UNIVERSITY OF PISA

DEPARTMENT OF COMPUTER SCIENCE

Song Lyrics Genre Classification

Group 18

HLT Project Report

Group members:

Claudia Gentili 581522

Simone Ianniciello 581201

Alessandro Stefanelli 686084

ACADEMIC YEAR 2023/2024

Contents

1 Introduction

The task of song classification is of growing importance given the exponential increase in the quantity of songs available on the web and the rising popularity of music streaming platforms. These platforms are now an integral part of daily life for millions of users worldwide, making it crucial to effectively categorize and recommend music to enhance user experience. Recommendation systems employed by these platforms rely heavily on the classification of songs, which can significantly influence user satisfaction and engagement.

In the context of these vast and ever-expanding music collections, the need for automated music classification becomes evident.

Manual classification is not only impractical due to the sheer volume of data but also subject to inconsistencies and subjective biases. As a result, automated approaches leveraging advanced machine learning techniques have become essential.

One specific and challenging instance of music classification is the task of categorizing songs by genre. Genres serve as a fundamental organizational structure within music libraries, providing users with a framework to discover and explore music. However, this task is complicated by the fact that many songs do not fit neatly into a single genre. Often, songs embody characteristics of multiple genres. This ambiguity makes genre classification a non-trivial problem.

Traditionally, genre classification has relied on various features extracted from music audio signals, such as melody, harmony, rhythm, and timbre. However, another promising approach is to classify songs based on their lyrics. Lyrics encapsulate the thematic elements, linguistic styles, and cultural contexts of songs, offering rich information that can be used for classification purposes.

With the advent of advanced natural language processing techniques and deep learning models, the analysis of song lyrics for genre classification has gained considerable attention.

This work aims to explore music genre classification based on lyrics by leveraging the capabilities of pre-trained language models, which have demonstrated exceptional performance in various language comprehension tasks. Using these models, we can achieve a robust representation of texts, which facilitates more accurate genre classification.

Furthermore, by comparing the lyric-based approach with the traditional one based on audio signals, this work aims to highlight their respective strengths and difficulties, offering a comprehensive vision of future directions in song classification.

2 Prior work

Classifying music genres using audio signals is a well-studied area of research. Many papers have focused on classifying music genre using audio features describing the song, such as spectral, rhythmic, and tonal features.

Also neural methods have been used to tackle the music genre classification task on audio data. Costa et al. [1] compare the performance of CNNs in genre classification through spectrograms with respect to results obtained through hand-selected features and SVMs. Jeong and Lee [2] learn temporal features in audio using a deep neural network and apply this to genre classification. Convolutional neural networks and hand-crafted features have both shown to yield success, as well as when these methods are combined in an ensemble (Bahuleyan, [3]).

Less research has looked into the performance of these methods with respect to the genre classification task on lyrics. Mayer et al. [4] investigate how well some lyrics features work for the task of genre recognition by training and evaluating different models based on various sets of textual features computed on song lyrics (50 pre-computed textual features and 10 audio features grouped into five categories: rhymes features, statistical features, statistical features, time features, explicitness features, audio features). Lima et al. [5] propose a BILSTM model to classify a set of Brazilian song lyrics. Tsaptsinos in [6] use word embeddings of raw lyrics as input to a hierarchical attention network (a hierarchical recurrent neural network model) to predict genres, exploiting the hierarchical layer structure that song lyrics exhibit: words combine to form lines, lines form segments, and segments form a complete song. Boonyanit et al. in [7] train their own GloVe embeddings of the song lyrics and use them to train LSTM models (bidirectional and non), achieving an accuracy of 68% in the best case.

Recently, pre-trained large language models have been used in learning language representations using a large amount of unlabeled data. Akalp et al. in [8] investigate the usage of BERT and DistilBERT in music genre classification on lyrics by comparing the results with a traditional deep neural network based on a BILSTM. In particular they compare models both in terms of classification scores and complexity, by analyzing their computation times. The results show that BERT would be a reasonable solution in a real-time and real-world application, outperforming other models in terms of accuracy and time.

