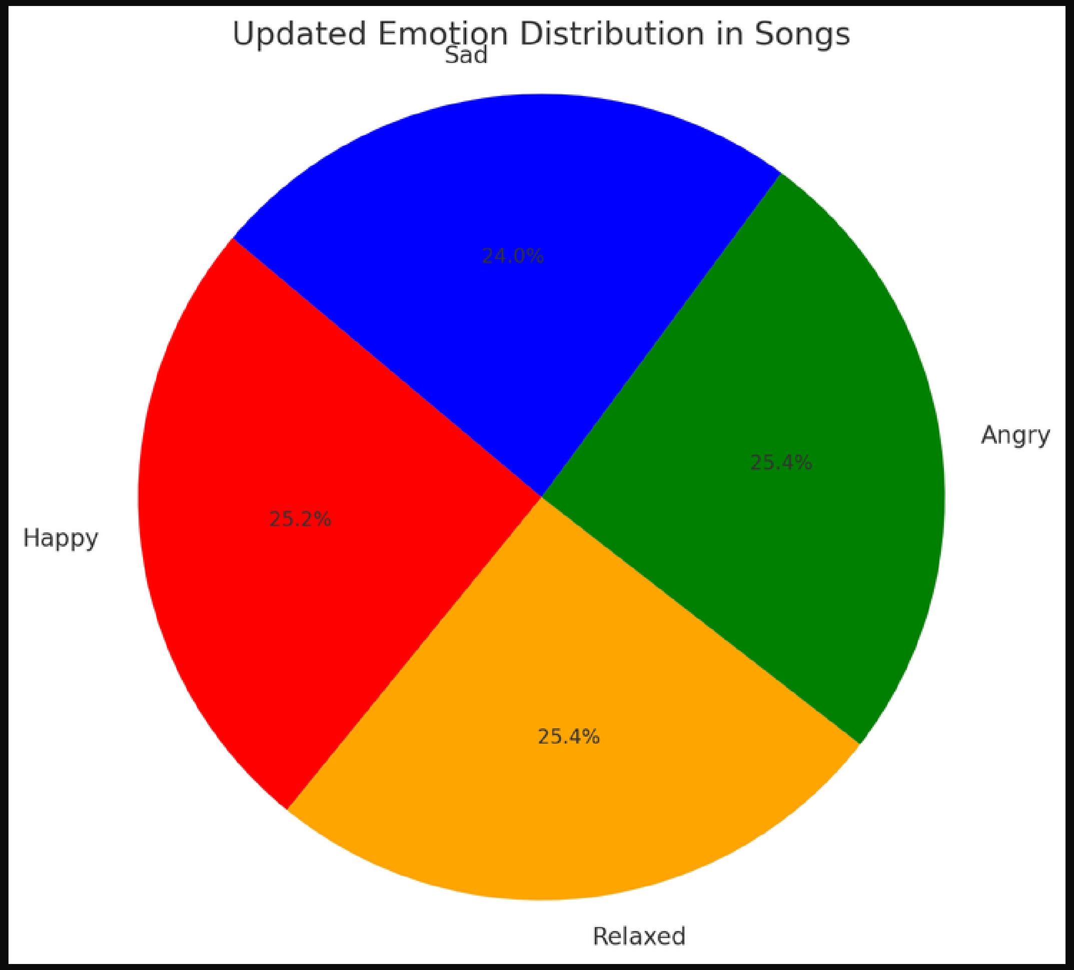
The ***Final Dataset2*** is 1576

Happy:394（25.2%）

Relaxed:396（25.4%）

Angry:396（25.4%）

Sad:375（24%）



Methodology

Reproducing the paper

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%

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***
   - Adhered to the paper's specified hyperparameters and structure.
   - Used pretrained GloVe 100-dimensional vectors 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(NB), K-Nearest Neighbors(KNN), Support Vector Machine(SVM), Convolutional Neural Network(CNN), Long Short-Term Memory Network(LSTM), and Bidirectional Long Short-Term Memory Network(Bi-LSTM).
   - Achieved accuracy similar to the original paper for each model.

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**

Reproducing paper result

Methodology

Experiment Design

Main objective of experimental design

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

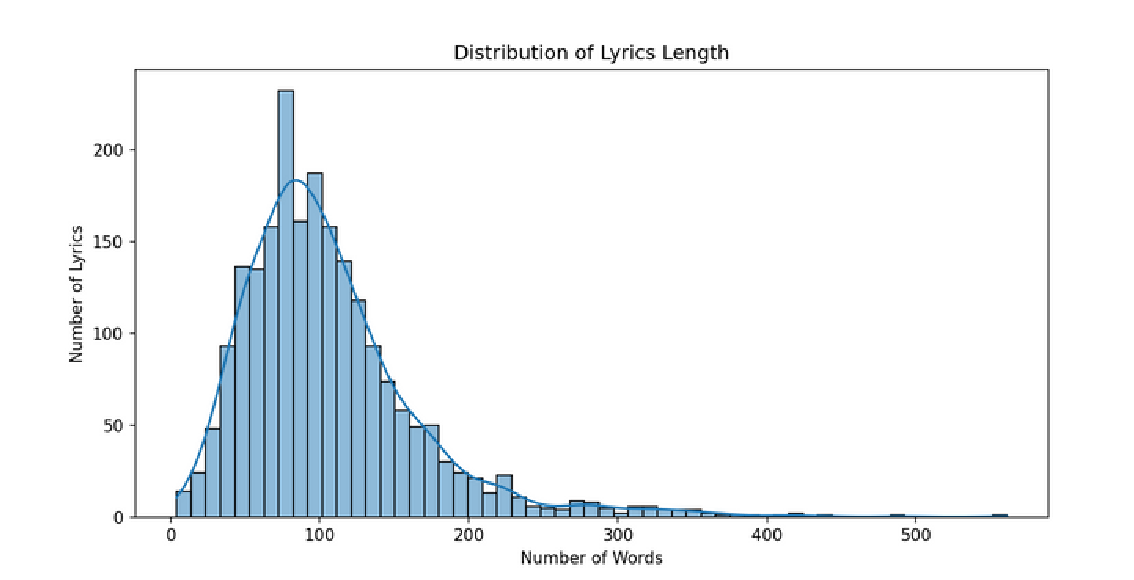
2.Preprocessing
Implement efficient data preprocessing strategies to optimize model inputs.

