Jack Blake
PM5: Multimodal Music Genre Classification
12/18/2025
https://github.com/halftimejack/multimodal_music_genre_classification

Problem Statement

Genre labels are essential tools for organization and retrieval within massive music databases like Spotify. However, the sheer scale of these libraries makes manual labeling impractical, necessitating robust automated classification systems. While many existing models take a unimodal approach, only considering the raw text or audio to make the prediction, this project investigates a multimodal framework to enhance predictive performance. By integrating lyrics, chord progressions, and rhyme schemes, this study evaluates the distinct contribution of each modality and identifies potential synergies between them. Beyond classification accuracy, this work aims to uncover insights about the structural definitions of specific musical genres.

Approach

To integrate the three modalities, this project utilizes an intermediate fusion architecture consisting of parallel transformer branches. Unlike early fusion, which combines raw inputs, or late fusion, which aggregates final predictions, intermediate fusion concatenates the learned representations immediately prior to the classification layer. This allows the model to capture non-linear interactions between modalities while maintaining modality-specific feature extraction.

The lyrics branch consists of a pre-trained DistilBERT base model, fine-tuned to classify genres based on the semantics of the lyrics. The chords branch consists of a custom RoBERTa encoder trained from scratch. To avoid overfitting on the smaller chord vocabulary, the architecture was scaled down to 6 layers, 8 attention heads, and a hidden size of 512, with a vocabulary size of 5000. The rhyme branch mirrors this custom RoBERTa architecture but utilizes a vocabulary size of 1,000 to accommodate the limited set of rhyme scheme tokens.

For fusion, the final [CLS] token from each branch is concatenated into a single multimodal feature vector, which is passed through a linear layer to project the fused features onto the target genres. Model performance was evaluated using the macro-averaged F1 score, with the best-performing checkpoint selected for final analysis.

Data

The data for this project was aggregated from two primary sources. The textual data comes from the Genius Song Lyrics dataset [1], consisting of 5.1 million entries scraped from genius.com. This dataset contains raw lyrics, genre labels, and metadata including title, artist, and language. To capture structural features, the dataset was augmented with rhyme schemes generated by analyzing the phonetic pronunciation of line-ending words with the CMU Pronouncing Dictionary [2].

Harmonic data was sourced from the Chordonomicon dataset [3], comprising approximately 680,000 chord sequences with associated genre labels and Spotify IDs. The dataset was enriched with standardized title and artist metadata retrieved via the Spotify Web API [4] to make it compatible to merge with the lyrics dataset.

To construct the final experimental dataset, an inner join was performed between the lyrics and chord datasets based on the standardized artist and title metadata. This resulted in approximately 200,000 songs which had full lyrical and harmonic data. In all datasets, the label distribution is heavily skewed towards pop and rock, necessitating stratification before undertaking experiments.

Preprocessing

Data preprocessing was performed in multiple stages to standardize the input for each modality, using the Python pandas and re libraries. The lyrics dataset was filtered to remove non-English songs, genres marked as miscellaneous, and null data. Text cleaning was performed using regular expressions to remove bracketed song metadata (eg., [Chorus], [Verse 1]) and normalize excess whitespace. For the scale analysis experiments, a stratified subset of approximately 430,000 songs was generated to ensure genre balance within the text-only and rhyme-only domains.

The chord dataset was filtered to remove null data and duplicate entries. Songs with genres that could not be easily mapped to the 5 target genres in the lyrics dataset were excluded. Regular expressions were employed to clean the chord sequences, removing angle-bracketed metadata and standardizing spacing. A stratified chord-only dataset of approximately 37,000 songs was created for the scale experiments on chords-only data.

To facilitate the intersection of the two datasets, a strict normalization pipeline was applied to the artist and title columns. Strings were lowercased, stripped of whitespace, and cleaned of parenthetical or bracketed sections (e.g., [Radio Edit]). This ensured that minor formatting differences did not prevent successful joining. The final merged corpus was stratified to create the primary 20,000-sample balanced dataset.

Lyrics were tokenized using the pre-trained DistilBertTokenizer. Tokenization on rhyme scheme labels and chord labels were done on word boundaries using custom WordLevel tokenizers.