This study will tackle the song genre classification task based on lyrics, exploiting the power of pre-trained large language models.

3 Dataset

In order to address the task, the Genius Song Lyrics Dataset, available on Kaggle, has been used.

This dataset contains information of songs released up to 2022 scraped from Genius, a platform where users can upload and annotate songs, poems, and books (though it primarily consists of songs). It expands on the 5 Million Song Lyrics Dataset by utilizing models to determine the native language of each entry.

The format of the Genius lyrics requires preliminary processing due to:

- Song metadata often embedded within square brackets in the lyrics.
- The retention of the original lyric structure, including numerous line breaks, which poses reading challenges when processing the data for modeling.

3.1 Features overview

Table 1 contains details about each feature made available by the dataset.

Feature Name	**Desctiption**
`title`	Title of the piece.
`tag`	Genre of the piece. There are six unique genre labels: `pop`, `rap`, `rock`, `rb` (rhythm and blues), `misc` (miscellaneous) and `country`.
`artist`	Author(s) of the piece.
`year`	Year of release of the piece.
`views`	Number of views of the piece.
`features`	Other artists that contributed to the piece.
`lyrics`	Text of the piece.
`id`	Unique Genius identifier.
`language_cld3`	Lyrics language according to the cld3 model. Not reliable results are NaN.
`language_ft`	Lyrics language according to the FastText language identification model. Values with low confidence (<0.5) are NaN.
`language`	Combines language_cld3 and language_ft. Only has a non NaN entry if they both agree.

Table 1: Dataset columns and relative descriptions.

Most entries are songs, but there are also books, poems and other stuff. Most non-music pieces are labeled with the `misc` tag.

As shown in Fig. 1 and Fig. 2, the dataset is unbalanced with respect to the genre and with respect to the language.

3.2 Data Preprocessing

The preprocessing of the dataset consists in the following steps:

- Initially, all incomplete entries, i.e., those containing 'None' values, were removed to prevent skewing the analysis results.

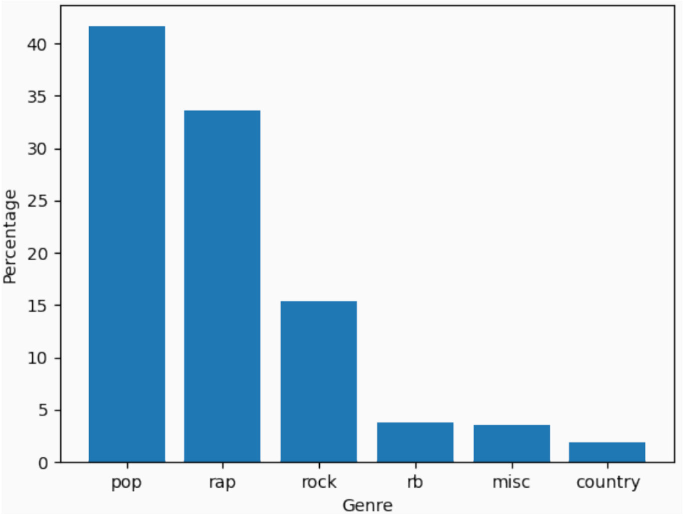


Figure 1: Percentage of songs in the dataset for each genre.