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

Methodology

Word Embedding

Sequence Length



Sequence Length Optimization

Reduced max sequence length from **1000 to 250** to minimize padding noise impact.

Word Embedding Approaches

Applied Bag of Words (BoW), TFIDF, and Word2Vec300d with uniform preprocessing steps.

Model Tuning Strategies

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

Embedding Technologies Performance

BoW, TFIDF, and Word2Vec outperformed GloVe, surpassing baseline accuracy.

Embedding Operation

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

Preprocessing Strategies and Model Performance

Evaluated the impact of Stemming(Stem), Lemmatization(Lemma), Noise Removal(NR), and Stopword(SR) Removal on four top-performing models: Naive Bayes, SVM, Text-CNN, and BiLSTM.

Stability in Experimental Results

Applied a loop testing method for deep learning models to ensure result stability, accounting for random weight initialization.

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

After finalizing embedding, tuning, and optimal preprocessing, notably surpassed baseline accuracy SVM reached **94%** in accuracy and F1 score, demonstrating the effectiveness of our refined approach.

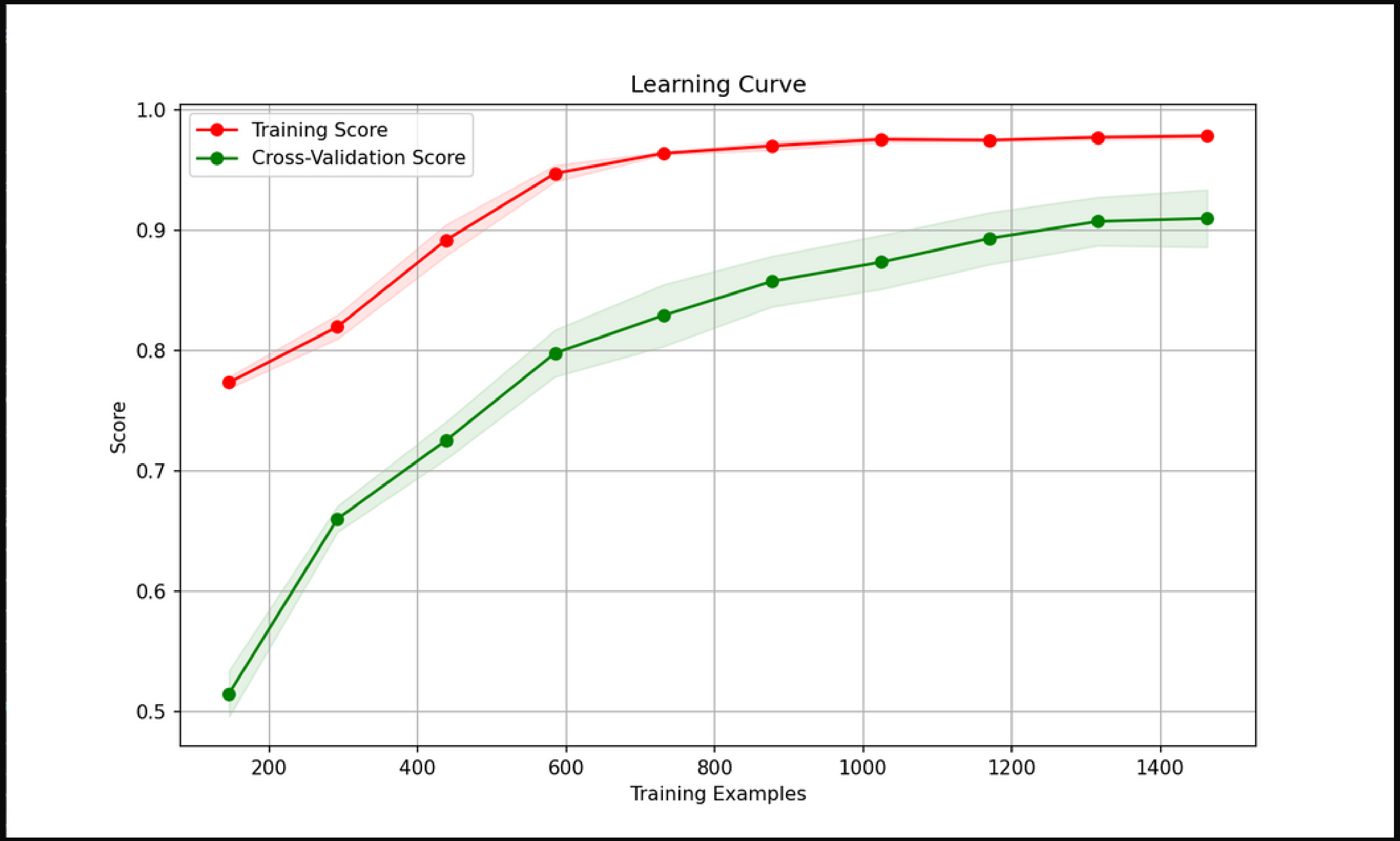
Best Preprocessing Model

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

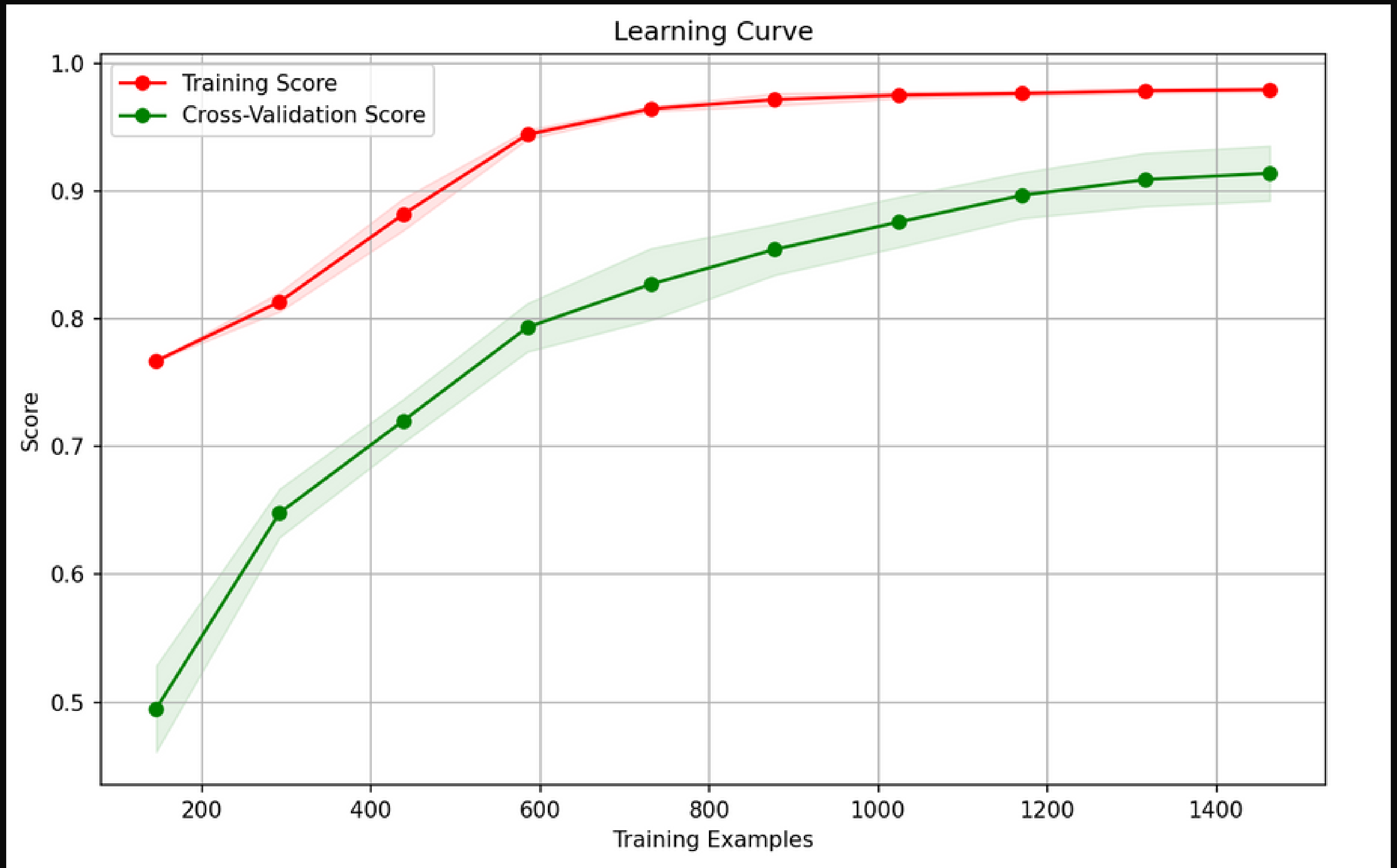
Preprocessing

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 integrating 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 in data points, indicating potential correlation loss in high-dimensional data.
3. ***Correlation Analysis of Audio Features***
   - Audio features showed varied correlations with emotional categories, with some like **energy** and **loudnes**s in "Angry" showing positive correlation, but others less significant.
4. ***Exclusion of Certain Audio Features***
   - Decided to exclude features like **key, mode**, and **time signature** due to limited contribution to emotion classification and concerns about the **curse of dimensionality.**
5. ***Training with Audio-Only Features***
   - Conducted training and prediction using only audio features, but results were suboptimal.

