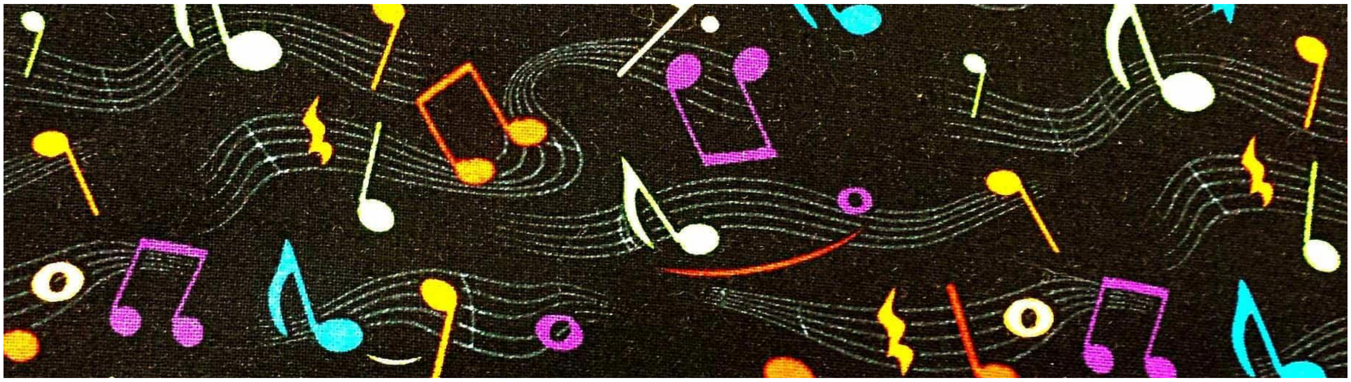


TuneType

Predicting Song Genres using Audio and Lyrics

Overview

This project builds and compares machine learning models to classify songs into ten major music genres using a large Spotify dataset containing audio features and song lyrics. An audio-only baseline model using Spotify's engineered acoustic features established initial performance and revealed strong separability for some genres (such as Classical and Hip-Hop) alongside substantial overlap among others (notably Rock and Pop). A lyrics-only model based on TF-IDF representations captured genre-specific linguistic patterns and improved performance for language-driven genres, while highlighting limitations for genres with sparse or theatrical lyrics.

A multimodal model combining scaled audio features with TF-IDF lyric representations achieved the best overall performance, demonstrating that acoustic and linguistic information provide complementary signals for genre classification. Hyperparameters were tuned using stratified cross-validation on a representative subset of the data to balance computational efficiency and performance. Model interpretability analyses examined audio feature coefficients, top genre-specific lyric terms, and confusion matrices, revealing meaningful genre characteristics and consistent ambiguities between closely related styles. This project demonstrates an end-to-end machine learning workflow on large real-world data, emphasizing multimodal learning, interpretability, and principled model evaluation.

Introduction

Background

Music genre classification is a foundational problem in music information retrieval, with applications in recommendation systems, catalog organization, and music discovery. Modern streaming platforms such as Spotify provide rich, multimodal data describing songs, including engineered audio features derived from signal processing as well as textual information in the form of song lyrics. While genre classification has traditionally relied on either acoustic features or metadata alone, genres are inherently defined by a combination of musical structure and lyrical content, motivating the use of multimodal approaches.

The availability of a large, cleaned Spotify dataset with standardized audio features and aligned English-language lyrics enables systematic investigation of how different data modalities contribute to genre discrimination at scale. This project was motivated by the question of whether combining audio and lyrical information meaningfully improves genre classification performance over single-modality models, and how these complementary signals can be interpreted to better understand the acoustic and linguistic characteristics that distinguish musical genres.

Objectives

a. To build and compare machine learning models for music genre classification using Spotify audio features, song lyrics, and a combination of both modalities.

b. To evaluate the benefit of multimodal learning, specifically whether integrating acoustic and lyrical information improves genre classification performance compared to audio-only or lyrics-only models.

c. To interpret model predictions and errors, identifying key audio features and lyrical patterns that characterize different genres and analyzing systematic confusions between closely related genres.

Problem

Music genre classification is a challenging task because genres are not defined by sound or lyrics alone, but by a combination of musical structure, production style, and lyrical themes. Many existing approaches rely on a single modality, either audio features or metadata, leading to inconsistent performance and confusion between closely related genres such as Rock and Pop or Folk and Country. This makes it difficult to accurately organize large music catalogs and limits the effectiveness of recommendation and discovery systems.

Recommendations

1. Use a multimodal approach for best performance.
2. Prioritize macro-F1 (not just accuracy) and handle imbalance.
3. Include interpretability in deployment decisions.

Additional Information

TF-IDF stands for Term Frequency–Inverse Document Frequency. It is a method used to convert text (like song lyrics) into numerical features that machine learning models can understand.

- Term Frequency (TF) measures how often a word appears in a document. Words that appear frequently in a song's lyrics get higher TF values.
- Inverse Document Frequency (IDF) down-weights words that appear in many documents (such as "the" or "love"), because they are less useful for distinguishing between genres.

TF-IDF assigns high values to words that are frequent in a specific song but rare across the entire dataset, making them more informative for classification. In this project, TF-IDF was used to represent lyrics in a way that highlights genre-specific vocabulary and themes, allowing the model to learn linguistic patterns associated with different music genres.

