Lyrics and Metadata for Genre Classification: A Comparative Study of Transformer models

Candidates: 112, 104, 111

Abstract

Transformer-based encoder models have shown strong performance in genre classification tasks, yet decoders and the inclusion of metadata remains underexplored in this field. This project explores how well transformer-based models can classify music based on song lyrics. The models tested include two encoders, BERT and DistilBERT, and one decoder model, LLaMA 3.1 8B Instruct. Additionally, the study examines whether simple metadata, specifically release year and danceability, can improve model performance. We trained the encoder models on genre-labeled lyrics, both with and without metadata. The decoder model was evaluated using zero-shot prompting techniques. Our results show that the encoder models significantly outperformed the decoder model, with macro F1-scores reaching up to 0.48 compared to LLaMA's 0.13. Inclusion of metadata led to slight improvements in performance and more balanced predictions across the genre classes. While the decoder model LLaMA 3.1 performed poorly in this task, it showed some signs of potential.

1 Introduction

Music genre classification is an active research area within Music Information Retrieval (Li et al., 2023). While audio features are often used, lyrics can also reveal genre through language patterns and themes. Earlier research found lyrics less effective than audio (McKay et al., 2010), but the rise of transformer models like BERT; interest in lyric-based classification has grown as these models can capture deeper semantic meaning and context in text (Li et al., 2023)

In this project, we explored genre classification using lyrics, with and without simple metadata. We compared encoder models Bidirectional Encoder Representation from Transformers (BERT) and DistilBERT with the decoder model Large Language Model Meta AI 3.1 (LLaMA). BERT served as our baseline, and we tested whether lyrics alone are sufficient and whether metadata improves classification performance.

With this in mind, we have defined two research questions; To what extent are song lyrics alone sufficient for genre classification, and how does the inclusion of metadata affect model performance? How do different transformer architectures, encoder-based or decoder-based, perform on lyric based genre classification?

To explore these questions, we used a labeled dataset of lyrics and metadata, we trained each model with and without metadata and evaluated performance using macro F1-score.

Our results show that while the models capture some meaningful patterns in the lyrics, overall performance is lower than expected. However, we observed small improvements in performance when metadata is included, and notable differences in how the models behave across the classification task.

We begin with a review of related work on genre classifications. Then we present our dataset. Following this, we introduce the three models used in our study, and outline our methodology. We proceed by presenting and analyzing the results, including a discussion of model performance, limitations, and error analysis. Finally, we summarize our key findings and reflect on the potential for future work

2 Related work

Historical research on music genre classification has predominantly focused on either audio or lyrical features independently, with audio features historically receiving greater attention. Tsaptsinos (2017), revisited the potential of using lyrics for genre classification. He used a Hierarchical Attention Network (HAN) and showed that lyrics can provide valuable information when modeled with advanced neural networks. It was also noted

that research involving song lyrics has faced challenges from copyright restrictions, which may limit progress in this area.

Li et al. (2023) created a multimodal framework that combined audio and lyrics. They found that BERT outperformed older lyric representation methods like GloVe, Bag-of-Words, and HAN, achieving an F1-score of 0.74 using only lyrics. The main finding was that combining audio data and lyrics, the model performance went up to 0.87. While they did not directly test the same metadata as in this project, they demonstrated that including more contextual information in the input led to better results. (Li et al., 2023)

Akalp et al. (2021) compared BiLSTM, BERT, and DistilBERT for genre classification using song lyrics, based on a dataset with six music genres. They tested both single-label and multi-label classification. BERT gave the best results, with 77.63% accuracy for single-label classification. DistilBERT followed with 74.38%. Their findings show that transformer models outperformed the older recurrent model. While class imbalance was a significant challenge in the study, BERT and DistilBERT appeared more robust to this problem. (Akalp et al., 2021)

Martinez et al. (2024), demonstrate the effectiveness of DistiliBERT in accurately classifying music genres solely on song lyrics. They achieved comparatively good results, including an accuracy of 65% across five genres, highlighting the models capability to interpret linguistic patterns while being more lightweight and efficient than BERT. (Martinez et al., 2024)

Researchers at the Swedish Royal institute of technology used metadata like year, energy and valance to classify explicit music content. Their discoveries indicate that metadata can be utilized together with song lyrics to enhance the classification capabilities of models (Bergelid, 2018).