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

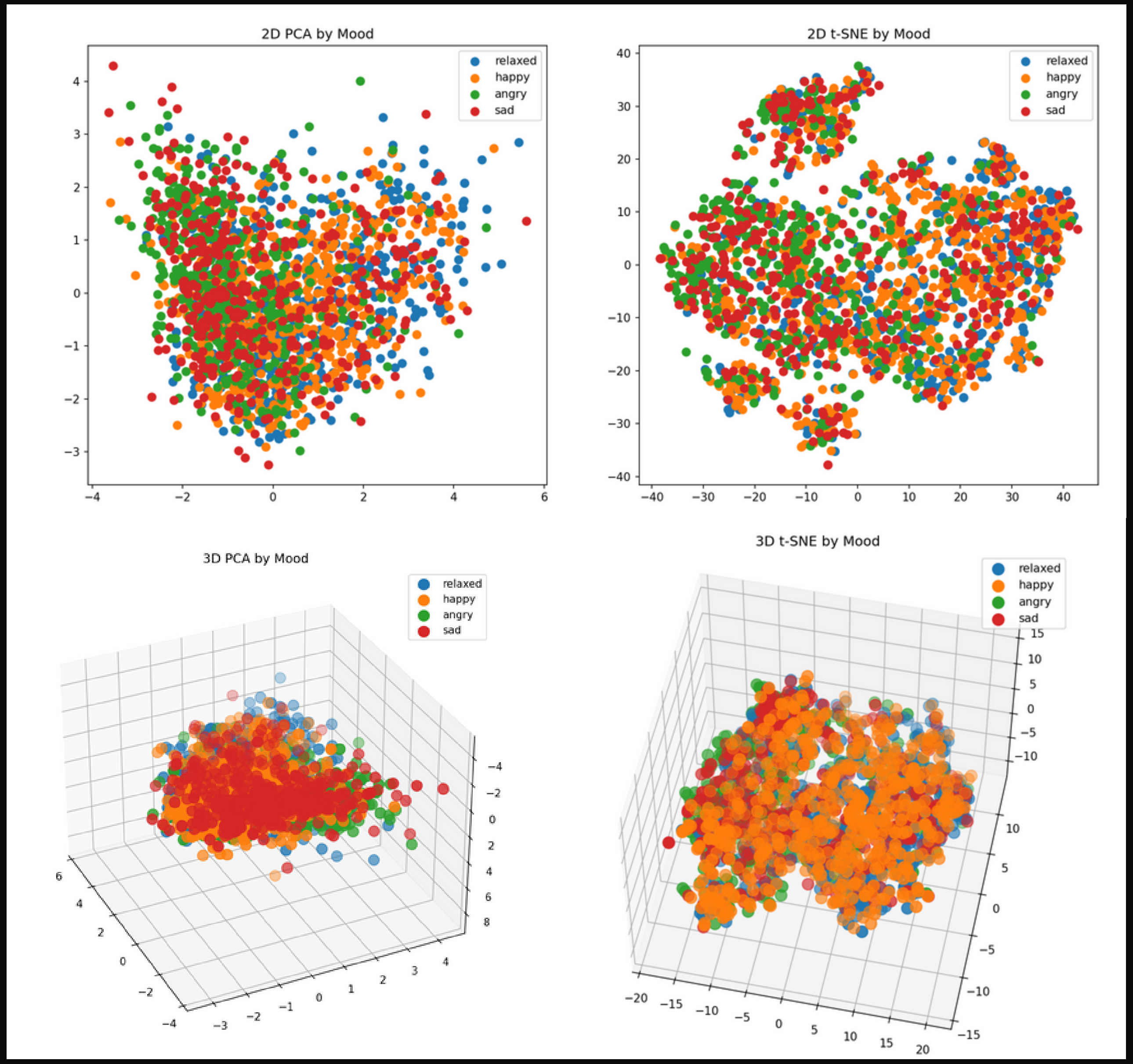
Audio Only Model Performance

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

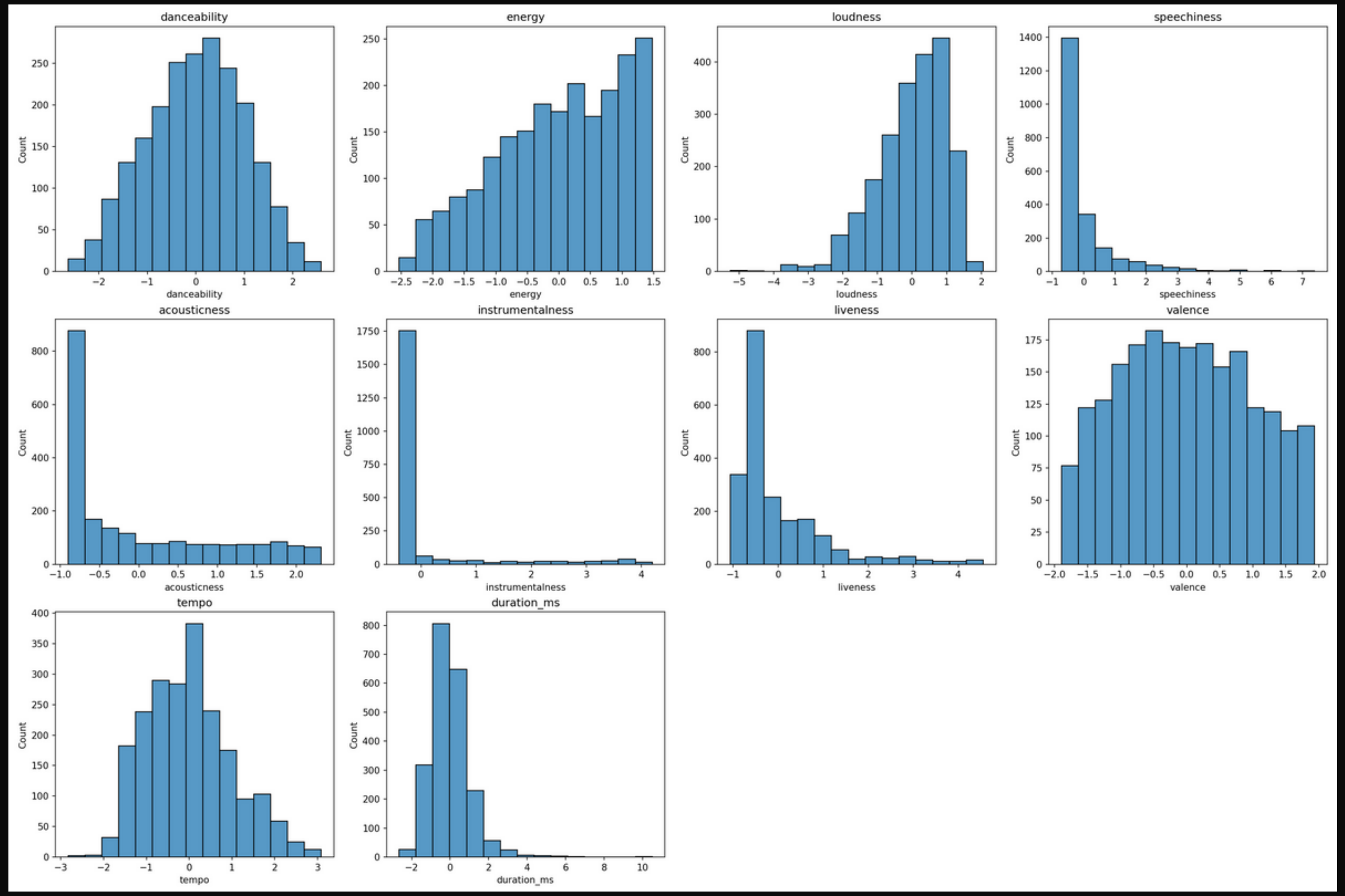
Methodology

Audio Feature

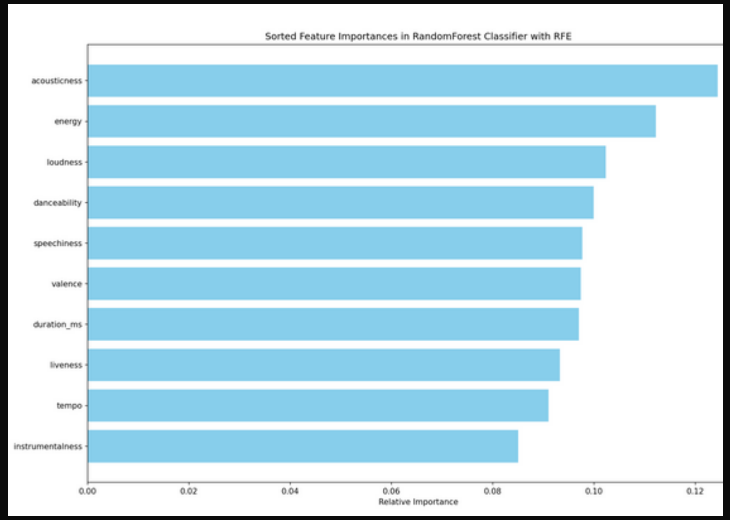
Dataset1 : PCA T-SNE



Dataset1 : Histogram data distribution



Dataset1 : RF



Dataset1 : Heat Map

	danceability	energy	speechiness	acousticness	instrumentalness	valence	tempo	duration_ms	loudness	liveness	key_0.0	key_1.0	key_2.0	key_3.0	key_4.0	key_5.0
mood_angry	-0.08	0.28	0.12	-0.27	0.05	-0.08	0.08	0.08	0.22	0.03	-0.02	0.02	-0.01	-0.02	-0.00	-0.01
mood_happy	0.08	-0.11	-0.06	0.11	-0.09	0.07	-0.04	0.04	-0.06	0.03	0.03	-0.01	-0.06	0.02	-0.02	0.02
mood_relaxed	0.03	-0.22	-0.10	0.20	-0.03	0.02	-0.07	-0.06	-0.19	-0.04	-0.01	-0.02	0.02	0.02	0.00	0.01
mood_sad	-0.03	0.06	0.05	-0.05	0.08	-0.02	0.04	-0.06	0.05	-0.02	0.00	0.01	0.06	-0.02	0.02	-0.02

	key_6.0	key_7.0	key_8.0	key_9.0	key_10.0	key_11.0	mode_0.0	mode_1.0	me_signature_1.0	me_signature_3.0	me_signature_4.0	me_signature_5.0	mood_angry	mood_happy	mood_relaxed	mood_sad
mood_angry	-0.01	-0.02	0.03	0.01	-0.02	0.05	0.09	-0.09	-0.04	-0.08	0.08	0.01	1.00	-0.35	-0.34	-0.30
mood_happy	0.03	0.02	-0.01	-0.00	0.01	-0.01	-0.05	0.05	0.00	0.01	-0.01	-0.00	-0.35	1.00	-0.36	-0.32
mood_relaxed	0.00	0.01	-0.00	-0.02	0.04	-0.04	-0.07	0.07	0.03	0.06	-0.07	0.02	-0.34	-0.36	1.00	-0.32
mood_sad	-0.03	-0.01	-0.02	0.02	-0.03	-0.00	0.03	-0.03	0.01	0.00	0.00	-0.03	-0.30	-0.32	-0.32	1.00