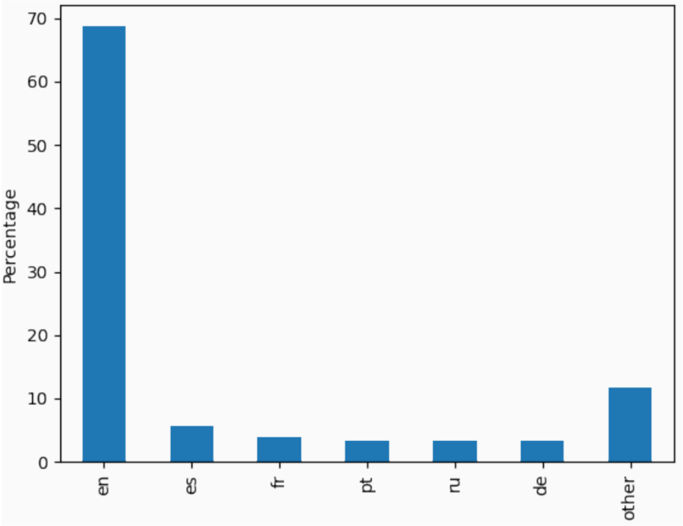


Figure 2: Percentage of songs in the dataset for each language.

- Additionally, entries tagged as `misc` were excluded, thus removing non-musical pieces that could compromise the specificity of our study related to songs.

- To focus on linguistically homogeneous and easily analyzable data, the selection was limited to English texts.

- The decision to only retain 'popular' tracks, defined as those with a view count exceeding 1,000, aims to focus the analysis on content that has garnered significant interest from listeners and has been labeled more reliably. Popular lyrics tend to be more meticulously curated in terms of content and metadata. This is because such songs receive more attention and resources, allowing for a more precise management of tags.

- Then the dataset was balanced with respect to the tag attribute through subsampling. This step is crucial to avoid the overrepresentation of specific musical genres and ensures that the developed model can generalize well across various musical categories.

- Finally the content between square brackets (i.e. the metadata) was removed from each lyric.

The result of this preprocessing is a tag-balanced dataset of 70,178 entries.

4 Models, experiments and results

This study focuses on analyzing the performance of pre-trained large language models in their behavior under both partial and complete fine-tuning.

Subsequently, the results are compared across various dimensions including F1 score, precision and recall.

4.1 Models overview

4.1.1 Baselines

To evaluate our methods effectively, it is essential to establish a baseline model. For this purpose, we employ a feedforward neural network which processes inputs that have undergone the following preprocessing steps:

- **Tokenization:** This is performed using the NLTK word tokenizer.

- **Removal of Non-Alphabetic Tokens:** We eliminate tokens that do not contain alphabetic characters.

- **Punctuation Removal:** All punctuations are removed from the text.

- **Stopwords Removal:** Common stopwords are excluded to focus on more meaningful words.

- **Conversion to Lowercase:** All text is converted to lowercase to maintain consistency.

- **Embedding Calculation:** We compute FastText (300-dimentional) embeddings for the processed tokens.

- **Vector Averaging:** The mean of these 300-dimensional FastText embedding vectors is taken.

In particular the selected baseline model architecture is composed of a single hidden layer of 300 units, a batch normalization layer utilizes ReLU activation function.

4.1.2 Recurrent Neural models

Some tests were performed on Recurrent Neural Networks scanning the lyrics and using the last token's context as *song embedding* to feed at a linear classifier. In particular the (300-dimentional) token embeddings were computed using GloVe and the recurrent layers were based on GRU units.

Given the model size and the troubles of parallelizing recurrent operations, the (left) context length was limited to 150 tokens.

4.1.3 Pretrained language models

The pretrained large language models taken into consideration are

- DistilRoBERTa, a discriminative pretrained model of size 82.8M parameters, and
- DistilGPT2, a generative pretrained model of size 88.2M parameters.

These models have been chosen in order to compare the performances of a contextualized pretrained model (i.e. pretrained with bidirectional context) and the performances of a causally pretrained model.

4.2 Experiments

The training process for the baseline model involves a series of epochs up to 1000, although this maximum is never reached. A stopping criterion based on validation loss is utilized, and the model is optimized using cross entropy loss and the Adam optimizer. The dataset is divided as follows: 60% for training, 20% for validation, and 20% for testing. Model evaluation and selection are performed using an hold-out validation set, while a separate hold-out test set is used for overall model assessment.