Genre classification using decoder-only models remains an underexplored area, with most prior research focusing on encoder-based architectures like BERT.

3 Dataset

For the classification task, we used the "Music Dataset: Lyrics and Metadata from 1950 to 2019" (Sahane, 2022) from the paper "Temporal Analysis and visualisation of Music (Moura et al., 2021). The dataset is provided as a comma-separated values (CSV) file and contains a total of 28,372 songs spanning the years 1950 to 2019. On average, each song contains approximately 75 words of lyrics.

The dataset includes a variety of features, but for this project, the relevant features are the song's lyrics, release year and danceability. The danceability values along with other audio metadata was retrieved using the Echo Nest API. (Sahane, 2022)

The primary feature for classification is the song lyrics, which are all in English and have been preprocessed. This preprocessing includes tokenization, stopword removal, and lemmatization. The target variable was the song's genre, selected from the following seven classes: pop, country, blues, jazz, reggae, rock and hip hop.

3.1 Metadata Explanation

As previously mentioned, this project explored two types of metadata: release year and danceability. Release year provides contextual information that may help the model interpret lyrics in light of temporal trends. For example, pop songs from 2005 may differ linguistically from those in the 1960s. Year has been proven to increase performance in other classification by Bergelid (2018), and could improve our multi genre classification task.

Danceability is a measure of 'how suitable a track is for dancing based on a combination of musical elements that include tempo, rhythm stability, beat strength, and overall regularity.' (Moura et al., 2021). It is a continuous value between 0 and 1, evenly distributed across the dataset, allowing the model to detect subtle genre-related differences along this spectrum.

3.2 Concerns about the data

As noted in related work, copyright restrictions have been an ongoing challenge for studies involving song lyrics (Tsaptsinos, 2017). These limitations also shaped the choice of dataset for this project, as there are no openly available resources that combine raw lyrics, genre labels, and appropriate licensing. The dataset used in this project included genre labels and licensing, but only provided heavily preprocessed lyrics. This introduced several challenges for genre classification as it removed important parts of the language, such as word forms, writing style, and grammar. This loss of detail made capturing meaningful patterns in the lyrics more difficult, especially for the decoder-based model.

Attempts were made to reverse the lemmatizations, using spaCy's Lemminflect (spaCy) and prompting LLaMA 3 via Ollama to regenerate more natural text. However, these efforts did not improve the quality or informativeness of the input.

4 Methods

In this section, we will present the models used in our experiment and explain the reasoning behind our model choices. We decided to test both encoder-based and decoder-based models to obtain a broader representation of the LLM field. Encoders process text in both directions, meaning each word is influenced by the words before and after it. In contrast, decoders read the text in one direction, from left to right, where each word can only be influenced by the words that came before (Touileb, 2025).

4.1 Choice of Models

As demonstrated in our related work, BERT has shown strong performance across different genre classification tasks and is therefore used as a baseline in this project. DistilBERT was chosen not only because it has performed well in earlier studies (Martinez et al., 2024; Akalp et al., 2021), but also to examine whether a lighter model can match or outperform BERT.

In addition to encoder-based models, we investigated several decoder-based approaches by inspecting popular instruct models on Hugging Face. Early experiments with T5 and Phi-3 Mini Instruct were abandoned due to poor learning behavior and inconsistent or unreliable classification results. Instead we selected LLaMA 3.1 8B instruct, a decoder-only model. It is optimized for instruction-following tasks. LLaMA was chosen to examine whether instruction-tuned decoder models can be effectively prompted to perform genre classification based on lyrics.

4.2 BERT (Baseline)

BERT is a model introduced by (Devlin et al., 2019). It is a transformer-based encoder model that captures bidirectional context. It is pre-trained using two tasks: Masked Language Modeling and Next Sentence Prediction, allowing it to capture context from both directions in a sentence (Devlin et al., 2019).