Methodology

Audio Feature

6.***<u>Integration of Audio Features into Models</u>***

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

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

- Initial experiments showed limited improvement from audio features, hypothesized due to dataset's focus on lyrical emotion.

8.***<u>Comparative Performance Tests</u>*****(Dataset2 Test)**

- Conducted tests using a second dataset; composite models outperformed single models in F1 scores, especially CNN with an increase in F1 score up to **38%**.

9.***<u>Benchmarking Against XL-NET Study</u>*****(Dataset2 Train and Test)**

- Compared with a benchmark study using XL-NET**(59%)** and Lemmatization, composite models achieved higher F1 scores, with CNN reaching up to **67%.**

10. ***<u>Visualization and Analysis Findings</u>***

- Found strong correlations between audio features and emotional categories; PCA indicated clustering patterns, validating the effectiveness of integrating lyrics and audio features for emotion classification.

Dataset1 Train Test with text and audio feature

Model	Accuracy (AC)	F1 Score
SVM+ TFIDF (early)	89	89
TEXT CNN+DESEN	92	92
BILSTM+DESN	90	90
EMSAMBALEARNING (stacking) (SVM (lyrics) + RF (audio) + XGBOOST (final))	91	91

Dataset2 Test

Method	TEST F1 Moody4Q
SVM+TFIDF (lyrics only)	32%
SVM+ TFIDF (early)	**34%**
TEXT CNN	36% Angry F1(49)
TEXT CNN+DESEN	**38% Angry F1(54)**
BILSTM+word2vec	35%
BILSTM+DESN	**37%**
Ensemble	**37%**
NB-bow (lyrics only)	35%