Analysis

Research methods

This project employed a supervised machine learning approach to music genre classification using a large, cleaned Spotify dataset containing audio features and song lyrics. Data preprocessing included filtering for English-language tracks, removing missing or invalid entries, and selecting a standardized set of Spotify audio features alongside cleaned lyrical text. The target variable was a mapped main genre label comprising ten broad genre categories.

Data Preprocessing

The dataset was cleaned by removing songs with missing genre labels or lyrics and filtering out tracks with very short lyrics. A fixed set of Spotify audio features was selected and standardized so all numerical inputs were on the same scale. Song lyrics were converted into numerical features using TF-IDF with basic text cleaning and stop-word removal. Genre labels were encoded as numeric classes, and stratified train–test splits were used to preserve genre distributions during modeling and evaluation.

Audio-Only Genre Classification

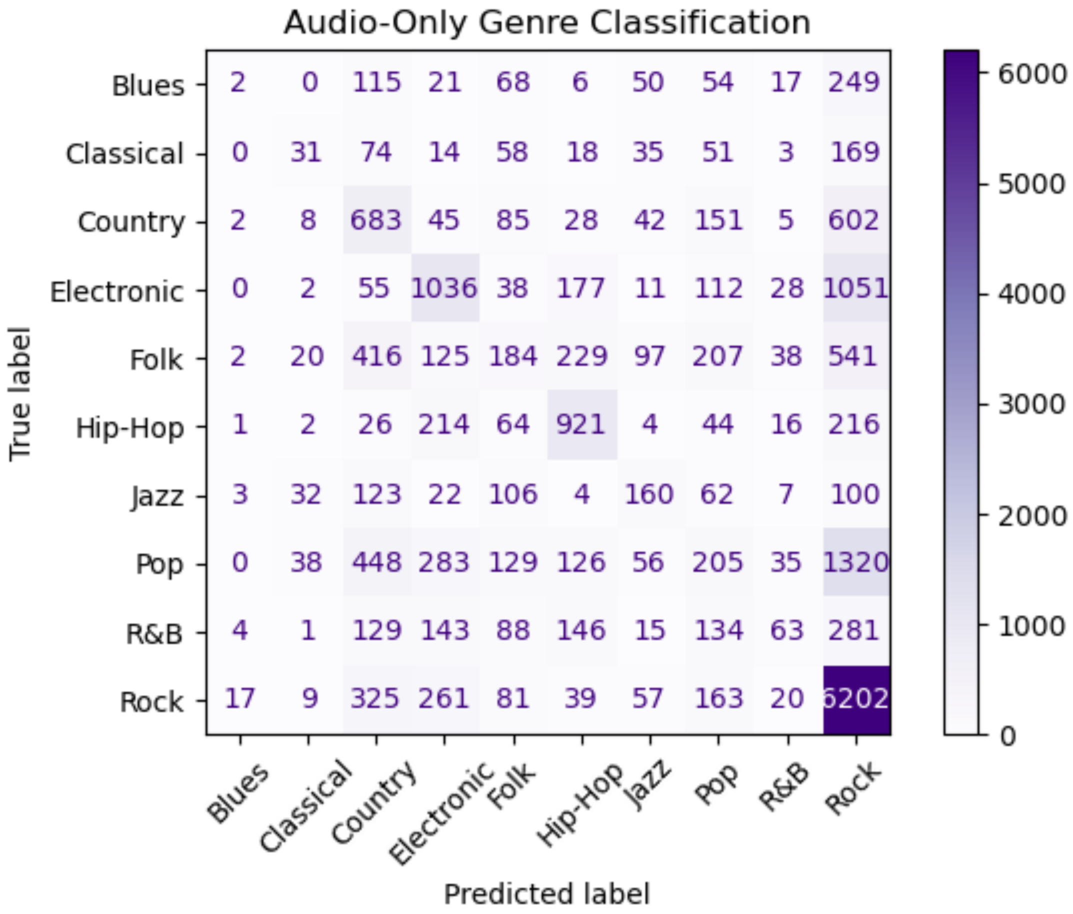
The audio-only genre classification model used Spotify's engineered audio features, such as danceability, energy, tempo, loudness, and valence, to predict a song's genre. These features were standardized and modeled using multinomial logistic regression as a baseline classifier. This approach demonstrated that certain genres with distinctive acoustic characteristics, particularly Classical, Electronic, and Hip-Hop, are more easily separable based on audio information alone.

However, the audio-only model showed consistent confusion between acoustically similar genres, including Rock and Pop as well as Folk and Country. While the model provided a strong and interpretable baseline, its limitations highlighted that audio features alone do not fully capture the stylistic and semantic differences between all genres, motivating the inclusion of lyrical information in subsequent models.

```
Accuracy: 0.47435

              precision    recall  f1-score   support

       Blues       0.06      0.00      0.01       582
   Classical       0.22      0.07      0.10       453
     Country       0.29      0.41      0.34      1651
  Electronic       0.48      0.41      0.44      2510
        Folk       0.20      0.10      0.13      1859
     Hip-Hop       0.54      0.61      0.58      1508
        Jazz       0.30      0.26      0.28       619
         Pop       0.17      0.08      0.11      2640
         R&B       0.27      0.06      0.10      1004
        Rock       0.58      0.86      0.69      7174

    accuracy                           0.47     20000
   macro avg       0.31      0.29      0.28     20000
weighted avg       0.40      0.47      0.42     20000
```

The audio-only model achieved an accuracy of 47%, showing that Spotify audio features contain useful information for genre classification, but are not sufficient on their own. The model performed best on Rock and Hip-Hop, which have strong and consistent acoustic patterns, and showed moderate performance for Electronic and Country. Rock in particular had very high recall, meaning the model frequently identified songs as Rock based on their sound.