4.3 DistilBERT

DistilBERT is a smaller, faster version of BERT and follows the same general transformer architecture, but it uses fewer layers and simplified training objectives. It retains approximately 97% of BERT's performance while being 40% smaller and 60% faster (Sanh et al., 2019).

4.4 LLaMA 3.1

LLaMA 3.1 8B instruct is an openly available, and open source decoder model developed by Meta, released in 2024. LLaMA has 8 billion parameters, and it follows a standard decoder architecture and is optimized for instruction-following tasks (Meta, 2024).

5 Methodology

This section describes how we structured our approach to the genre classification task. We explain how the data was prepared, how the models were trained and tested, and how we included metadata in the process. In addition, we outline how we evaluated model performance and analyzed the results.

5.1 Data preparation

To improve data quality, we removed all songs with 15 words or fewer in their processed lyrics, as these were unlikely to provide sufficient linguistic patterns for the models. We also dropped irrelevant columns from the dataset to reduce noise and simplify processing.

A fully balanced dataset would have limited us to 904 songs per category, due to the low number of hip hop samples. To avoid this loss, we applied a stratified split to maintain genre proportions in the dataset. The dataset was split into 70% training (19 510), 15% validation (4 181), and 15% test (4 181). This ensured that all models are trained and evaluated on the same data, allowing for a fair comparison of model performance.

5.2 Encoder model training

We fine-tuned the pre-trained transformer models (BERT and DistilBERT) on our genre classification task. Both models use pre-trained transformer embeddings, where the [CLS] token from the final layer is used as representation of the input lyrics. This 768-dimensional vector is passed to a classifier for genre prediction.

When metadata was used, it was first normalized using MinMaxScaler, then passed through a linear

layer to project it into a 768-dimensional vector. This vector was then concatenated with the [CLS] embedding from the lyrics before being passed to the classifier. By making the metadata the same size as the lyrics representation, the aim was to give the model a fair chance to learn from both types of input equally.

Lyrics were tokenized using DistilBertTokenizer or BertTokenizer with max length of 256 tokens. We implemented a custom PyTorch dataset class to combine tokenized lyrics, labels and optional metadata. Both models were trained using the AdamW optimizer with a learning rate of 2e-5 and batch size of 16. DistilBERT was trained for 5 epochs, while BERT was trained for 3 epochs due to its larger capacity. To avoid overfitting, we monitored the validation loss and implemented a "save best model" strategy that stored the checkpoint with the lowest validation loss during training.

The implementation was done using PyTorch and HuggingFace transformers, with most of the inspiration coming from lecture code. We also used ChatGPT to help with debugging and designing the model architecture, especially for combining text and numerical data on the encoder models.

5.3 Prompting strategy for LLaMA

We applied prompt engineering techniques for LLaMA 3.1 based on the Prompt Engineering Guide (Guide, 2025). Multiple few-shot prompting strategies were tested, including one-shot, two-shot and three-shot setups, each providing one or more genre-labeled lyric examples. However, these consistently led to biased predictions. We also tested a seven-shot with all the genres. This made the prompt contain too much information, causing the model to overfit.

As a result, we settled on a zero-shot prompting approach, which proved to be more stable. In this setup, the prompt was phrased to guide the model to select one genre from a fixed list, without seeing any prior examples. When metadata was included, a short explanation was added in the prompt to define its meaning. The lyrics were appended at the end of each prompt, along with any relevant metadata.

We used the HuggingFace pipeline with LLaMA 3.1 in generator mode to generate predictions. do_sample was set to False to produce deterministic, repeatable outputs, while max_new_tokens=1 was used to ensure the model returned only a single predicted genre (HuggingFace, 2025). All evaluations were performed on the same test set used for the other models to ensure comparability.

The implementation also drew inspiration from the lecture codes. ChatGPT was used to help with debugging.