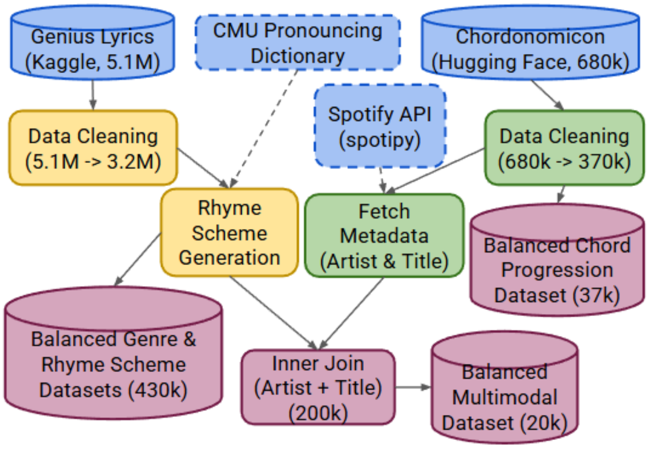
Genius Lyrics (Kaggle, 5.1M)
CMU Pronouncing Dictionary
Chordonomicon (Hugging Face, 680k)
Spotify API (spotipy)
Data Cleaning (5.1M -> 3.2M)
Data Cleaning (680k -> 370k)
Rhyme Scheme Generation
Fetch Metadata (Artist & Title)
Balanced Chord Progression Dataset (37k)
Balanced Genre & Rhyme Scheme Datasets (430k)
Inner Join (Artist + Title) (200k)
Balanced Multimodal Dataset (20k)

Figure 1: Data Processing Pipeline

Experiments and Results

The primary experiment consisted of a comprehensive ablation study to isolate the predictive contribution of each modality. The primary experiment consisted of a comprehensive ablation study to isolate the predictive contribution of each modality. Prior to fusion, each modality was evaluated individually to optimize branch-specific hyperparameters. To resolve conflicting optimization requirements during joint training, the most conservative settings (lowest learning rate and batch size) were adopted for the fusion model. All models were evaluated against a random baseline using accuracy and macro-averaged F1 score.

Model	Modalities	Accuracy	F1
Baseline (Random)	0	0.2	0.2
Rhyme only	1	0.364	0.332
Chords only	1	0.38	0.379
Text only	1	0.531	0.533
Chords and Rhyme	2	0.454	0.461
Text and Rhyme	2	0.547	0.546
Text and Chords	2	0.567	0.57
Full Fusion	3	0.574	0.575

Table 1: Ablation Study Results

As shown in Table 1, every added modality resulted in a strict increase in model performance, with the full fusion model achieving the best results (0.575 F1). This trend supports the conclusion that semantics, harmony, and structure provide complimentary information for genre classification.

A secondary scale study was conducted to assess the trade-off between feature complexity and dataset size. Unimodal models were trained on the maximally available balanced datasets for lyrics (430k) and chords (37k) to determine if raw data volume could outperform the architecturally superior but data-constrained fusion model.

Model	Max Balanced Dataset	F1
Chords Only (Max)	37,000	0.393
Rhyme Only (Max)	430,000	0.42
Full Fusion	20,000	0.575
Text Only (Max)	430,000	0.632

Table 2: Scale Study Results

While the fusion model is highly data-efficient (0.575 F1 on only 20k samples), the Text Only (Max) model achieved the highest overall performance (0,632 F1). This demonstrates that while multimodal features add significant predictive power per sample, sacrificing volume to produce quality multimodal data can hinder results.

Analysis

The experimental results verify that musical genre is a composite of semantics, harmony, and structure. Model performance improved with every added modality, demonstrating that each feature captures complimentary information that the others miss. Combining lyrics with chords was the strongest example of this effect, boosting individual F1 scores of 0.533 and 0.379 to a fused F1 score of 0.57. A key finding regarding redundancy was revealed when comparing bimodal pairings. Adding rhyme to the text model did not significantly improve the model compared to the effect of adding chords. This suggests that textual structure is partly implicit in the lyrics themselves (e.g., through line breaks and vocabulary), whereas harmonic progression offers a truly distinct signal. Alternatively, the current method of strictly evaluating rhyme pairs overlooked all near rhymes and slant rhymes, possibly degrading the rhyme modality into one more resembling repetition. In either case, harmony appears to be the more valuable complementary feature to semantics.