However, the model struggled with genres such as Blues, Pop, R&B, Folk, and Classical, which had low recall and F1-scores. These genres tend to share similar acoustic characteristics or rely more on lyrical content and stylistic context than on sound alone. Overall, the results show that audio features provide a useful baseline but motivate the need for lyrics or multimodal models to improve genre classification performance.



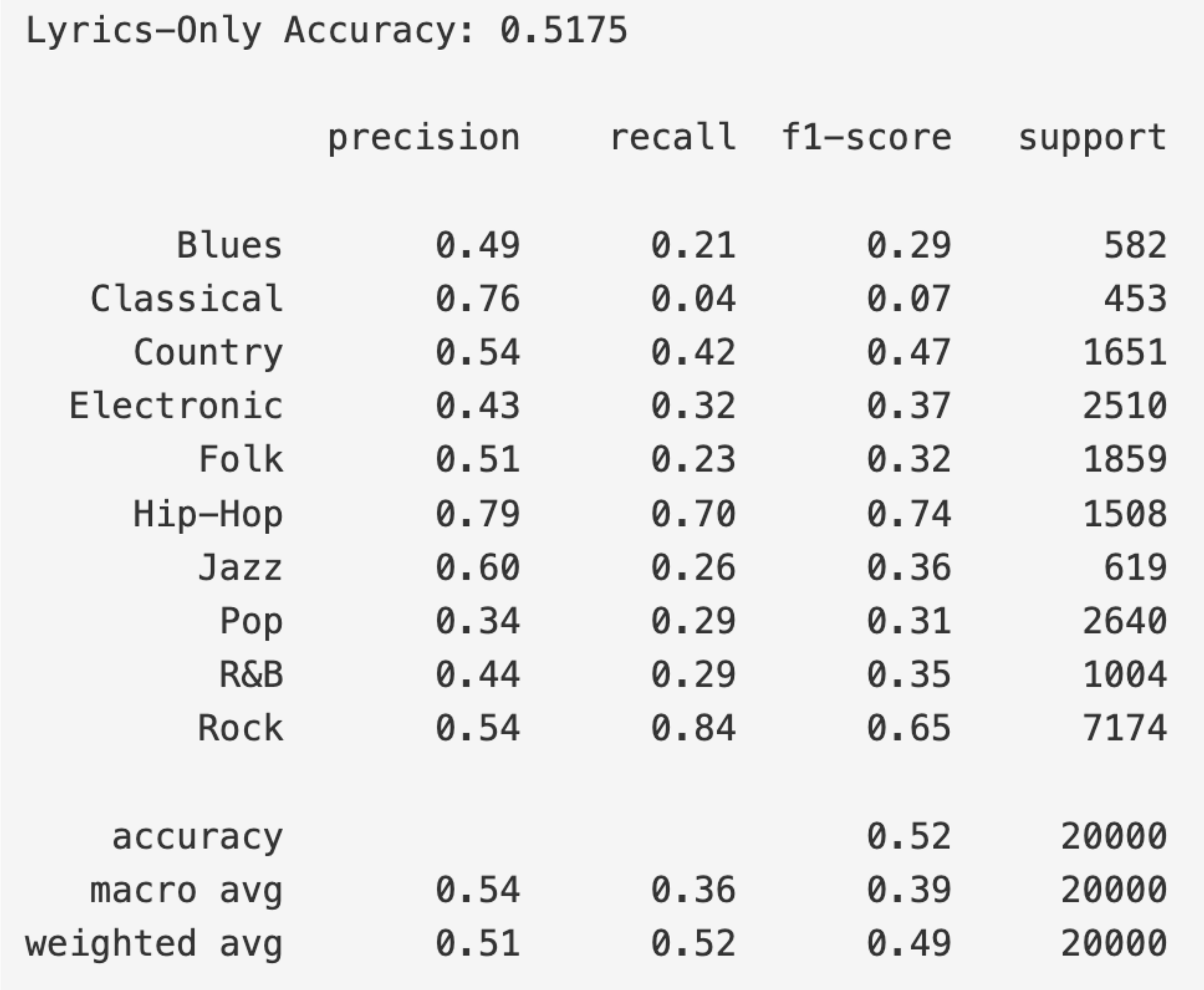
The confusion matrix shows that the audio-only model frequently predicts Rock, even when the true genre is something else. Many Blues, Pop, Folk, and Country songs are misclassified as Rock, indicating strong overlap in their acoustic features. This explains Rock's high recall but also shows that the model is biased toward this dominant genre.

Genres with clearer acoustic patterns, such as Hip-Hop and Electronic, are classified more accurately, with more correct predictions along the diagonal. In contrast, genres like Blues, Jazz, and Classical have very few correct predictions and are often confused with other styles. Overall, the confusion matrix highlights that audio features alone capture broad sound characteristics but struggle to distinguish between closely related genres, motivating the need for lyrics or multimodal models.