5.4 Evaluation metrics

To evaluate model performance, we chose to rely on macro F1-score. This choice was made because our dataset is imbalanced across genres. F1-score metric treats all classes equally by calculating the F1-score independently for each class and then averaging, making it suitable for multi-class classification with uneven class distributions (Jain, 2025). The metric was computed using the classification_report function from scikit-learn. (Scikit-itLearn)

In addition to metric-based evaluation, we have also analysed the confusion matrices as our qualitative analysis to better understand model behavior. We focused on interpreting prediction patterns, such as class biases, genre overlap and how prediction distributions changed with metadata. This gave us insight into how models handled different genres in practice, even in cases where the overall F1-score remained stable.

6 Results

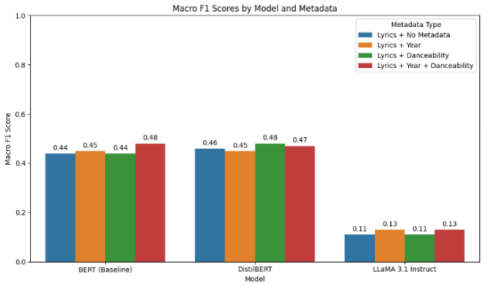

Figure 1: Macro F1-scores for BERT (Baseline) (left), DistilBERT (middle), and LLaMA 3.1 Instruct (right) across different metadata settings.

This section presents the results from our experiments with three different transformer-based models: BERT (baseline), DistilBERT, and LLaMA 3.1 Instruct. Because training these models involves some randomness, we noticed that the results could vary a bit from run to run. The differences were minor and the same patterns appeared in each run.

6.1 BERT (Baseline)

The baseline BERT classification model achieved moderate performance. Using only lyrics as input, the model reached a macro F1-score of 0.44. Considering the complexity of a seven-class classification task, this is a solid baseline.

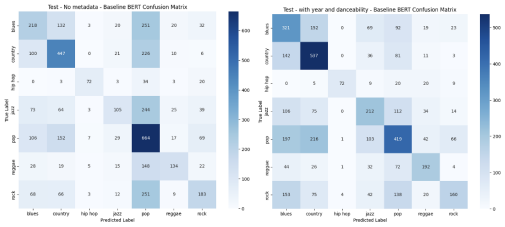

Figure 2: Confusion Matrices for BERT. BERT with no metadata (left), BERT with both year and danceability (right).

The confusion matrix to the left in figure 2 shows that predictions were relatively evenly distributed across the genre classes, suggesting that BERT was able to generalize from lyrics alone. However, it also reveals a tendency to over-predict pop, with notable misclassification from blues (251) and rock (251) into the pop category.

Adding metadata led to modest improvements, especially for underrepresented genres. The highest score 0.48 was achieved when using both release year and danceability, with improved classification for genres like country, blues and jazz as seen in figure 2. Although pop remained a frequent prediction class, the overall distribution of predictions became more balanced after adding metadata.

Overall, metadata helped improve performance and support underrepresented genres, though the model continued to favor dominant ones like pop in uncertain cases. These results serve as a baseline for evaluating the efficiency and adaptability of the other models.

6.2 DistilBERT

Figure 1, shows that DistilBERT achieved a slightly higher macro F1-score than BERT when using only lyrics (0.46 vs. 0.44). While this is slightly higher than BERT´s score of 0.44, results across multiple runs have shown fluctuations, with the performance between the two models often switching places. This suggests that the difference between them is minor, and that both models are capable of extracting relevant patterns from the lyrics.

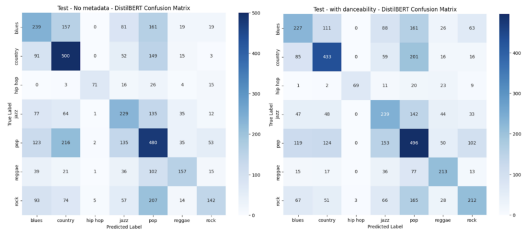

Figure 3: Confusion matrices for DistilBERT. DistilBERT with no metadata (left), DistilBERT with only danceability (right).

