Lyrical Mood Classification

Team MoodSwing –

Members:
Aravind Narayanan
Akshith Rajkumar

Permission to post video: YES
Permission to post final report: YES
Permission to post source code: YES
Word Count: 1885
Word Penalty: 0%

Introduction

The goal of our project is to accurately classify song lyrics into emotions such as happy, sad, angry, and relaxed. The motivation behind our goal is to understand the emotional impact of songs on listeners and consider the possibility of using song data to analyze the mental state/health of people. When achieved, this could be used to enhance music recommendation systems to satisfy the mood of the user. We believe this to be an important and refreshing improvement in the industry of music applications with the growth of NLP, opening considerable potential for mood-based recommendation system that analyses lyrical information.

Illustration / Figure

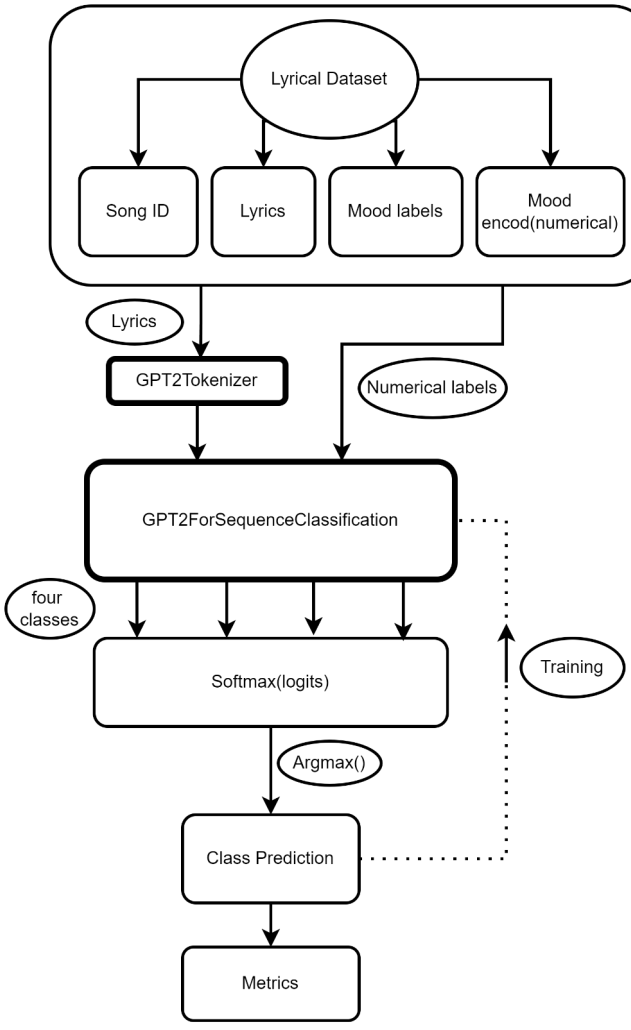

Figure 1: GPT2 Model Architecture

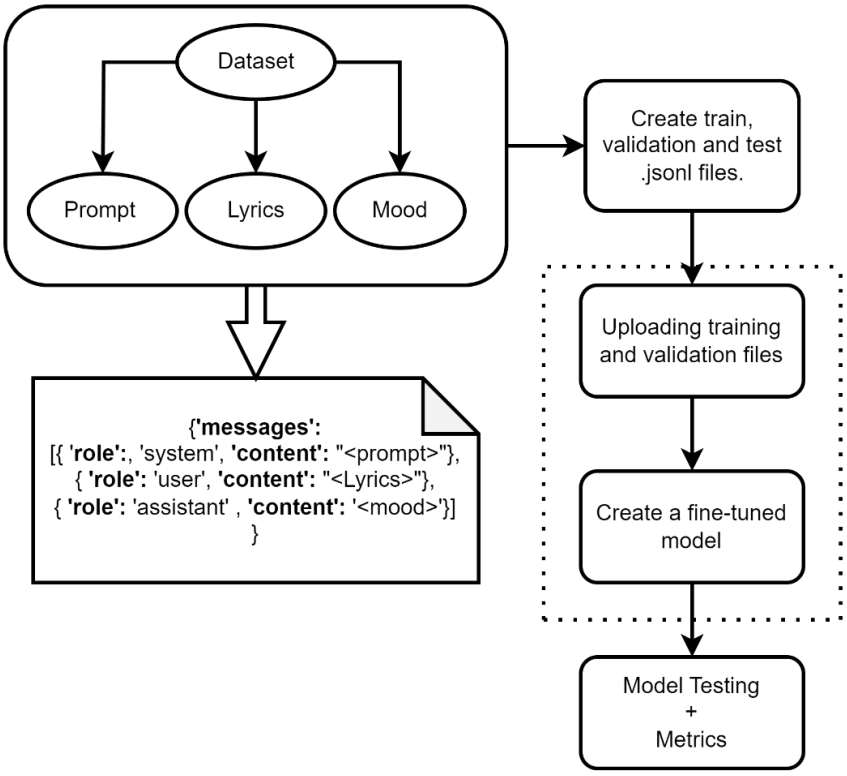

Figure 2: GPT3.5 Model Architecture

Background & Related Work

In the realm of mood classification from lyrics, classical machine learning methods prevailed, leveraging algorithms like Random Forest and Naïve Bayes. However, recent strides in Natural Language Processing (NLP) reveal the prowess of transformer models. [1] introduces a pioneering Bi-LSTM approach with GloVe weighting, demonstrating superior emotion classification efficacy. Comparative analyses highlight Bi-LSTM's robustness against methods like Naïve Bayes and Convolutional Neural Networks.

Additionally, recent developments in the field emphasize the transformative influence of attention networks. Noteworthy is the work by [2], which places a distinctive focus on the textual dimension of songs. Recognizing the pivotal role of lyrics in conveying the mood of a song, the authors employ NLP techniques to extract nuanced features. This study exemplifies the utilization of transformers through BERT, featuring multi-head attention mechanisms that capture semantic relationships between input vectors in diverse ways. The BERT-based model attains an overall accuracy of 58.08%, excelling particularly in predicting the 'Aggressive' class with an F1 score of 0.58. These findings underscore the nuanced performance variations across distinct mood categories, highlighting the significant impact of attention networks on hybrid models.