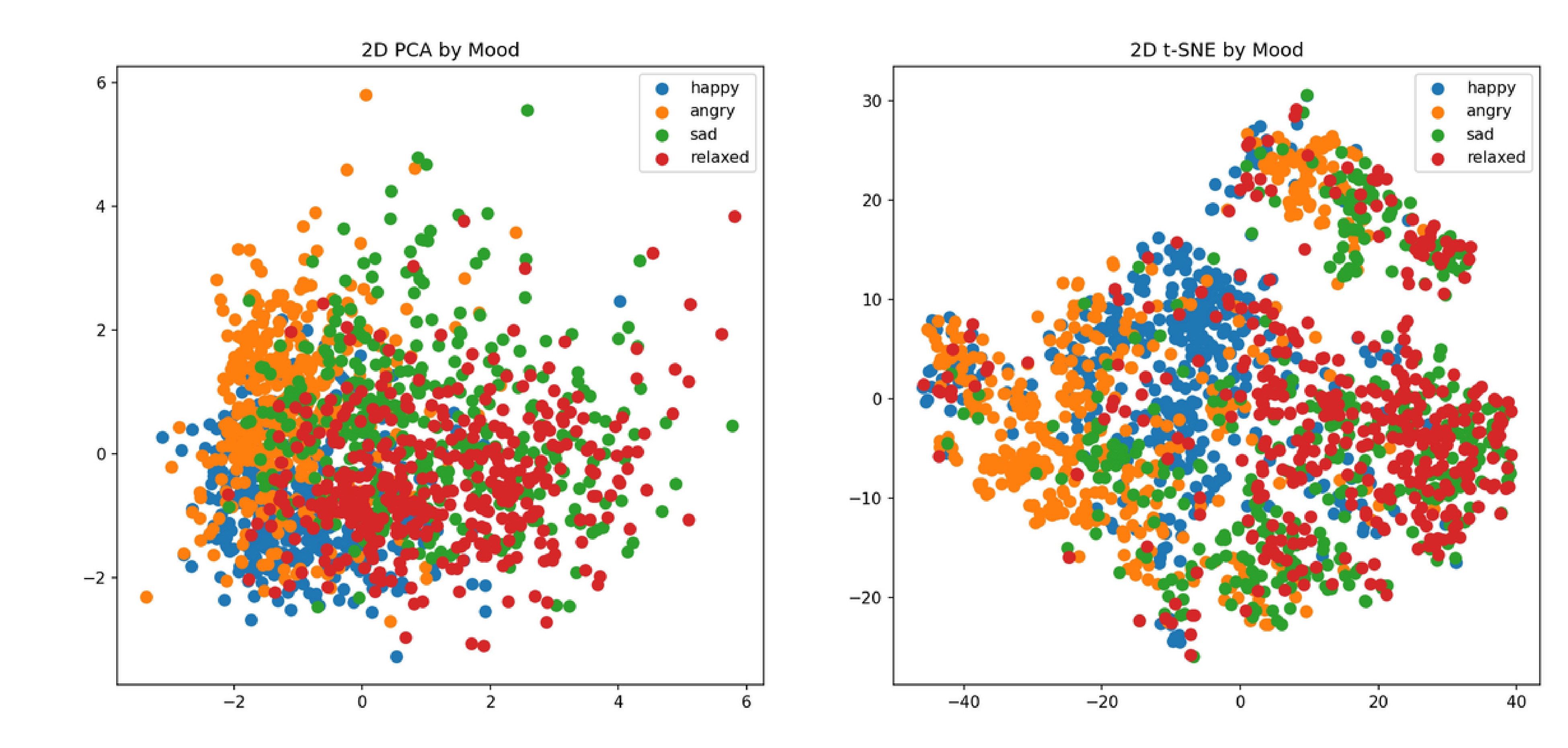
Dataset2 Train and Test

Method	Training Test F1
XL-NET+Lemma benchmark (lyrics only)	59%
SVM+TFIDF (lyrics only)	54%
SVM+TFIDF	**62%**
CNN (lyrics only)	57%
CNN+DESEN best	**67%**
Ensemble	**64%**
Nb-bow (lyrics only)	52%
bilstm (lyrics only)	53%
Bilstm	64%
SVM (audio only)	61%