From figure 3 we can see that, like BERT, DistilBERT also tends to over-predict pop. However, it shows slightly better class balance overall, with predictions more evenly distributed across genres and less dominant bias toward pop.

Of the two metadata types, danceability contributed the most to improved performance. With danceability and lyrics, DistilBERT achieved its highest macro F1-score of 0.48, and showed improvements in classifying genres like jazz and rock as seen in figure 3. The predictions were more evenly distributed across genres, with fewer incorrect predictions into pop.

In contrast, adding release year alone led to a slight drop in overall performance and the model introduced more bias toward dominant genres like pop. The combination of both metadata types, release year and danceability, resulted in a slightly lower score of 0.47 as seen in figure 1.

Despite being a smaller and more lightweight model, DistilBERT consistently matched or slightly outperformed BERT, demonstrating its suitability for music genre classification tasks. This aligns with findings by Martinez et al (2024), who demonstrated DistilBERTs efficiency in genre classification from lyrics alone.

6.3 LLaMA 3.1

LLaMA underperforms compared to the encoder-based models like BERT and DistilBERT. Without metadata, the model achieved a macro F1-score of only 0.11, meaning it correctly labeled approximately ¼ as many songs as the BERT baseline. The weak performance is not unexpected, as zero-shot classification is more challenging for decoder models. Unlike fine-tuned classifiers, zero-shot prompting relies on the model´s prior knowledge, which may not include sufficient distinctions between music genres based solely on lyrics.

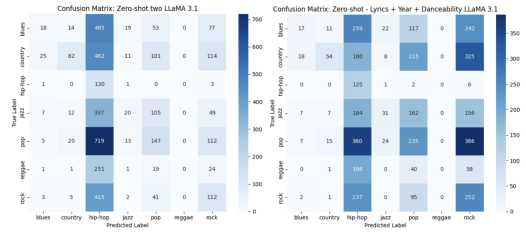

Figure 4: Confusion matrices for LLaMA 3.1. LLaMA with no metadata (left), LLaMA with both year and danceability (right).

Figure 4 shows that when using only lyrics as input, the model becomes heavily biased toward predicting hip hop, regardless of the true genre. The majority of predictions fall into the hip hop category, even for genres like country and blues which typically differ in both themes and vocabulary. This suggests that LLaMA struggles to distinguish between genres based on lyrics alone, likely due to biases present in LLaMAs pre-training where certain genres might be overrepresented and limited contextual understanding.

Adding both metadata combined led to a small increase in macro F1, from 0.11 to 0.13, but this modest numerical improvement does not fully capture the shift in model behavior. As shown in figure 4, the confusion matrices reveal a more even distribution of predictions when metadata is included. Without metadata, the model strongly defaulted to predicting hip hop for most inputs. With the combination of both metadata, LLaMA became more likely to consider other genres, such as pop, rock and jazz, and made fewer incorrect predictions of hip hop and pop. The underrepresented genres like blues and reggae saw little to no improvement, suggesting that the model still struggles to capture patterns for these classes.

Despite overall weak performance, these results indicate that metadata can help LLaMA towards more balanced prediction behavior. However, without fine-tuning, more training data or access to unprocessed lyrics, it remains far less effective than the encoder-based alternatives for this classification task.

7 Error Analysis

Model performance, particularly for LLaMA, fell below expectations. A key factor was the use of heavily processed lyrics, which likely limited the models ability to capture meaningful patterns.

In addition, we believe that lyric-based genre classification is an inherently difficult task, as lyrics alone may not provide enough distinctive information to consistently differentiate between genres. Several factors likely limited the performance of the models in this project, including legal restrictions, the simplified nature of the data set, and the general difficulty of classifying genres based solely on lyrics.

7.1 The role and effect of metadata

When metadata was included in the classification task, we saw little to no improvements. Although metadata was added as a 768-dimensional vector to match the size of the [CLS] embedding of the lyrics, it is possible that the model still prioritized the richer semantic content from the lyrics when training. This might explain why metadata had limited influence on the final predictions. However, when we analysed the confusion matrices, it became clear that metadata did affect model behavior. In particular we observed shifts in prediction distributions, for example reduction in over-prediction of pop and improved recognition of the underrepresented genres. This suggests that while metadata may not have boosted the overall score significantly, it contributed to more balances and informed predictions.