Data and Data Processing

Data:

For lyrics mood classification, our initial step involved utilizing the MoodyLyricQ dataset. This dataset is a carefully curated collection comprising 2000 song titles and associated artist names, spanning various genres. Each song is tagged with a single mood from four categories - Happy, Sad, Angry, and Relaxed. Notably, these 2000 songs are evenly distributed, ensuring class balance across the four mood tags. However, it's important to highlight that this dataset lacks the lyrics of the songs due to copyright restrictions. To address this limitation, we turned to the Genius APIs' search feature, allowing us to retrieve song lyrics by using the song title and artist name.

Mood	Total Lyrics Scraped
Happy	500
Relaxed	423
Sad	415
Angry	383
Total	1721

Table 1: Statistics of Lyrics Scraped by Mood Category

Data Preprocessing:

To prepare our data for analysis, we initially removed metadata lines, including the first and last lines, as well as an advertisement line, resulting in cleaned data. Further refinement involved eliminating unnecessary extra lines in certain lyrics. While our initial approaches involved removing stop words and applying lemmatization, our final and best-performing model demonstrated optimal results without these steps for both the baseline and top-performing models.

Subset Selection:

For training and testing the final model, we ensured data integrity by meticulously verifying 100 song lyrics, maintaining class balance with 25 songs from each class. An 8:2 train-validation split was implemented, resulting in 80 samples for training and 20 for validation. This initial subset was used for training and validation, and the dataset was expanded by verifying additional songs for a comprehensive evaluation. The final test metrics, detailed in the results section, were derived from this expanded dataset, totaling 80 testing samples.