Lyrics-Only Genre Classification

The lyrics-only classification model used TF-IDF representations of song lyrics to predict genre, focusing on linguistic patterns rather than acoustic features. This model performed well for genres with strong and distinctive language use, such as Hip-Hop, Country, and Folk, where themes, vocabulary, and repetition provide clear genre signals. The results show that lyrics contain meaningful information for genre discrimination, particularly for genres defined by storytelling or cultural expression.

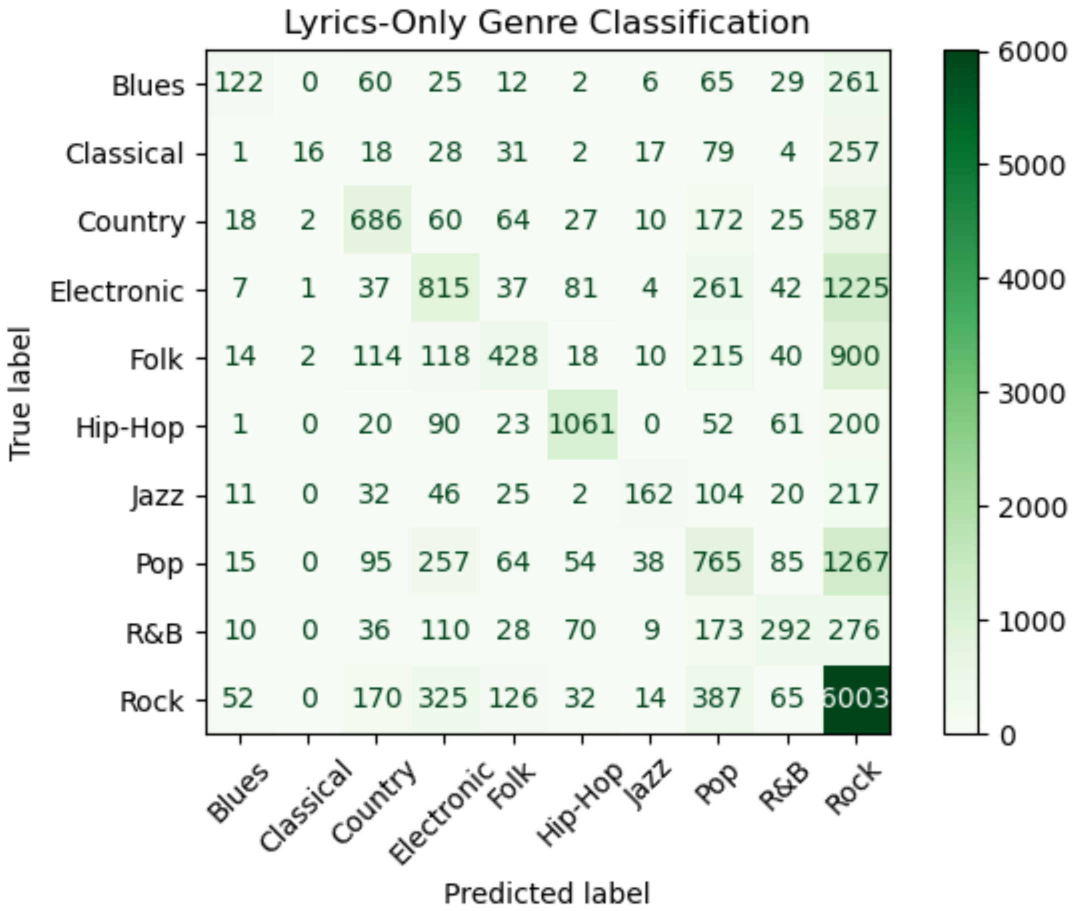
However, the lyrics-only model struggled with genres where lyrics are sparse, generic, or less central to genre identity, such as Classical, Electronic, and Jazz. Confusions between genres with similar language, including Pop and Rock, were common. Overall, the lyrics-only model complemented the audio-only approach by capturing semantic differences that audio features miss, but it was insufficient on its own, reinforcing the value of combining lyrics with audio features in a multimodal model.



```
Lyrics-Only Accuracy: 0.5175

              precision    recall  f1-score   support

       Blues       0.49      0.21      0.29       582
   Classical       0.76      0.04      0.07       453
     Country       0.54      0.42      0.47      1651
  Electronic       0.43      0.32      0.37      2510
        Folk       0.51      0.23      0.32      1859
     Hip-Hop       0.79      0.70      0.74      1508
        Jazz       0.60      0.26      0.36       619
         Pop       0.34      0.29      0.31      2640
         R&B       0.44      0.29      0.35      1004
        Rock       0.54      0.84      0.65      7174

    accuracy                           0.52     20000
   macro avg       0.54      0.36      0.39     20000
weighted avg       0.51      0.52      0.49     20000
```

The lyrics-only model achieved an accuracy of about 52%, performing better than the audio-only model and showing that lyrics are an important signal for genre classification. The model performed especially well for Hip-Hop and Rock, which have strong and distinctive lyrical patterns. Country and several other genres also showed moderate performance, indicating that language and themes help distinguish these styles.

However, the model struggled with Classical, which had very low recall, likely because many classical tracks have little or no meaningful lyrics. Overall, the results show that lyrics alone can capture important genre information but are not sufficient for all genres, reinforcing the benefit of combining lyrics with audio features.



