ECE1786H: Creative Applications in Natural Language Processing

PROJECT PROPOSAL DOCUMENT

Team MoodSwing

Aravind Narayanan, Akshith Rajkumar

Word Count:

1. *Figure 1: 53 words*
2. *Figure 2: 31 words*
3. *Figure 3: 23 words*
4. *Total (excluding references) = 1069 + 107 words =* ***1144 words***

Word Penalty: 0%

1. Introduction

In the digital age dominated by platforms like Spotify and Apple Music, efficient music categorization has become vital. Mood-based classification, surpassing traditional genre sorting, responds to users' needs for navigating a vast musical landscape. Research in analyzing the relationship between music and human emotion, with the aid of Natural Language Processing (NLP) and psychological components, is increasingly valuable in delving into an individual's mental state. This evolving approach recognizes the profound influence of music on emotions and the reciprocal impact of emotional states on one's musical preferences.

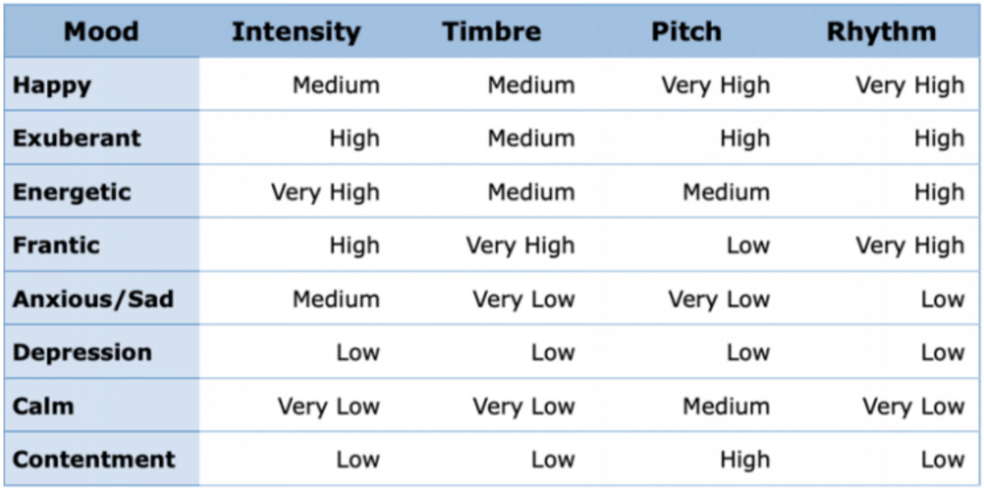
While conventional approaches predominantly focus on the acoustic attributes of music, the new frontier emphasizes the utilization of lyrics and NLP for emotional analysis. Lyrics in songs deviate from standard texts by incorporating stylistic qualities like rhyming and other literary forms, contributing significantly to their emotional essence [1]. Leveraging sentiment mining and transformer models heralds a new phase in music sentiment analysis. Our project explores this uncharted territory, conducting a thorough analysis that integrates both lyrical and audio features. By employing deep learning and NLP, our aim is to unify textual and audio data, revealing the intricate relationship between lyrics and emotional context in music classification.

2. Background

Past research extensively explores music for mood and emotion classification, considering acoustic and textual features. Emotion perception in music is shaped by individual backgrounds and cultural influences. Prior approaches to music-based mood classification typically revolve around three strategies: acoustic features only, lyrics only, and a combined yet independent analysis of both features in synchronization.

2.1. Audio to Mood

In [2], a hierarchical framework is presented to automate the task of mood detection from acoustic music data, by following some music psychological theories in western cultures and classified in categories as shown in Figure 1.



Mood	Intensity	Timbre	Pitch	Rhythm
Happy	Medium	Medium	Very High	Very High
Exuberant	High	Medium	High	High
Energetic	Very High	Medium	Medium	High
Frantic	High	Very High	Low	Very High
Anxious/Sad	Medium	Very Low	Very Low	Low
Depression	Low	Low	Low	Low
Calm	Very Low	Very Low	Medium	Very Low
Contentment	Low	Low	High	Low

Figure 1: Mood Classification according to audio features

Existing music recommender systems rely heavily on audio features, overlooking the emotional analysis of song lyrics.