Architecture and Software

The analysis done in mood classification of lyrics was achieved by using two models. Our first architecture is based on GPT2 model and is illustrated in Figure 1. The work was then progressed to a larger model architecture based on GPT3.5 and is illustrated in Figure 2.

GPT2 Architecture:

Our GPT-2 pretrained model implementation begins by structuring the model architecture with a dataframe containing Song ID, preprocessed lyrics, mood labels, and corresponding encoded numerical labels. The dataset is then split into 80% training and 20% validation sets, processed through the GPT-2 tokenizer, and fed into the GPT2ForSeqeunceClassification pretrained model from Hugging Face, accompanied by numerical labels. Training involves Cross Entropy Loss with an Adam optimizer, utilizing a StepLR scheduler for learning rate optimization (initial rate: 5e-5, gamma: 0.9). A batch size of 4, 20 epochs, and optimization techniques like gradient accumulation and mixed precision are employed. Post-training, the model undergoes evaluation using specified metrics.

GPT3.5 Architecture:

The GPT3.5 model architecture mandates a specific input data format, detailed in Figure 3, which includes a system prompt in addition to lyrical text and mood labels. We initiate the process by creating a prompt array that organizes the data accordingly. After formatting, the data is divided into training, validation, and test datasets. Subsequently, we generate the necessary train, validation, and test jsonl files essential for GPT3.5 model finetuning on OpenAI. Following file creation, the data is uploaded, and a finetuning job for the GPT3.5 model is initiated with default hyperparameters. Once the finetuned model is available, we assess its performance using a comprehensive test dataset and specific metrics.

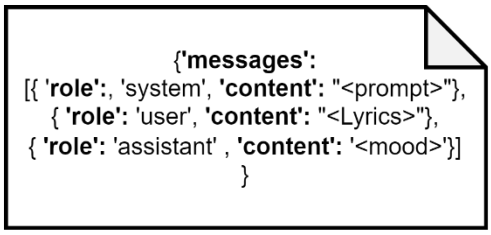

Figure 3: GPT3.5 Model's Input Data format

Prompt: “You are a chatbot that, when prompted with song lyrics, predicts one of the emotions ('Happy', 'Sad', 'Angry', or 'Relaxed') without providing any explanation. Reply with only the emotion name. You do not retain any previous information regarding the lyrics given to you. You specialize in analysing the given song lyrics and predicting the emotion of the song.”