The confusion matrix shows that the lyrics-only model classifies Hip-Hop and Rock particularly well, with strong diagonal values indicating many correct predictions. This reflects the distinctive vocabulary, themes, and structure commonly found in the lyrics of these genres.

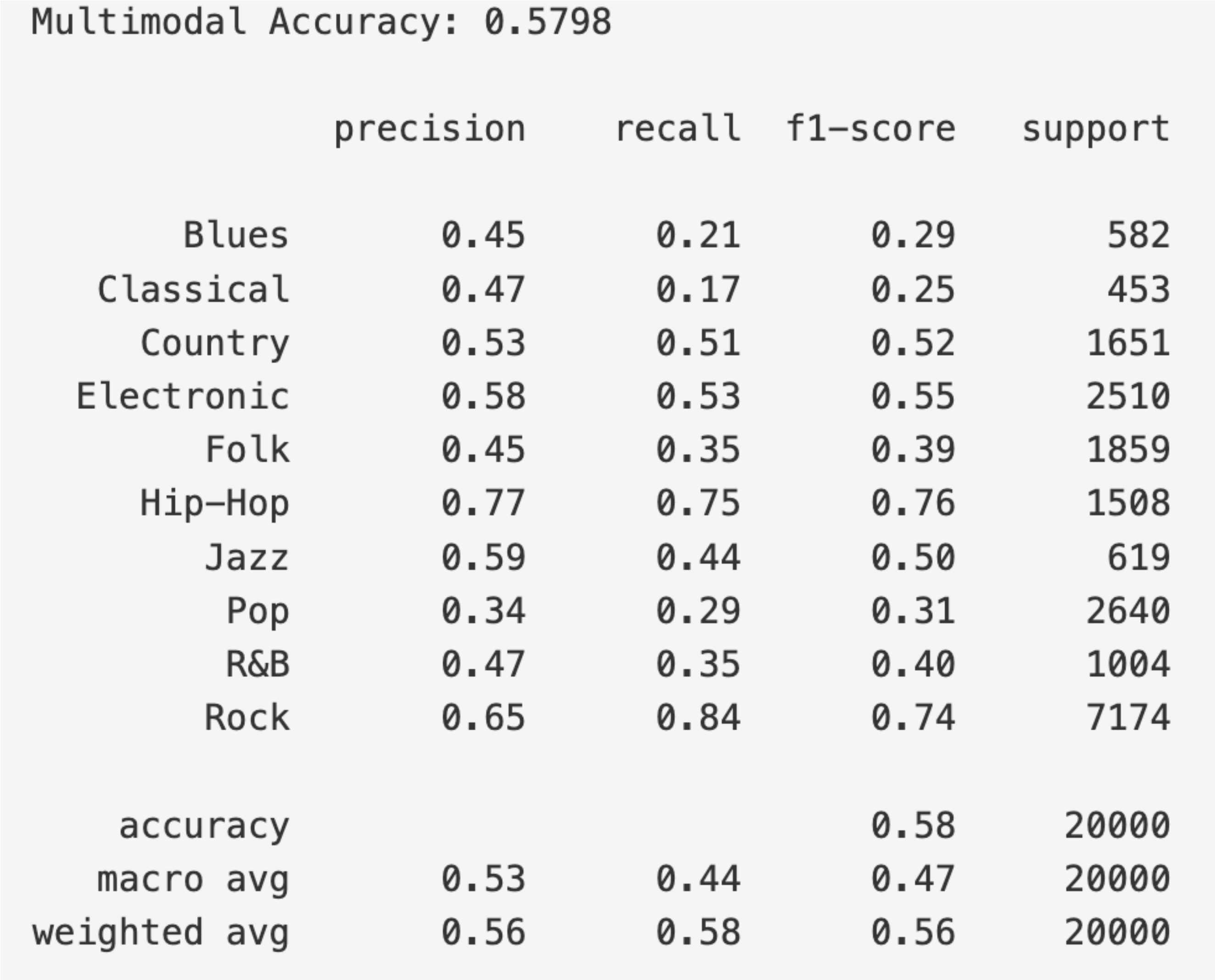
Genres such as Country, Pop, Electronic, and Folk show moderate performance but also substantial confusion with one another. This suggests overlap in lyrical themes and language, making them harder to separate using text alone. Pop in particular is frequently confused with Rock and Electronic, reflecting its broad and generic lyrical style.

Classical again shows very weak performance, with most classical tracks being misclassified as other genres, especially Rock and Pop, due to sparse or non-representative lyrics. Overall, the confusion matrix highlights that lyrics provide strong genre signals for some styles but struggle to distinguish genres with overlapping language or minimal lyrical content, reinforcing the need for a multimodal approach.

Multimodal (Audio + Lyrics) Genre Classification

The multimodal genre classification model combined Spotify audio features with TF-IDF representations of song lyrics to leverage both acoustic and linguistic information. This approach achieved the best overall performance among the models, improving classification accuracy and F1-scores across most genres compared to the audio-only and lyrics-only models. By integrating complementary signals, the model was better able to distinguish genres that are acoustically similar but lyrically distinct, and vice versa.

