CS230

LyricNet: Multimodal Deep Learning for Emotion, Genre, and Lyric Continuation in Music (Natural Language Processing / Speech Recognition)

David Maemoto (dmaemoto)
Department of Computer Science
Stanford University
davidmaemoto@stanford.edu

Ethan Harianto (ethanhhr)
Department of Electrical Engineering
Stanford University
ethanhhr@stanford.edu

Sarah Dong (sarahjd)
Department of Computer Science
Stanford University
sarahjd@stanford.edu

1 Problem Statement and Motivation

Modern music analysis often focuses on audio signals and listening patterns but overlooks the emotional and semantic depth in lyrics. This project builds a multimodal deep learning model that integrates lyrics and audio features to predict a song's emotion, genre, and popularity and cluster songs by mood or vibe. The goal is to bridge language and sound to capture how they jointly convey emotion. This approach enables more emotionally grounded analysis of music and can extend to lyric generation conditioned on mood or genre for creative applications.

2 Project Challenges

The main challenges involve aligning lyrics and audio data from different sources, which can cause duplication or missing entries. Label imbalance across emotion, genre, and popularity may bias results, and the subjective nature of emotion adds noise. Maintaining a large, diverse dataset after preprocessing is essential for stable performance.

3 Dataset and Data Collection

Our dataset combines the Spotify Web API and the 900K Spotify Songs with Lyrics, Emotions & More dataset from Kaggle. The Spotify API provides metadata and audio features (valence, energy, danceability), while the Kaggle dataset adds lyrics and emotion labels for multimodal training. We will collect data by querying the Spotify API for track features and metadata, then merge it with the Kaggle dataset using Spotify track IDs to align songs and build a unified dataset for analysis.

4 Proposed Methodology

We propose a multimodal deep learning model that combines lyrics and audio features to learn unified representations of a song's emotion and style. The model processes (1) lyric text, encoded with a BERT-based model for semantic and emotional meaning, and (2) Spotify audio features (valence, energy, danceability) as numerical vectors. These modalities are fused through a CNN-based network

that models relationships between sound and language. While we will build on existing BERT and CNN modules, we will modify them by adding shared projection layers and a joint contrastive loss to align audio and text embeddings in a common latent space. The model predicts emotion, genre, and popularity, performs cosine-based mood clustering, and enables mood-consistent lyric generation through embedding-guided decoding.

5 Background and Related Work

We build on prior work in music emotion recognition and multimodal learning combining lyrics and audio. "Developing a Benchmark for Emotional Analysis of Music" (Aljanaki et al. 2017) and "Multi-Modal Music Emotion Recognition" (Panda et al. 2013) introduced early multimodal benchmarks. Yang et al. (2008) advanced fusion methods, and Pyrovolakis et al. (2022) apply deep learning for cross-modal mood detection. These studies inform our use of transformer lyric embeddings and Spotify audio features to model emotion and style.

6 Evaluation Plan (Quantitative and Qualitative)

We will evaluate our model both quantitatively and qualitatively for accuracy, clustering, and generation quality. Quantitatively, we will use accuracy, F1-score, and ROC-AUC for emotion and genre classification, and RMSE, MAE, and R^2 for popularity prediction. For lyric continuation, we will assess fluency (perplexity, cross-entropy) and semantic similarity with BERTScore or BLEU, while mood clustering will be measured using Silhouette scores and cosine similarity. Qualitatively, we will visualize cluster plots with t-SNE or UMAP and assess playlist coherence and lyric continuations for emotional consistency and alignment with predicted mood or genre.

References

[1] Aljanaki, Anna, et al. "Developing a Benchmark for Emotional Analysis of Music." *PLoS ONE*, vol. 12, no. 3, 2017, e0173392. Public Library of Science, https://doi.org/10.1371/journal.pone.0173392

[2] Panda, R., et al. "Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis." *CMMR 2013*, Centre for Informatics and Systems of the University of Coimbra (CISUC), 2013

[3] Pyrovolakis, Konstantinos, et al. "Multi-Modal Song Mood Detection with Deep Learning." *Sensors*, vol. 22, no. 3, 2022, p. 1065. MDPI, https://doi.org/10.3390/s22031065

[4] Yang, YH., et al. (2008). Toward Multi-modal Music Emotion Classification. In: Huang, YM.R., et al. *Advances in Multimedia Information Processing* - PCM 2008. PCM 2008. *Lecture Notes in Computer Science,* vol 5353. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-89796-5_8

TA Meeting & Proposal Form

We attended our TA meeting with Cody Ho, and we have also submitted our Project Proposal through the Google Form.