Baseline Model or Comparison

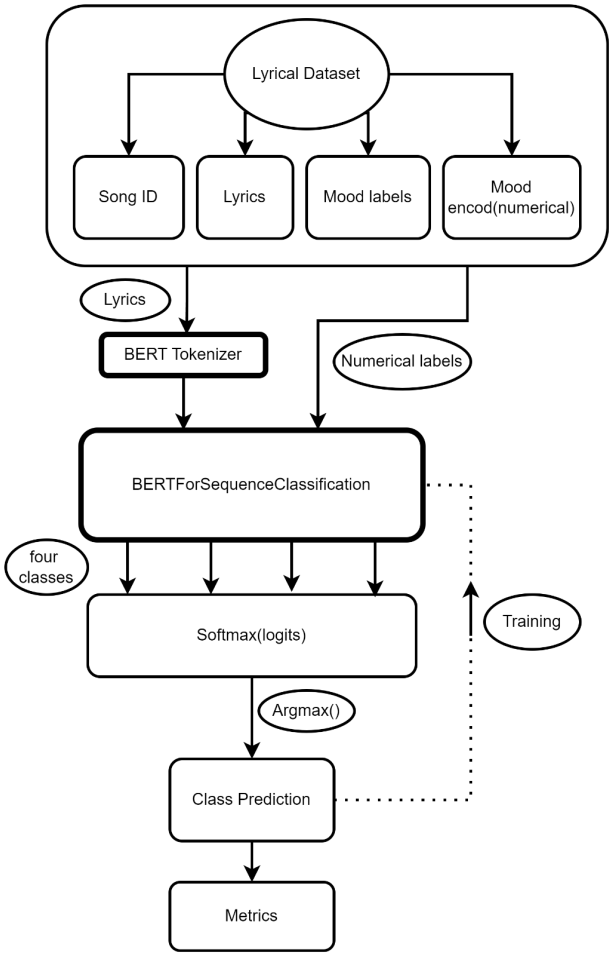

Figure 4: Baseline BERT Architecture

We selected a BERT-based architecture as our baseline model for performance comparison, inspired by prior work in our base paper [2], which extensively evaluates various models for mood classification in lyrics. To ensure comparable results and gain insights into model performance, we implemented our version of the baseline model, following the structure depicted in Figure 4. The primary difference from our GPT2 model lies in the tokenizer and pretrained model from Hugging Face. The loss function, optimizer, epochs, and scheduler mirror those of the GPT2 model, with the sole distinction being a batch size of 8 instead of 4.

Quantitative Results

The results attached below are metrics used to evaluate our models quantitatively. We focus our analysis on the training and validation curves, and the classification report.