2.2. Lyrics to Mood

In [3], XLNet, a large bidirectional transformer model is used for music emotion analysis only using lyrics. This model categorizes songs based on Russell's Valence-Arousal (V-A) circumplex model, illustrating emotions in a two-dimensional space (Figure 2).

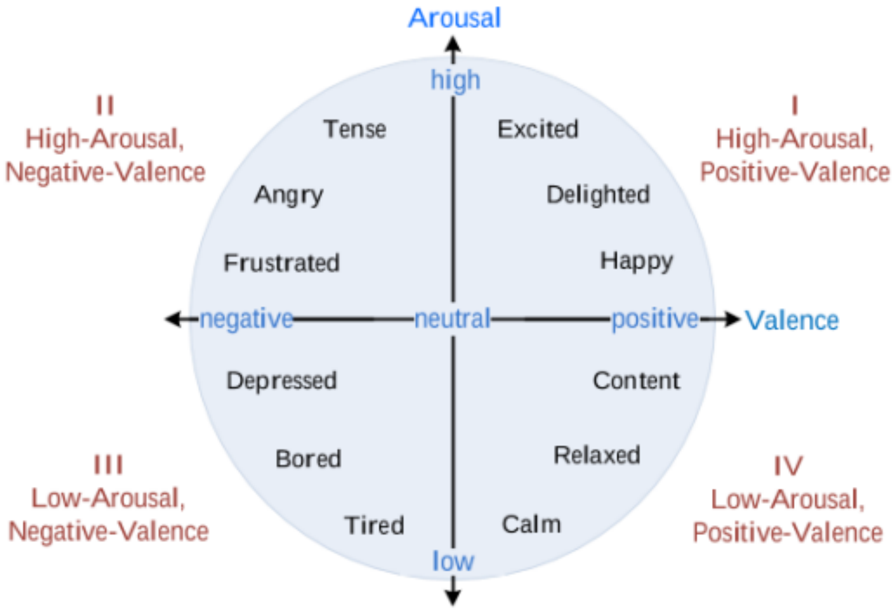


Figure 2: Core emotions established in the circumplex model.

The study in paper [4] introduces a multi-modal transformer-based approach that combines audio and textual data extracted from song lyrics to comprehensively interpret the emotional intricacies in music. Utilizing transformer-based models within a multi-modal framework, the paper suggests an advanced system for improved music mood classification. This system demonstrates superior performance compared to models lacking attention and those relying solely on either audio or lyrical features.

3. Source of Data and data processing

We've extensively explored available datasets for audio features and lyrics. Our main sources include the Million Song Dataset (MSD) and its associated datasets Last.fm and Musixmatch.

We will 1% subset of the extensive 280GB Million Song Dataset, containing diverse audio features tailored to our project's needs. Each song's tags, encompassing various moods and genres, are stored separately, necessitating linkage queries. Last.fm's extensive tag-to-song mapping surpasses typical 4-quadrant mood tags. We plan to categorize these varied emotional tags into 4 quadrants after a comprehensive literature review.

The Musixmatch dataset presents lyrics in a Bag of Words (BoW) format. However, we'll scrape lyrics from the internet to ensure a sequential arrangement, crucial for effective use of BERT or GPT2 embeddings. Our accuracy in tagging will be validated by aligning them with the established four quadrants for consistency.

4. Model Architecture

Our proposed model architecture is a multi-modal system comprising of CNN and GPT2 models. The CNN trains the acoustic features, while transfer learning with GPT2 handles the lyrics. After combining the models, integrated features are processed to classify moods into categories such as Happy, Sad, Angry, and Relaxed, as depicted in Figure 2.

GPT-2 (Generative Pre-trained Transformer) enhances text understanding and emotion prediction. By leveraging GPT-2, our model predicts emotions from text inputs, broadening the emotional classification alongside musical features in the dataset.