One of the most notable findings was the performance of the No-Text Model (Chords + Rhyme). With an F1 score of 0.461, it more than doubled the performance of the random baseline. This demonstrates that genre classification does not strictly require semantic data to be effective. Even without access to a single word of the lyrics, the combination of harmonic mood and sonic structure provides a robust, language-independent signal for genre definition.

The scale study highlighted a critical trade-off between feature complexity and data volume. While the fusion model is the superior architecture per-sample, the Text-Only model achieved the highest absolute score (0.632) simply because it had 21 times more training data. However, the Fusion model proved exceptionally efficient, achieving comparable results (0.575) with only 20k samples. This demonstrates that investing in complex feature engineering can largely compensate for data scarcity.

The Rhyme-Only model peaked on rap (F1: 0.59), significantly outperforming other genres. This demonstrates unlike melodic genres, rap is structurally distinct enough to be identified by its rhyme scheme alone.The Chord-Only model performed best on these country and R&B (F1: 0.47 & 0.45), validating their reliance on specific harmonic signatures, such as standardized I-IV-V progressions for country and extended chord voicings (7ths/9ths) for R&B.

Pop was consistently the weakest performer across all models. Error analysis suggests that pop acts as a "catch-all" class, borrowing harmonic complexity from R&B and lyrical themes from country, lacking the distinct signature required for accurate classification. Confusion matrices revealed a high rate of misclassification between pop and rock. Without audio spectrograms to detect instrumentation (e.g., distorted guitars vs. synthesizers), these genres appear compositionally similar, sharing identical verse-chorus structures and simple triads.

Related Work

Prior research in music classification has shifted toward multimodal architectures. Wadhwa and Mukherjee (2021) demonstrated the efficacy of this approach by fusing lyrics with audio spectrograms, utilizing co-attention mechanisms to achieve 90.4% accuracy [5]. This project builds on this multimodal foundation but pivots from audio fusion (combining semantics with production timbre) to symbolic fusion (combining semantics with compositional theory).

This approach is made possible by the recent release of the Chordonomicon dataset [3]. This corpus represents a massive data engineering undertaking, constructed by scraping and parsing over 680,000 crowdsourced chord charts from Ultimate Guitar. The authors successfully converted unstructured web content into standardized harmonic sequences linked with Spotify metadata. This work builds directly on this foundation, leveraging their large-scale aggregation to treat chord progressions as a structured language sequence equivalent to lyrics.

Conclusion

This project successfully demonstrated that music genre classification can be effectively approached through the deconstruction of symbolic features. By engineering a custom pipeline that fuses DistilBERT (lyrics) with specialized transformers for harmony and structure, the final multimodal model achieved an F1 score of 0.575, significantly outperforming the unimodal baselines and validating the hypothesis that genre is a composite of semantics, harmony, and structure.

The results highlight distinct roles for each modality: while lyrics provide the dominant signal, harmonic progression and rhyme scheme serve as critical differentiators for specific genres like country and rap. Notably, the No-Text model (Chords + Rhyme) achieved more than double the accuracy of the random baseline, proving that genre identity is compositionally robust even in the absence of semantic content. Furthermore, the scale analysis revealed a compelling efficiency trade-off: while massive unimodal datasets currently yield the highest absolute performance, multimodal fusion allows for competitive modeling using only a fraction of the data.

Ultimately, this work establishes that the fusion of symbolic features offers a viable, computationally efficient alternative to traditional audio signal processing, capable of capturing the essence of musical style.

Future Work

Future improvements to this work would look to integrate more modalities to this fusion architecture, such as audio data or symbolic representations extracted from raw audio data. The weakest of this project's modalities, rhyme scheme, could also be improved by selecting a method that takes into account slant rhymes or near rhymes. These changes would likely enhance performance in the model's weak area of distinguishing pop and rock.

References

[1] https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] https://huggingface.co/datasets/ailsntua/Chordonomicon

[4] https://developer.spotify.com/documentation/web-api

[5] L. Wadhwa and P. Mukherjee, "Music genre classification using multi-modal deep learning based fusion," 2021 Grace Hopper Celebration India (GHCI), Bangalore, India, 2021, pp. 1-5, doi: 10.1109/GHCI50508.2021.9514020.