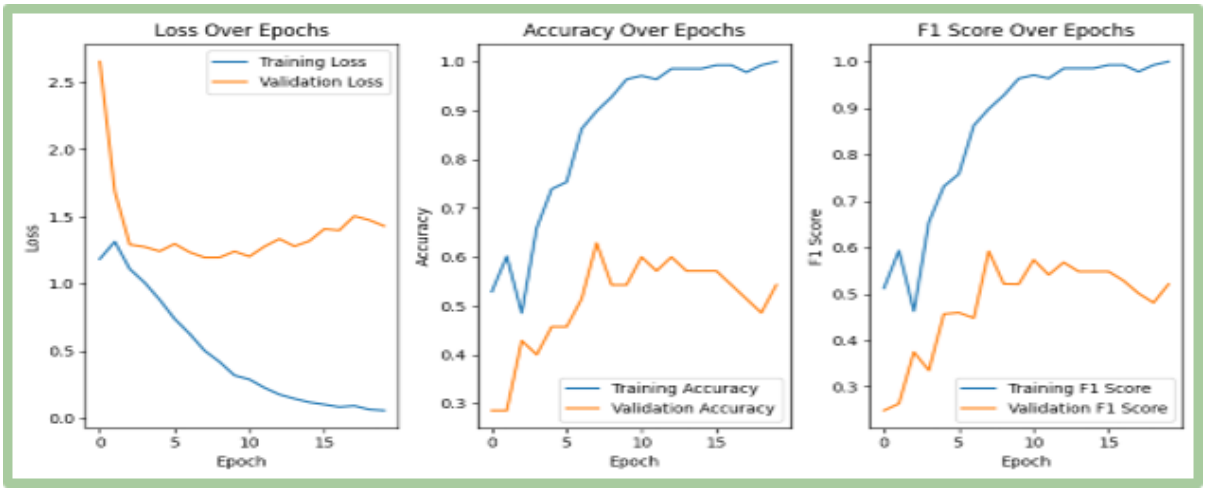

Figure 5: GPT2 Model Training and Validation Curves

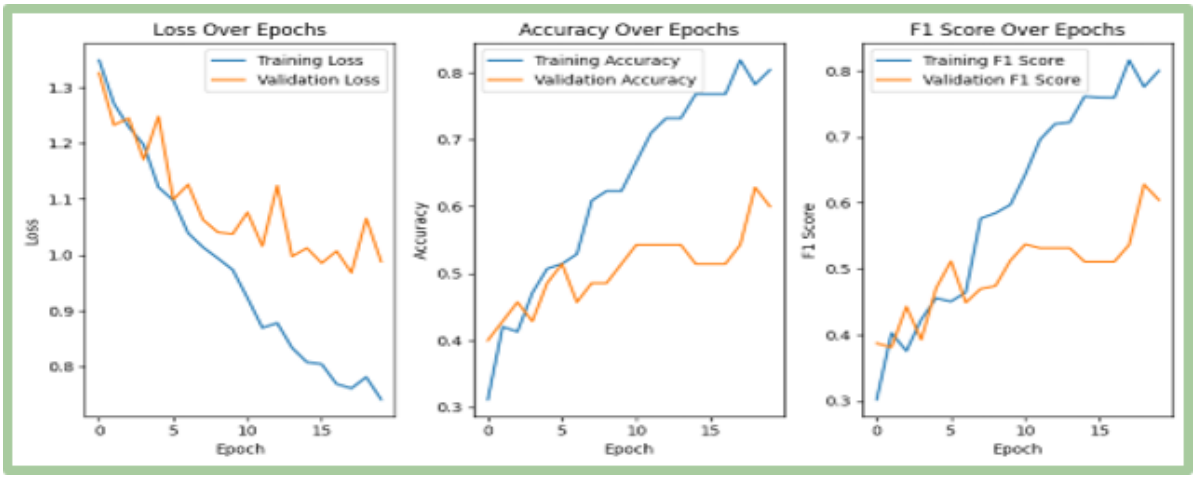

Figure 6: BERT Model Training and Validation Curves

Displayed in Figures 5 and 6 respectively are the training and validation metrics of the previously mentioned GPT2 and BERT models. Both models exhibit improvement over epochs, but a notable observation is the overfitting trend in the GPT2 model. While these graphs were instrumental in tracking our training process, a deeper analysis of model performance required additional metrics.

To delve further, we compiled a classification report for each model attached as Figure 7 and 8. The overall accuracy for BERT is 60%, while GPT2 yields an overall accuracy of 54%. Examining individual classes, we observe strong performance in identifying 'happy' and 'anger.' However, a noteworthy trend emerges, indicating lower performance in the 'relaxed' class compared to others in both models.

Classification Report - Validation:

	precision	recall	f1-score	support
happy	0.67	0.86	0.75	14
sad	0.25	0.40	0.31	5
angry	0.67	0.44	0.53	9
relaxed	0.33	0.14	0.20	7
accuracy			0.54	35
macro avg	0.48	0.46	0.45	35
weighted avg	0.54	0.54	0.52	35

Figure 7: GPT2 Model Classification Report

Classification Report - Validation:

	precision	recall	f1-score	support
happy	0.71	0.71	0.71	14
sad	0.50	0.60	0.55	5
angry	0.71	0.56	0.63	9
relaxed	0.38	0.43	0.40	7
accuracy			0.60	35
macro avg	0.58	0.57	0.57	35
weighted avg	0.62	0.60	0.60	35

Figure 8: BERT Model Classification Report

Efforts to enhance GPT2 led us to explore GPT3.5 for improved classification. The attached Figure 9 illustrates the training and validation loss curves during the fine-tuning of the GPT3.5 model. Limited control over the training process led us to analyze the finetuned model using an 80-lyric test dataset. The subsequent classification report attached as Figure 10 emphasizes strong identification of 'anger' and 'happiness,' with 'relaxed' exhibiting decent recall but lower precision. The F1score of 'sad' seems to indicate difficulty in classification.