Release date had limited impact on performance and sometimes even worsened results, particularly for DistilBERT. This may be because most genres are spread across many decades, making year a weak distinguishing feature.

7.2 Investigating LLaMA through prompting

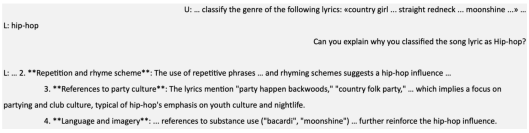

Figure 5: snippet of prompting LLaMA 3.1 through Ollama for classification explanations

We explored LLaMA's predictions by feeding the original prompts back into the model using Ollama and asking follow-up questions. While its predictions were often incorrect, the model confidently provided seemingly logical justifications. A recurring pattern was to abstract specific words. For example "moonshine" as a substance and linking it to hip hop rather than country. It also described the preprocessed lyrics as "simplistic", reinforcing its bias toward hip hop. These observations suggest the model rationalizes its outputs, even when

they are incorrect, though this might be a case of confirmation bias on our part.

7.3 Misclassification trends in encoder models

One pattern of misclassification in the encoder-models was that many blues and rock songs were incorrectly predicted as pop, especially when no metadata was included. This suggests that the model tends to default to the majority class, in our case pop. This is likely due to the class imbalance in the dataset. Additionally, it can also reflect some overlap in lyrical patterns between pop and these genres, making it harder for the model to distinguish between them. However, after adding metadata, the predictions became more balanced, and the number of misclassifications into pop decreased.

8 Limitations

This project faced several limitations that likely affected model performance.

8.1 Preprocessed lyrics

Most notably, the lyrics in the dataset were already preprocessed and lemmatized, which made it harder for models, especially the decoder model, to pick up on genre-specific cues. Due to copyright restrictions, we could not use unprocessed lyrics, which would likely have given the models more informative input. Attempts to reverse the preprocessing did not improve the results. Martinez (2024), was able to get DistilBERT up to a solid 65% accuracy, by using a dataset with unprocessed lyrics, which seems to have an effect on performance. To solve this limitation, an ethically sourced dataset could be created to make genre prediction more precise, but also more realistic.

8.2 Model transparency

Another limitation is related to model transparency. All the models used are complex black-box systems, which means we cannot fully understand why they make certain predictions. Although we explored LLaMA's reasoning through prompting, we did not apply formal explainability techniques such as LIME or SHAP. This could have provided deeper insights into the decision-making of encoder models like BERT and DistilBERT. Such methods might have helped us identify which words or metadata features were most influential in classification. This could potentially reveal systematic errors, model bias, or overlooked patterns in the data.

8.3 Time and Resource limitations

We only experimented with two features, due to time and computational constraints. A broader selection of metadata could have provided more context and improved the model´s ability to distinguish between similar genres. Future work could explore which types of metadata have the greatest impact on genre classification performance.

In terms of model selection, our exploration of decoder-based approaches was limited to LLaMA 3.1. While this model offered a useful point of comparison, it was not fine-tuned due to limited computing resources and was instead evaluated in a zero-shot prompting setup. This likely restricted its ability to adapt to the genre-classification task. Furthermore, we were unable to test newer or more advanced LLMs, because of licensing, that have since shown improvements since LLaMAs release (LLMStats, 2025).

9 Ethical considerations

The project raised some ethical considerations, particularly related to data licensing and model transparency. As discussed previously, copyright restrictions on song lyrics prevented the use of unprocessed, full-text datasets. Although using such data could have improved model performance, it would have conflicted with legal and ethical guidelines, we chose to work with a licensed dataset.

The imbalance in the dataset may have introduced biases, causing the models to favor certain genres, which in our case was the pop genre. Although our stratified split aimed to preserve class distribution and fairness, the imbalance in genre representation may still introduce performance biases. Models like LLaMA are trained on huge datasets we don't have insights into, so they might include hidden biases. In contrast, BERT and DistilBERT are trained on more controlled data, which could make them more general. Still all the models are complex and work like black boxes, so it's hard to fully understand why they make certain decisions.