The multimodal model reduced many of the systematic confusions observed in the single-modality models, particularly between closely related genres such as Rock and Pop or Folk and Country. While some ambiguity remained for inherently overlapping genres, the combined approach provided more balanced performance across classes and demonstrated that genre is most effectively characterized through both sound and language.

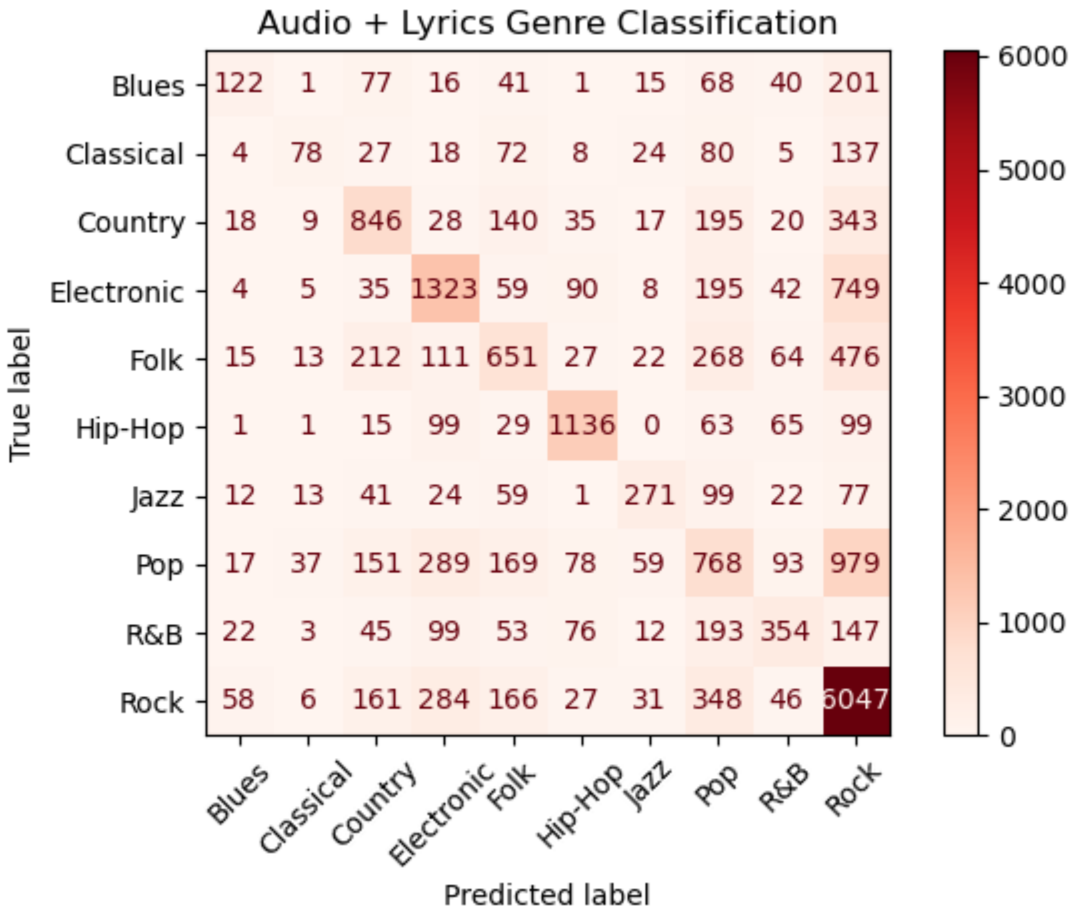


```
Multimodal Accuracy: 0.5798

              precision    recall  f1-score   support

       Blues       0.45      0.21      0.29       582
   Classical       0.47      0.17      0.25       453
     Country       0.53      0.51      0.52      1651
  Electronic       0.58      0.53      0.55      2510
        Folk       0.45      0.35      0.39      1859
     Hip-Hop       0.77      0.75      0.76      1508
        Jazz       0.59      0.44      0.50       619
         Pop       0.34      0.29      0.31      2640
         R&B       0.47      0.35      0.40      1004
        Rock       0.65      0.84      0.74      7174

    accuracy                           0.58     20000
   macro avg       0.53      0.44      0.47     20000
weighted avg       0.56      0.58      0.56     20000
```

The multimodal model achieved an overall accuracy of about 58%, the highest among all models tested, confirming that combining audio features with lyrical information improves genre classification performance. Both macro-F1 (0.47) and weighted F1 (0.56) increased compared to the audio-only and lyrics-only models, indicating more balanced performance across genres.

The model performed especially well for Hip-Hop and Rock, which showed high recall and F1-scores, reflecting the benefit of capturing both strong acoustic patterns and distinctive lyrical content. Electronic, Country, and Jazz also saw clear improvements compared to single-modality models, showing that the combined features help separate genres that were previously confused.