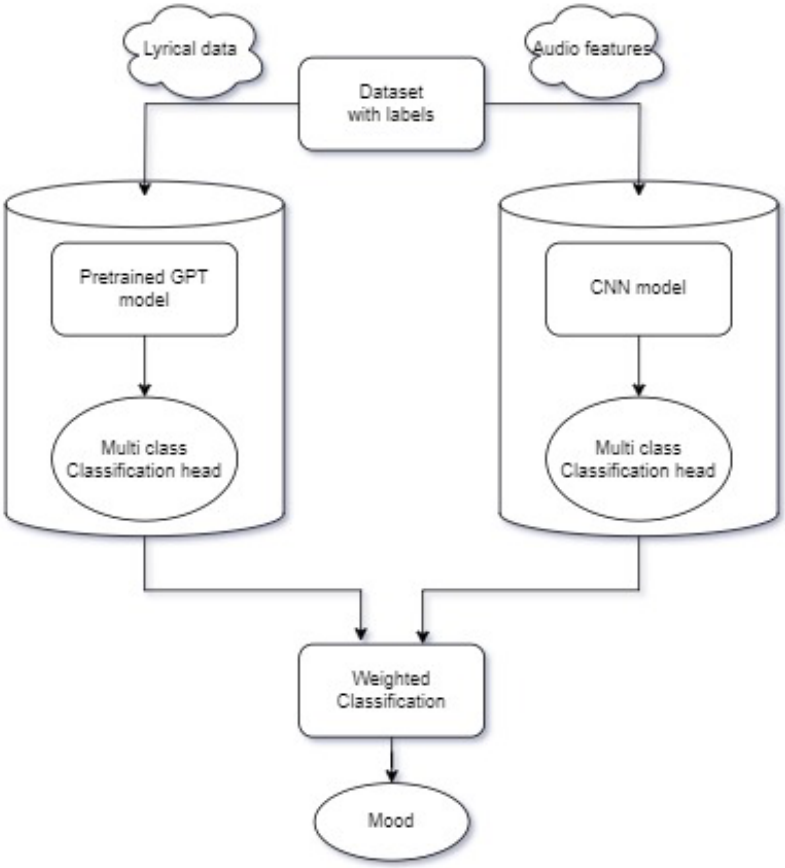


Figure 3: Proposed Architecture

5. Baseline Model and Comparisons

The project's baseline model for comparison is the paper [4], which evaluates results using a combination of audio and lyrical data. The paper presents performance metrics (Recall, Precision, F1 score) and accuracy for SVM, CNN, Transformer, and a combination of CNN + Transformer models. We aim to use this paper's findings as a benchmark for our own comparison. We will also try to do a qualitative analysis of the results to get a deeper understanding of how the model performs.

6. Plan

The tasks ahead of this project are primarily revolving around three major phases, namely the dataset, model implementation and then experimentation. The proposed plan for the project is tabulated in Table 1.

Phase	**Time**
Dataset Compilation – combining musical and lyrical data with labels	7 days (1 week)
Model Implementation – defining and training CNN and GPT2 models on the created dataset.	14 days (2 weeks)
Weighted Lyrics Mood Classification - this task experiments with model outcomes and combines them using weighted analysis to assess lyrics' significance in song mood classification.	7 days (1 week)

Table 1: Proposed Plan of Action with an estimate time

6.1. Proposed Timeline

1. **Projected duration:** Approximately 4 weeks (Nov 6th - Dec 3rd)
2. **Approach:** Tasks to be approached mostly in parallel.

6.2. Approach for Each Phase

1. **Phase 1:** One team member focusing on combining labels with music features, while another extracts lyrical data.
2. **Phase 2:** One team member works on the CNN model, while the other trains the GPT2 model.
3. **Phase 3:** Conducting experiments and comparative analyses with base models as required.

7. Risks

Following are the potential risks that we might encounter in the project:

1. **Limited Lyrics Access:** Copyright constraints may restrict gathering complete lyrics, resulting in a smaller dataset or mismatched audio-lyric pairs, causing uncertain performance effects.
2. **Audio Feature Reliability:** Pre-existing audio features from the dataset might underperform, leading us to rely solely on lyrical data in case of poor performance.
3. **Dataset Size and Complexity:** The large size of individual lyrics could cause computational challenges during model training. We'll explore optimization methods or reduce the dataset size if necessary.

8. References

1. Yang, Dan & Lee, Won-Sook. (2010). Music Emotion Identification from Lyrics. 624 - 629. 10.1109/ISM.2009.123
2. Lie Lu, D. Liu and Hong-Jiang Zhang, "Automatic mood detection and tracking of music audio signals," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 5-18, Jan. 2006, doi: 10.1109/TSA.2005.860344.
3. Agrawal, Y., Shanker, R.G.R., Alluri, V. (2021). Transformer-Based Approach Towards Music Emotion Recognition from Lyrics. In: Hiemstra, D., Moens, MF., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science (), vol 12657. Springer, Cham.
4. 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.