While genre misclassifications may carry cultural implications in large-scale commercial systems, for example by misrepresenting culturally significant music, this is not considered a direct ethical risk in our academic setting. However, if

similar models were used in future recommender systems, it would be important to consider ways to make the classification more culturally aware.

10 Conclusion and Future Work

In this project, we have investigated the use of transformer-based language models for genre classification based on song lyrics, with and without simple metadata. Our goal was to evaluate how well lyrics alone can support genre prediction, how the metadata affects performance, and how different transformer architectures perform on this task.

Our results show that encoder models like BERT and DistilBERT can classify genres from lyrics alone with reasonable performance with F1-scores of 0.44 and 0.46. Adding metadata gave small improvements and helped reduce over-prediction of dominant genres like pop, making predictions more balanced overall. In contrast, LLaMA performed poorly with an F1-score of 0.11, and defaulted to hip hop. While metadata made LLaMA's predictions more evenly distributed, it remained less suited for this task. Overall, encoder models proved more reliable for genre classification from lyrics, while decoder models may require fine-tuning or better prompting to be effective.

Although this study faced certain limitations, especially due to the use of heavily preprocessed lyrics and not being able to fine-tune LLaMA, we were still able to identify meaningful trends in model behavior. The loss of linguistic richness likely made the task more difficult, particularly for the decoder model.

For future work, we recommend fine-tuning decoder-based models like LLaMA on genre-labeled lyrics to improve their performance. We also suggest exploring less preprocessed lyric data, experimenting with richer metadata, and using explainability techniques to better understand model behavior. As LLMs continue to develop rapidly, testing newer architectures could offer improved results and better generalization for future genre classification tasks.

References

H. Akalp, E. F. Cigdem, S. Yilmaz, N. Bolucu, and B. Can. 2021. Language representation models for music genre classification using lyrics. *ACM Digital Library*.

L. Bergelid. 2018. Classification of explicit music content using lyrics and music metadata. Digitala Vetenskapliga Arkivet. Retrieved May 9, 2025.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. Retrieved May 9, 2025.

Prompt Engineering Guide. 2025. Prompt engineering guide. Retrieved May 14, 2025.

HuggingFace. 2025. Generation strategies. Retrieved May 14, 2025.

S. Jain. 2025. F1 score in machine learning. Retrieved May 9, 2025.

Y. Li, Z. Zhang, H. Ding, and L. Chang. 2023. Music genre classification based on fusing audio and lyric information. *Springer Nature Link*. Retrieved May 9, 2025.

LLMStats. 2025. Llm leaderboard. Retrieved May 9, 2025.

S. P. Martinez, M. Zimmermann, M. S. Offermann, and F. Reither. 2024. Exploring genre and success classification through song lyrics using distilbert: A fun nlp venture. *Arxiv*. Retrieved May 9, 2025.

C. McKay, J. A. Burgoyne, J. Hockman, and J. B. L. Smith. 2010. Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features. *ResearchGate*. Retrieved May 9, 2025.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. Retrieved May 9, 2025.

L. M. G. d. Moura, C. H. Q. Foster, E. P. Fontelles, V. A. Sampaio, and M. J. C. Franca. 2021. Temporal analysis and visualisation of music. *SBC Open Lib*. Retrieved May 9, 2025.

S. Sahane. 2022. Music dataset : 1950 to 2019. Retrieved May 9, 2025.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv*. Retrieved May 9, 2025.

ScikitLearn. classification$_r eport_ scikit - learn 1.6.1 documentation$. Retrieved May 9, 2025.

spaCy. lemminflect. Retrieved May 9, 2025.

S. Touileb. 2025. Session 6. Retrieved May 14, 2025.

A. Tsaptsinos. 2017. Lyrics-based music genre classification using a hierarchical attention network. *Arxiv*. Retrieved May 9, 2025.