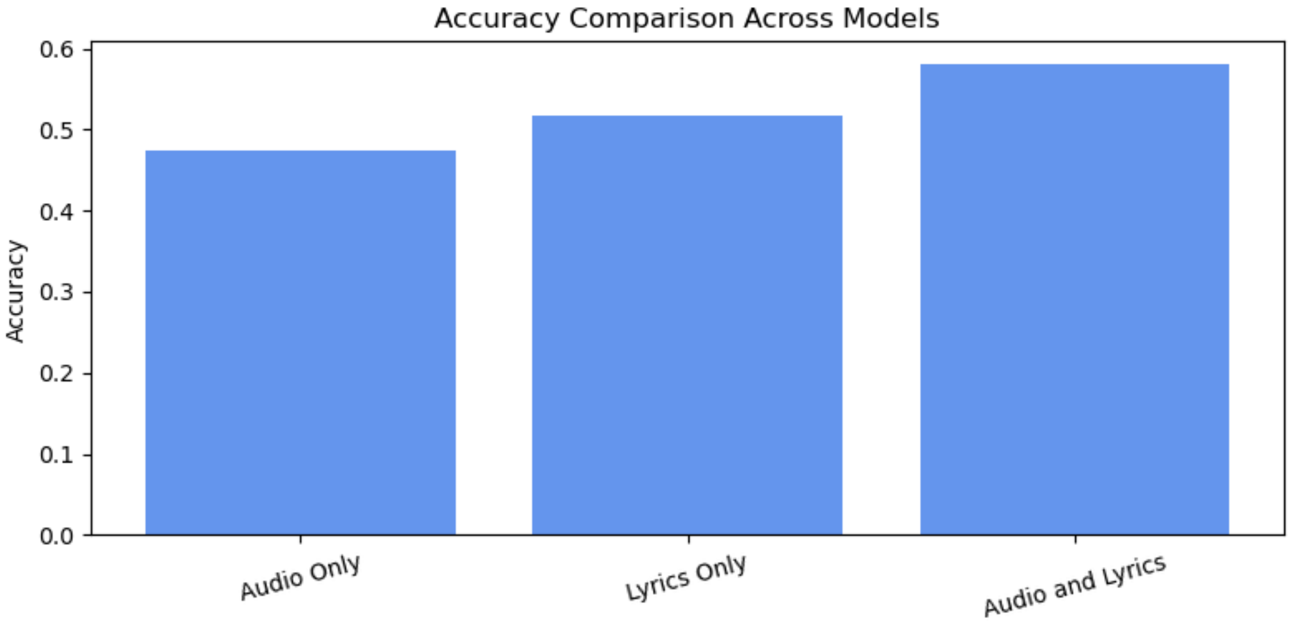
While genres such as Pop, Blues, and Classical remained challenging, their performance still improved relative to the audio-only model. Overall, these results demonstrate that genre classification benefits most from a multimodal approach, as audio and lyrics provide complementary information that reduces systematic confusion and leads to more reliable predictions.



True label \ Predicted label	Blues	Classical	Country	Electronic	Folk	Hip-Hop	Jazz	Pop	R&B	Rock
Blues	122	1	77	16	41	1	15	68	40	201
Classical	4	78	27	18	72	8	24	80	5	137
Country	18	9	846	28	140	35	17	195	20	343
Electronic	4	5	35	1323	59	90	8	195	42	749
Folk	15	13	212	111	651	27	22	268	64	476
Hip-Hop	1	1	15	99	29	1136	0	63	65	99
Jazz	12	13	41	24	59	1	271	99	22	77
Pop	17	37	151	289	169	78	59	768	93	979
R&B	22	3	45	99	53	76	12	193	354	147
Rock	58	6	161	284	166	27	31	348	46	6047

The multimodal confusion matrix shows that combining audio and lyrics improves genre classification across most genres. There are more correct predictions along the diagonal compared to the audio-only and lyrics-only models, especially for Hip-Hop, Rock, Electronic, and Country, which benefit from having both strong sound patterns and distinctive lyrical content.

Some confusion between similar genres, such as Pop and Rock or Country and Folk, still remains, but it is reduced compared to the single-modality models. Genres like Blues and Classical remain difficult, though their performance improves slightly. Overall, the confusion matrix confirms that using both audio and lyrics leads to more accurate and balanced genre predictions.