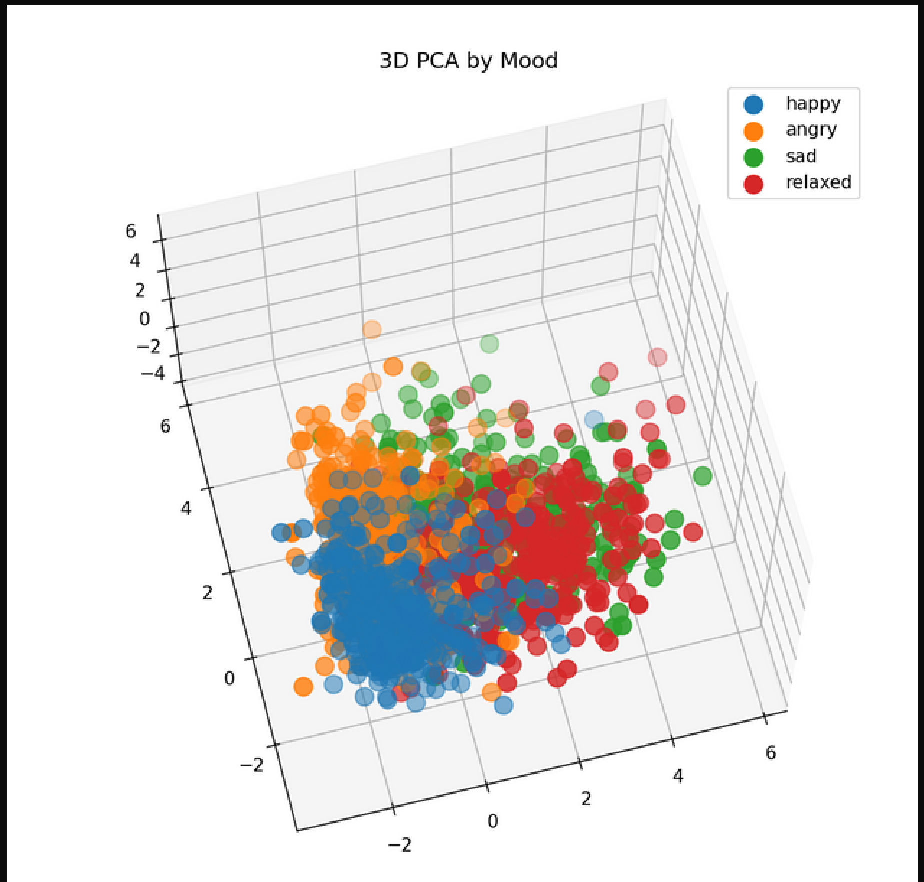
Methodology

Audio Feature

Dataset2: PCA T-SNE



Dataset2: 3D PCA

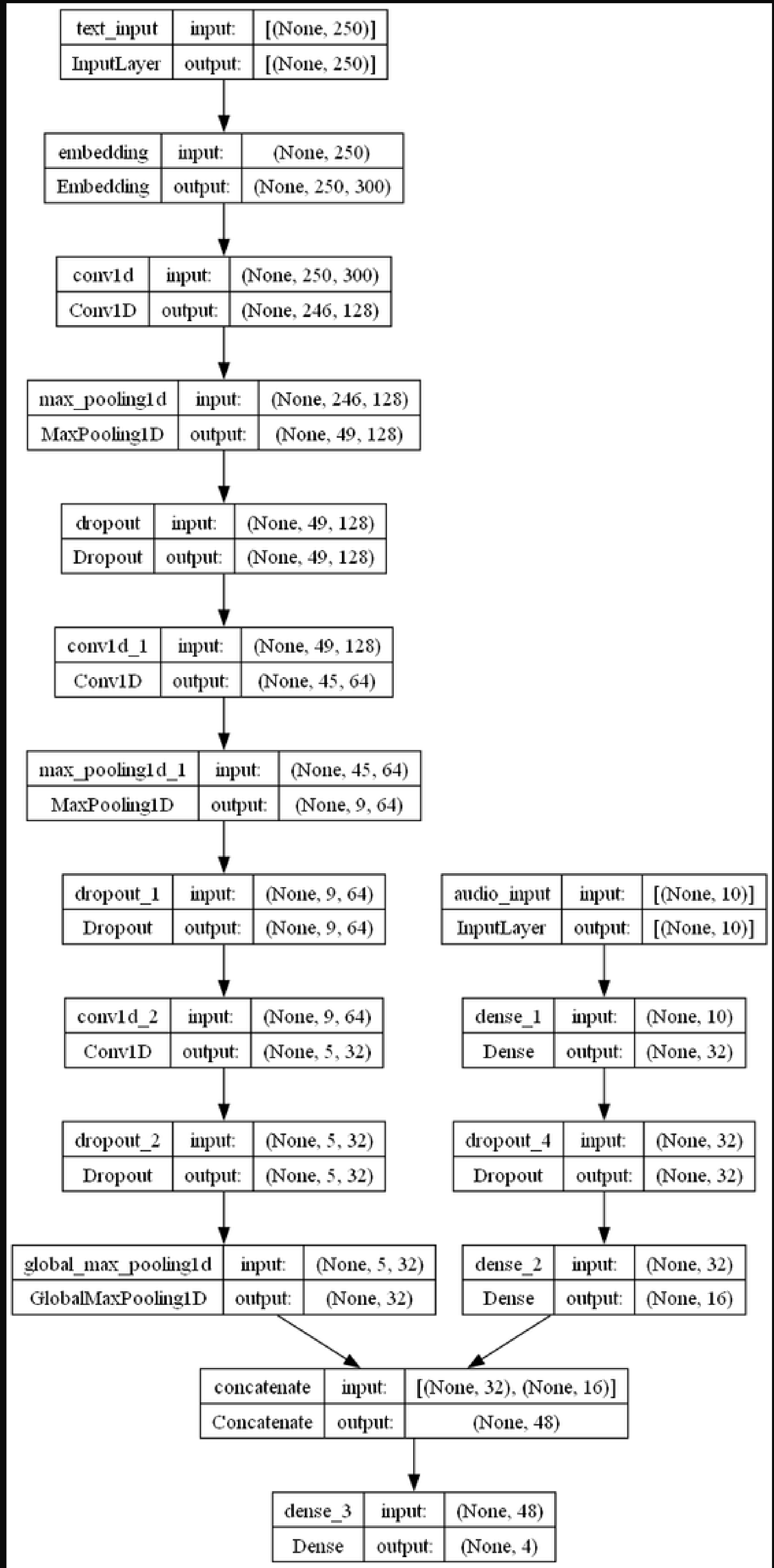


Dataset2 : Heat map

	danceability	energy	speechiness	acousticness	instrumentalness	valence	tempo	duration_ms	loudness	liveness	key_0.0	key_1.0	key_2.0	key_3.0	key_4.0	key_5.0
mood_angry	-0.22	0.49	0.40	-0.38	0.07	-0.10	0.11	-0.04	0.32	0.13	-0.04	0.04	-0.02	-0.03	0.01	-0.05
mood_happy	0.32	0.24	0.02	-0.16	-0.16	0.55	0.11	-0.21	0.22	0.08	0.01	-0.03	0.01	-0.00	0.00	0.03
mood_relaxed	0.11	-0.45	-0.20	0.40	0.02	-0.12	-0.10	0.01	-0.35	-0.12	0.03	-0.02	0.00	-0.01	-0.01	0.05
mood_sad	-0.22	-0.29	-0.22	0.14	0.08	-0.34	-0.12	0.24	-0.20	-0.09	-0.00	0.01	0.00	0.05	-0.00	-0.03