The model selection process selects the best model on the macro averaged F1 validation score and a final retraining is applied on the selected model using both training and validation set.

The RNN uses the same 60-20-20 splits. It uses the Optuna Framework to search for the best hyperparameters (on a subsample of 20% of the training data to speedup the process) and then retrains the model with the found parameters on the full development set for assessment on the test set.

Regarding the pre-trained language models, both partial and complete fine-tuning have been explored, utilizing again the Optuna Framework.

Key details of the model training and selection are summarized as follows:

- The usual split 60% training, 20% validation and 20% test has been used.
- A hold-out validation set was employed for model evaluation and selection, while a separate hold-out test set was utilized for model assessment.

- The model selection process chooses the best model on the accuracy score and use a subsample of 20% of the training data for the initial phases.

- A final retraining is applied on the selected model using both training and validation set.

- The fine-tuning process consists of 3 epochs both for the hyperarameters selection phase and for the final retraining.

4.3 Results

Test results analysis

The performance of various models was assessed through precision, recall, and F1 scores across different musical genres. Here is a summary of the findings:

- Baseline model performance: The baseline model showed a decent performance with a macro-averaged F1 score of 0.55. Country and rap genres performed best in terms of F1 score within this model (see Table 2).

- RNN performace: The RNN does not show any improvements w.r.t. the baseline, with a macro-averaged F1 score of 0.55. (see Table 3)

- DistilGPT2 partial fine-tuning: This approach resulted in a significant drop in performance across all genres, with a macro-averaged F1 score of only 0.19, indicating issues in model adaptation to the dataset (see Table 4).

- DistilRoBERTa partial fine-tuning: There was a notable improvement with this model, achieving a macro-averaged F1 score of 0.59. The model performed particularly well for the rap genre (see Table 5).

- DistilGPT2 full fine-tuning: This full fine-tuning significantly improved the results, achieving a macro-averaged F1 score of 0.66. This indicates that a more extensive adaptation of the model parameters was beneficial (see Table 6).

- DistilRoBERTa full fine-tuning: The full fine-tuning of DistilRoBERTa resulted in the highest overall performance among the tested models with a macro-averaged F1 score of 0.66. This model showed robustness across most genres, particularly in rap and country (see Table 7).

Genres analysis

The confusion matrices provided (Figure 3) offer insight into specific model errors:

- Rap and country genres were consistently better recognized, suggesting that these genres have distinctive lyrical features that models can learn effectively.

- Pop genre showed the lowest recognition accuracy, likely due to its diverse and overlapping stylistic elements with other genres.

Comparative Analysis

The results underline the effectiveness of full fine-tuning over partial fine-tuning strategies. In particular DistilRoBERTa full fine-tuning showed a 10% improvement in the F1 average score compared to its partial fine-tuned counterpart.

Error Analysis

Our error analysis highlights frequent misclassifications between genres with similar themes, such as pop and rock, observed in the confusion matrices (Figure 3). This suggests a blending of stylistic elements that can confuse models trained solely on lyrics.

Critical Comparisons

Comparative analysis with audio-based genre classification models underscores the limitations of lyric-only models, which lack sensitivity to musical elements like melody and rhythm. This suggests the potential of hybrid models that integrate both lyrical and audio data to enhance genre classification accuracy.

The test results reveal distinct characteristics and identities of the different genres, with rap and country being the most distinguishable in lyrical content.

Class	Precision	Recall	F1 score
country	0.66	0.54	0.60
pop	0.33	0.50	0.40
rap	0.81	0.75	0.78
rb	0.53	0.48	0.50
rock	0.51	0.42	0.46
Macro avg	0.57	0.54	0.55

Table 2: Test results for the selected baseline model.

Class	Precision	Recall	F1 score
country	0.59	0.74	0.66
pop	0.39	0.17	0.23
rap	0.74	0.77	0.76
rb	0.52	0.58	0.55
rock	0.51	0.60	0.55
Macro avg	0.55	0.57	0.55

Table 3: Test results for the selected recurrent model.