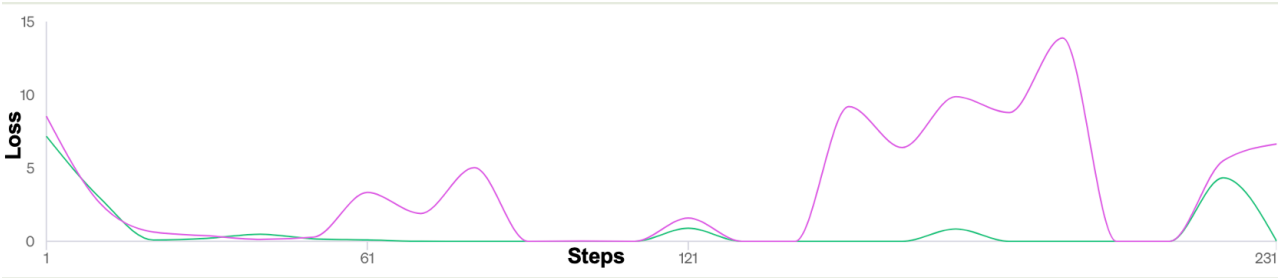

Figure 9: GPT3.5 Model' Training and Validation Curves
(purple, green curves represent validation and training respectively)

	precision	recall	f1-score	support
happy	0.78	0.58	0.67	24
sad	0.54	0.58	0.56	12
angry	0.74	0.74	0.74	19
relaxed	0.52	0.67	0.59	18
accuracy			0.64	73
macro avg	0.64	0.64	0.64	73
weighted avg	0.66	0.64	0.65	73

Figure 10: GPT3.5 Model Classification Report

Qualitative Results

Misclassified Example:

Actual Label: Happy

Predicted Label: Relaxed

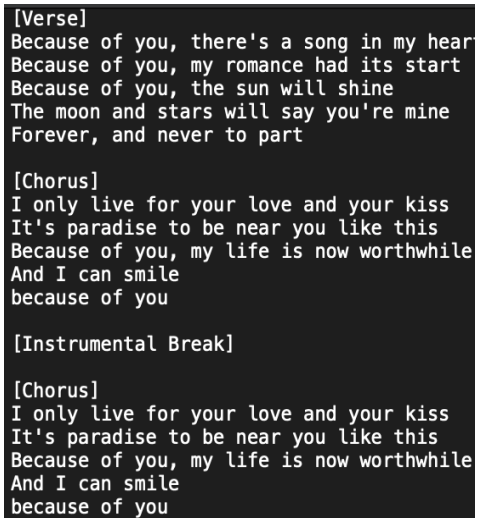

Figure 11: Misclassified Lyric Sample

Potential Reason for Misclassification: The mention of a "song in my heart," "paradise," and the overall soothing tone could contribute to a sense of relaxation.

Correctly classified Example:

Actual Label: Angry

Predicted Label: Angry

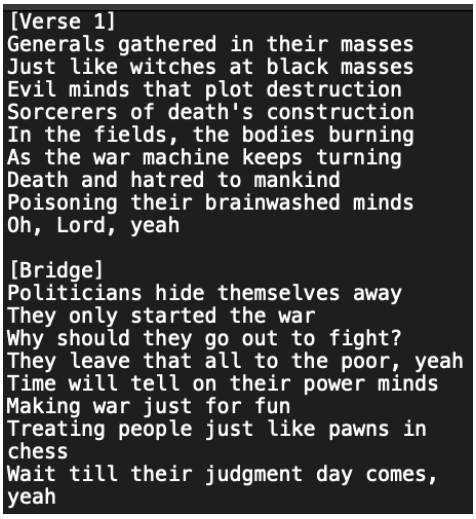

Figure 12: Correctly Classified Lyric Sample

Potential Reason for correct classification: The presence of phrases such as "evil minds," "destruction," and expressions of contempt for war elicits feelings of anger.

While these examples showcase both strengths and areas for improvement, it is evident that the model's performance is sensitive to the nuanced expressions of emotions in text.

Discussion and Learnings

The quantitative and qualitative analysis helped us explore certain interesting findings in the way our system worked. The explanation for why we believe we observe such trends is hypothesized below after analyzing the results deeply.