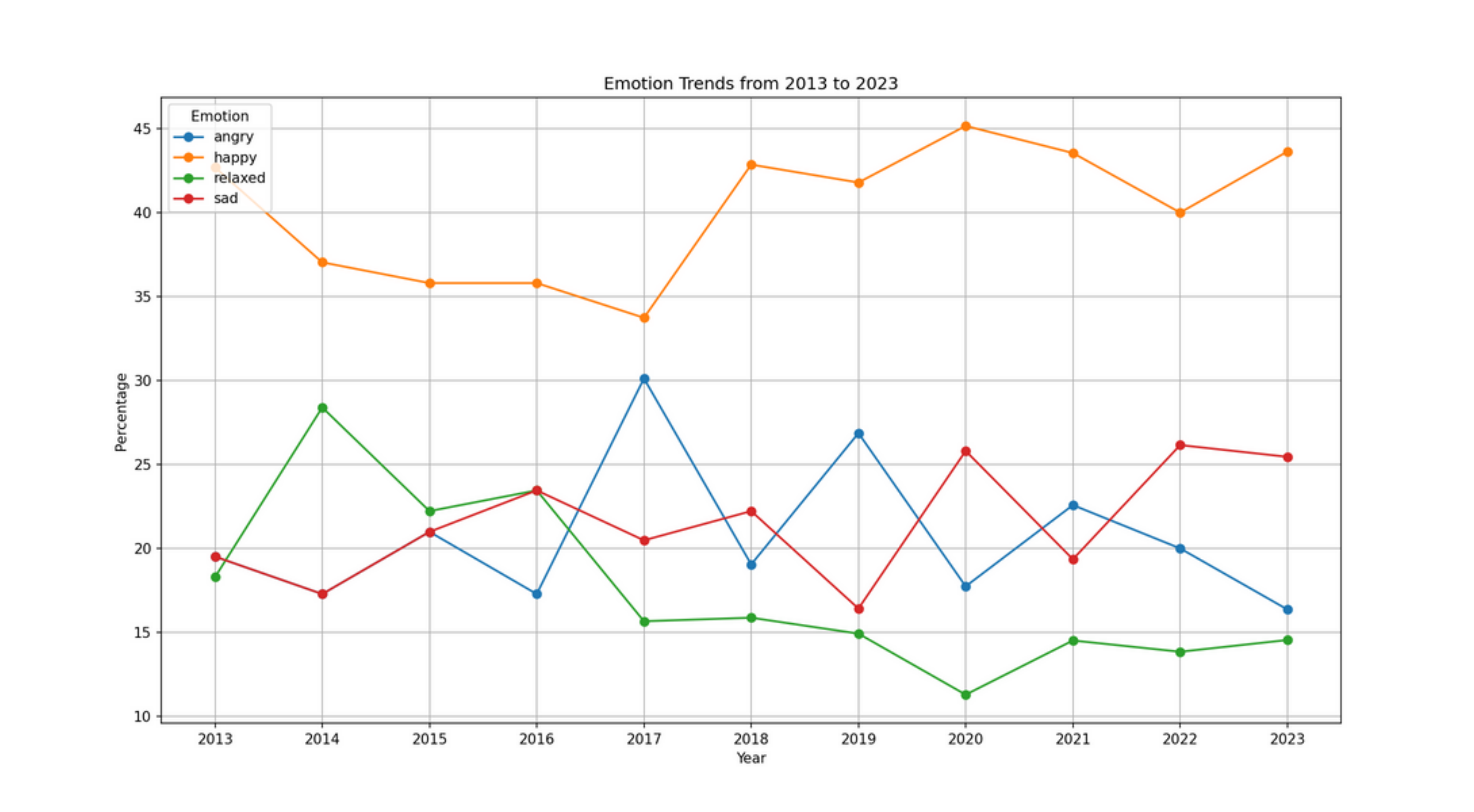
	key_6.0	key_7.0	key_8.0	key_9.0	key_10.0	key_11.0	mode_0.0	mode_1.0	time_signature_1.0	time_signature_3.0	time_signature_4.0	time_signature_5.0	mood_angry	mood_happy	mood_relaxed	mood_sad
mood_angry	0.02	0.01	0.01	0.01	0.00	0.02	0.00	-0.00	0.02	-0.08	0.07	0.01	1.00	-0.34	-0.34	-0.33
mood_happy	0.03	-0.05	0.00	0.02	-0.00	0.00	-0.03	0.03	-0.05	-0.14	0.15	-0.02	-0.34	1.00	-0.34	-0.33
mood_relaxed	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	-0.08	0.08	-0.00	0.10	-0.09	0.01	-0.34	-0.34	1.00	-0.33
mood_sad	-0.04	0.02	0.00	-0.01	0.01	-0.01	0.11	-0.11	0.04	0.12	-0.12	0.01	-0.33	-0.33	-0.33	1.00

Text-CNN With Word2vec and Dense Architecture



Methodology

Use case



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

Model Selection Process
Chose the model based on **cross-validation results and performance on two datasets**.
Selected the CNN model trained on Dataset 2 for its superior performance and generalization capabilities.

Analysis of Trends in Music Emotion
Observed a significant increase in "Sad" songs in 2020 and 2022, possibly linked to the global COVID-19 pandemic.
Noted a continuous decrease in "Happy" songs from 2020 to 2022, hitting a five-year low in 2022.
These trends may reflect the impact of global events like the Russia-Ukraine conflict in 2022, influencing widespread unrest and negative emotions.

Evaluation and Discussion:

1. ***<u>Ensuring Dataset Quality</u>***
   - Used "MoodyLyrics" and "MoodyLyrics4Q" datasets annotated by the Russell emotion model.
   - Adopted a two-stage method for lyric acquisition, 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. ***<u>Paper Replication and Insights</u>***
   - Replicated key research to validate original study's reliability and deepen understanding of methods.
   - Gained insights into BiLSTM combined with GloVe, identifying and addressing potential issues.
3. ***<u>Experiment Design: Focus on Word Embedding, Preprocessing and Audio feature</u>***
   - Emphasized word embedding techniques, reducing maximum sequence length for enhanced performance.
   - Discovered BoW, TFIDF, and Word2Vec outperformed GloVe in analyzing rhythmic and repetitive texts(Like lyrics), highlighting the need for task-specific embedding selection.
4. ***<u>Model Tuning and Evaluation</u>***
   - Deeply understood and tuned models using global searches and analysis of loss curves.
   - Found appropriate preprocessing improved model accuracy and efficiency.
5. ***<u>Audio Feature Integration Assessment</u>***
   - Audio features significantly enhanced model performance in extensive testing across datasets, underscoring their importance in multimodal music analysis.
6. ***<u>Practical Application and Real-World Testing</u>***
   - Applied model to Spotify's Top 100 songs over ten years, confirming model's potential and generalizability.
7. ***<u>Conclusion: Project’s Design and Insights</u>***
   - Meticulous design at each phase provided key insights, continuously optimizing the model.
   - Experiments highlighted the potential of word embedding techniques and audio features in enhancing music emotion analysis.

Next step

- Conduct a detailed emotional analysis of Spotify's Top 100 songs annually.
- Develop and finalize a comprehensive report on the findings.