Excluding the case of DistilGPT2 partial fine-tuning, in all other cases the first well recognised genre is the rap one, followed by country, highlighting the strong identity of this two genres. On the other hand the least well recognized genre is always pop, highlighting its mixed nature.

In summary the best models are the ones resulting from the complete fine-tuning of transformer models, whose macro averaged F1 score reaches and exceeds the performances of [8], in which a fine-tuning of BERT and DistilBERT was performed in

Class	Precision	Recall	F1 score
country	0.19	0.14	0.16
pop	0.22	0.22	0.22
rap	0.14	0.21	0.17
rb	0.25	0.20	0.22
rock	0.20	0.19	0.19
Macro avg	0.20	0.19	0.19

Table 4: Test results of DistilGPT2 partial fine-tuning.

Class	Precision	Recall	F1 score
country	0.61	0.76	0.68
pop	0.44	0.23	0.30
rap	0.84	0.78	0.81
rb	0.58	0.56	0.57
rock	0.51	0.68	0.58
Macro avg	0.60	0.60	0.59

Table 5: Test results of DistilRoBERTa partial fine-tuning.

Class	Precision	Recall	F1 score
country	0.71	0.79	0.75
pop	0.48	0.44	0.46
rap	0.88	0.80	0.84
rb	0.63	0.63	0.63
rock	0.59	0.62	0.61
Macro avg	0.66	0.66	0.66

Table 6: Test results of DistilGPT2 full fine-tuning.

Class	Precision	Recall	F1 score
country	0.73	0.79	0.76
pop	0.47	0.42	0.44
rap	0.88	0.80	0.84
rb	0.64	0.63	0.63
rock	0.58	0.66	0.62
Macro avg	0.66	0.66	0.66

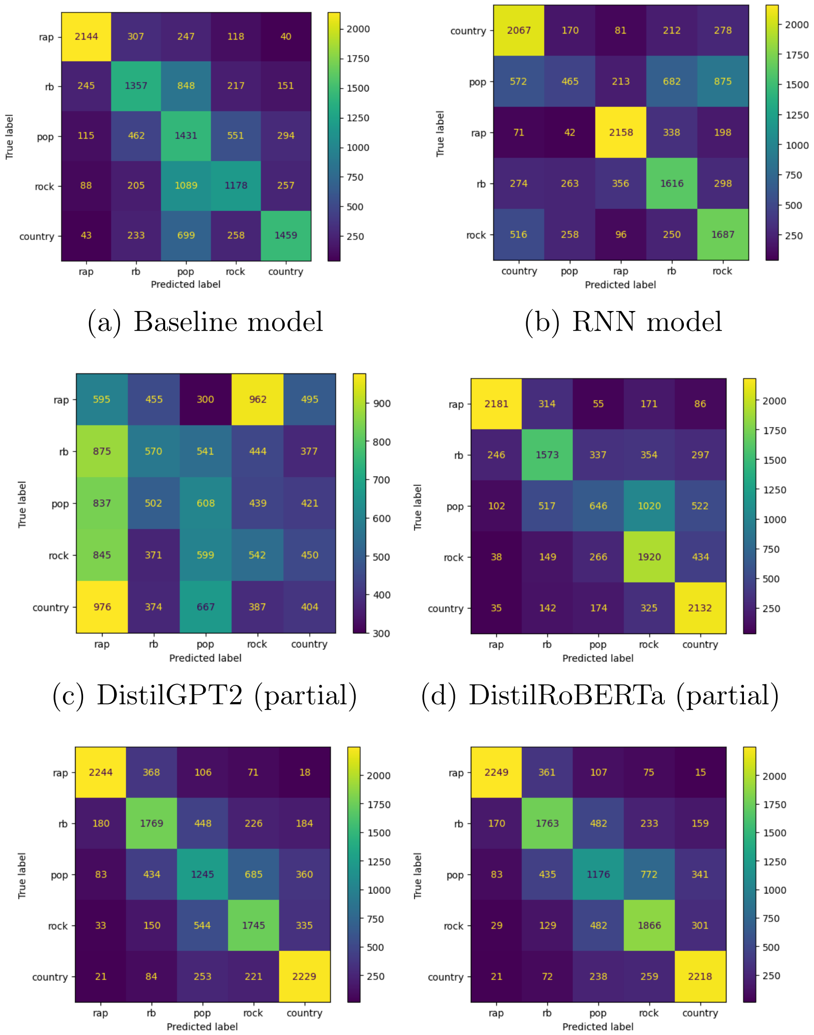
Table 7: Test results of DistilRoBERTa full fine-tuning.