1. **BERT shows better results than GPT2.**

   BERT may outperform GPT-2 in mood classification of lyrics due to its bidirectional context understanding. Understanding the relationships between words in both directions helps capture nuanced sentiment and context in lyrics. Additionally, BERT's token-level representations allow for a detailed analysis of individual word sentiments, which is crucial for accurately classifying the overall mood of lyrics where specific words heavily influence the mood. The fine-grained contextual understanding provided by BERT makes it well-suited for tasks like mood classification in text.

2. **Minimal Pre-processing shows better performance.**

   For tasks like mood classification in song lyrics, minimal pre-processing without stemming or lemmatization preserves detail and better addresses the complexity of understanding subtle emotional nuances unique to creative content. Song lyrics may have a vocabulary that standard rules of pre-processing fail to capture effectively.

3. **Class Relaxed consistently was misclassified.**

   The concept of being "relaxed" could be more nuanced and challenging to define with specific features. If "relaxed" instances share subtle similarities with "happy," the model may struggle to differentiate between them.

   This also shows the model might place a higher emphasis on valence (extent to which an emotion is positive or negative), it might prioritize features related to positive valence, which is common in both "relaxed" and "happy" instances. Emphasizing valence more might lead the model to downplay arousal (intensity of the associated emotional state) differences. Since valence is not sufficient to distinguish between "relaxed" and "happy," the model may default to predicting the more prevalent class, "happy," when faced with similar valence patterns.

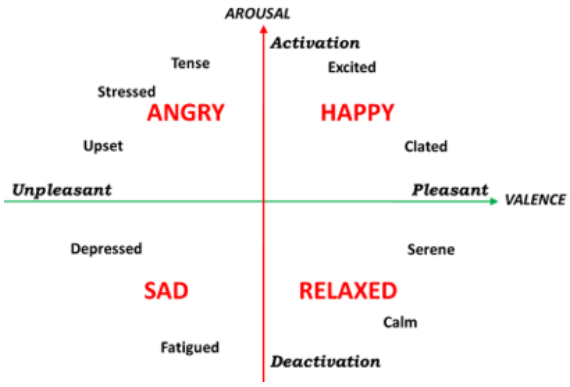

Figure 13: Valence-Arousal Model of Emotions

Recognizing the intricate emotional nuances in lyrics, a key improvement for future work involves adopting a multi-label classification approach. This adjustment acknowledges that lyrics often evoke multiple emotions simultaneously, allowing the model to assign multiple relevant labels and better capture the complexity of emotional expressions in songs. Careful consideration of a comprehensive emotion category set and refinement of the model architecture for effective multi-label handling are essential components of this enhancement.

Individual Contributions

Akshith's Contributions	Aravind's Contributions
1. Conducted data preprocessing. 2. Manually verified 50% of the dataset. 3. Developed GPT2 code and carried out experiments. 4. Designed experiments for GPT3.5 and prepared data format. 5. Compiled and analyzed quantitative results.	1. Collected data and performed web scraping of lyrics. 2. Manually verified the remaining 50% of the dataset. 3. Developed baseline BERT code and conducted experiments. 4. Wrote fine-tuning code for GPT3.5 using the OpenAI API. 5. Compiled and analyzed qualitative results.
Both worked on identifying the anomalies and interesting findings, reports, and presentation together.	

References

1. Abdillah, J., Asror, I., & Wibowo, Y.F. (2020). Emotion Classification of Song Lyrics using Bidirectional LSTM Method with GloVe Word Representation Weighting. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi).
2. S, Sujeesha & Rajan, Rajeev. (2023). Transformer-based Automatic Music Mood Classification Using Multi-modal Framework. Journal of Computer Science and Technology. 23. e02. 10.24215/16666038.23. e02.
3. Valenza, G., Citi, L., Lanatá, A. et al. Revealing Real-Time Emotional Responses: a Personalized Assessment based on Heartbeat Dynamics. Sci Rep 4, 4998 (2014). https://doi.org/10.1038/srep04998