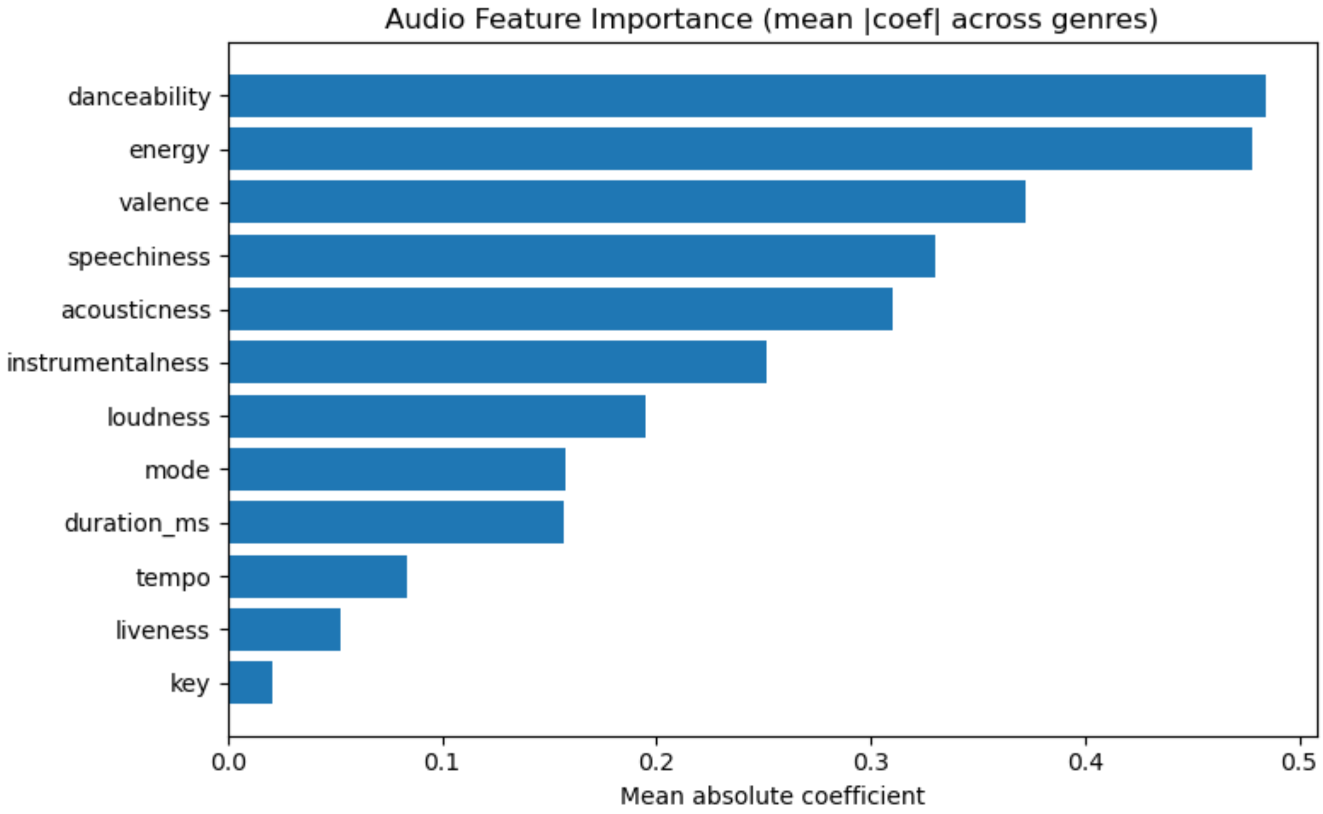


The bar chart shows a clear performance improvement as more information is added to the model. The audio-only model has the lowest accuracy (around 47%), indicating that sound features alone are not sufficient to fully distinguish between all genres. The lyrics-only model performs better (about 52%), showing that lyrical content provides stronger discriminative information for many genres.

The audio + lyrics (multimodal) model achieves the highest accuracy (about 58%), confirming that combining acoustic and lyrical features leads to the best overall performance. This comparison highlights that audio and lyrics capture different but complementary aspects of music genre, and that using both together results in more accurate and robust genre classification.

Interpretability

Interpretability analyses show that the models learn meaningful and genre-relevant patterns from both audio and lyrical features. Audio-based coefficients highlight clear acoustic distinctions for genres like Rock, Hip-Hop, and Electronic, while TF-IDF terms reveal genre-specific lyrical themes for language-driven genres such as Hip-Hop and Country. Confusion matrix analysis indicates that remaining errors occur primarily between closely related genres, reflecting genuine stylistic overlap rather than random misclassification.



This feature importance plot shows which audio features contribute most to genre classification on average across all genres. Danceability and energy are the most influential features, indicating that rhythm and intensity are key factors in distinguishing musical styles. Valence and speechiness also play important roles, reflecting differences in emotional tone and the presence of spoken vocals across genres.

Features such as acousticness and instrumentalness provide additional discriminative power, particularly for separating acoustic or instrumental genres from more produced styles. In contrast, key, liveness, and tempo have relatively low importance, suggesting they vary less systematically across genres or provide limited standalone information. Overall, the plot confirms that high-level rhythmic, energetic, and expressive characteristics are the most informative audio cues for genre classification.

Conclusion

This project evaluated music genre classification using audio features, song lyrics, and a combination of both modalities on a large Spotify dataset. The results show that audio features alone provide a useful but limited baseline, performing well for genres with strong acoustic signatures but struggling with stylistically overlapping genres. Lyrics-only models improved performance, particularly for language-driven genres, but were ineffective for genres with sparse or non-representative lyrics.

The multimodal model combining audio and lyrics achieved the best overall performance, delivering higher accuracy and more balanced results across genres. Interpretability analyses confirmed that the models relied on meaningful audio characteristics (such as energy and danceability) and genre-specific lyrical patterns, while confusion matrices revealed that remaining errors were primarily between closely related genres. Overall, the findings demonstrate that music genre is best captured through a combination of sound and language, and that multimodal learning provides a more robust and interpretable approach to genre classification than single-modality models.

References

Spotify. (n.d.). *Spotify Audio Features*. https://developer.spotify.com/documentation/web-api/reference/get-audio-features

— Source of engineered audio features such as danceability, energy, and valence used in this project.

Serkantysz. (2024). *550K Spotify Songs: Audio, Lyrics & Genres*. Kaggle. https://www.kaggle.com/datasets/serkantysz/550k-spotify-songs-audio-lyrics-and-genres

— Primary dataset used for model training and evaluation.

Salton, G., & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval*. Information Processing & Management, 24(5), 513–523.

— Foundational reference for TF-IDF text representation.

Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

— Machine learning library used for modeling, evaluation, and preprocessing.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Draft.

— Reference for text modeling and NLP concepts used in lyrics-based classification.