order to tackle a 6 genres lyrics based classification task on a (different) preprocessed dataset of 84,664 elements.

Finally the test results give some hint on the nature and the identity of the different genres.

5 Conclusions

This study has successfully demonstrated the use of pre-trained language models, specifically DistilGPT2 and DistilRoBERTa, for music genre classification based



(a) Baseline model

(b) RNN model

(c) DistilGPT2 (partial)

(d) DistilRoBERTa (partial)

(e) DistilGPT2 (full)

(f) DistilRoBERTa (full)

Figure 3: Confusion matrices of the different models on the test set

solely on lyrics. By employing two fine-tuning strategies—partial and complete—we were able to highlight the models' capability to discern patterns and semantic nuances within song lyrics, achieving a notable average F1 score of 0.66 in complete fine-tuning scenarios.

One promising direction is the integration of non-textual information, such as audio features, to enhance the performance of genre classification models. Audio signals carry critical information about melody, rhythm, harmony, and instrumentation, which are integral to the identity of musical genres. Combining these features with lyrical analysis could lead to a more accurate classification system.

Another valuable investigation could involve a comparative analysis between distilled pre-trained models and their larger counterparts. Such a comparison would not only provide insights into the trade-offs between model size and performance but also help in understanding the impact on time and memory requirements. Larger models, while potentially offering higher accuracy, may require significantly more computational resources, which is a crucial consideration for practical applications.

Moreover, exploring different fine-tuning techniques and optimization strategies could further enhance model performance. For instance, employing advanced regularization methods or data augmentation techniques might improve the generalization capabilities of the models.

References

[1] Y. M. Costa, L. S. Oliveira, and C. N. Silla, “An evaluation of convolutional neural networks for music classification using spectrograms,” *Applied Soft Computing*, vol. 52, pp. 28–38, 2017, ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2016.12.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494616306421.

[2] I.-Y. Jeong and K. Lee, “Learning temporal features using a deep neural network and its application to music genre classification,” Aug. 2016.

[3] H. Bahuleyan, *Music genre classification using machine learning techniques*, 2018. arXiv: 1804.01149 [cs.SD].

[4] M. Mayerl, M. Vötter, M. Moosleitner, and E. Zangerle, “Comparing lyrics features for genre recognition,” in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020, pp. 73–77.

[5] R. de Araújo Lima, R. C. C. de Sousa, S. D. J. Barbosa, and H. C. V. Lopes, *Brazilian lyrics-based music genre classification using a blstm network*, 2020. arXiv: 2003.05377 [cs.CL].

[6] A. Tsaptsinos, “Lyrics-based music genre classification using a hierarchical attention network,” Jul. 2017.

[7] S. CS224N, M. Leszczynski, A. Boonyanit, and A. Dahl, “Music genre classification using song lyrics,” 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235344286.

[8] H. Akalp, E. F. Cigdem, S. Yilmaz, N. Bölücü, and B. Can, “Language representation models for music genre classification using lyrics,” *2021 International Symposium on Electrical, Electronics and Information Engineering*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236145441.