Genre Mapping: Exploring Genre Classification through BERT

Kayla Derman, Kimsean Pen

W266 Natural Language Processing with Deep Learning
UC Berkeley School of Information

1. Abstract: The challenge of managing musical preferences amid a diverse and ever-expanding set of music genres calls for the need to accurately and efficiently classify songs into genres based on certain song-specific metadata. This paper proposes leveraging Natural Language Processing (NLP) techniques, specifically a fine-tuned BERT model, to perform music genre classification using song lyrics on a comprehensive set of genres. Our results demonstrate that despite the increased difficulty of classifying a larger subset of genres, BERT-based models combined with numerical song features outperform traditional methods, suggesting that a multi-modal approach offers a more robust solution for the evolving music landscape.

2. Introduction: Music, which begins as noise, evolves into rhythms, melodies, and meaningful lyrics, creating social cohesion that spans from cultural expression to personal creative perspective. Today, music has become an important part of people's daily lives with various music genres. These genres can differ significantly from one another, leading to a distinct preference for specific subgenres and subsets of music. As new songs are continuously created, people often struggle to manage their musical preferences. Music genre classification is a crucial task in the field of music information retrieval (MIR). This is particularly important for digital music platforms like Spotify and Apple Music, which rely on accurate genre classification to develop music recommendation systems, automate playlist generation based on users' listening history, and effectively catalog their music libraries [1].

Traditional approaches to genre classification have relied heavily on numerical features such as time, amplitude, and tempo, often neglecting the rich source of information found in lyrics. Lyrics can encapsulate the emotional and thematic essence of a song which correlates with its genre [2]. However, collecting and processing this vast amount of lyrical data can be challenging. With natural language processing (NLP) and machine learning techniques, it has become possible to incorporate lyrical context into genre classification. By using word embedding derived from lyrics, we can enhance the depth of genre classification models to improve accuracy.

Previous studies on genre classification have typically focused on models predicting a limited number of genres, often around four labels. However, as new musical styles emerge and diversify, the need for more comprehensive genre categorization becomes apparent. This expansion is crucial for accurately reflecting the evolving landscape of music. We aim to explore and implement genre classification models with expanded label sets. Specifically, we will develop and compare a model predicting 8 genres with another model predicting 11 genres to evaluate the effectiveness and scalability of genre classification as musical diversity continues to grow.

Our approach involves preprocessing lyrical data, tokenizing the text, and preparing it as input for a pre-trained BERT model that we will fine-tune. We will compare the performance of this model against other experiments including Navies Bayes (our baseline), logistic regression with BERT embeddings, and a multi-modal approach that incorporates numerical features for genre classification. We will evaluate all of our models using metrics such as accuracy, and F1-score, which combine precision and recall, to provide an assessment of the model's performance. By comparing these models, we aim to gain insight into the most effective strategies for genre classification. As we look to expand the classification model to predict more genres through lyrics, our findings can help contribute to ongoing efforts in the field of MIR and support of a more sophisticated music recommendation system and playlist generation on digital music platforms.

3. Background: As the field of MIR evolves, so does the research surrounding classification techniques. Methods that use non-text metadata have found success in machine learning models, specifically relying on deep learning models. In 2018, Hareesh Bahuleyan used 10-second audio clips to create spectrograms as input for convolutional neural networks (CNNs) and hand-crafted features for traditional classifiers like logistic regression, random forests, and support vector machines. Ultimately, they find the greatest success with deep learning models, specifically the CNNs with spectrograms, classifying the songs into 7 genres with up to 64% accuracy, an F-score of 61%, and AUC value of 89% [3].

In a different approach focusing on Part-of-Speech (POS), this study applies POS tagging as a textual feature for genre classification utilizing the LingpingPipe toolkit where lyrics are categorized into grammatical groups such as nouns, verbs, and adjectives. The classification methods include k-Nearest-Neightbor (k-NN), Naive Bayes, and Support Vector Machines (SVM). The dataset consists of songs labeled with 10 distinct genres, resulting in classification accuracies of 38% for k-NN, 37% for Naive Bayes, and 40% for SVM. These results highlight the challenge of lyric-based genre classification, primarily due to the subtle and overlapping lexical patterns across different genres. Despite these challenges, the study provides insight into the potential of textual analysis in enhancing genre classification in MIR [4].

Additional methods that utilized text were conducted in 2018, where deep learning models were explored this time with song lyrics as the model's input to attack the genre classification problem. The study uses two embedding methods: Word2Vec embeddings and Word2Vec embeddings with TF-IDF scores as weights and apply these to various machine learning models, ultimately finding a 3-layer DNN using the weighted embeddings to yield top results at 74% accuracy. In this instance, advanced word embedding techniques significantly improved genre classification accuracy, leading to our interest in exploring more complex embedding methods such as BERT. That said, they only focused on a subset of music in four distinctive genres: Christian, Metal, Country, and Rap [5].

By combining both audio and lyrical features, Mayer, Neumayer, and Raber significantly enhance genre classification accuracy. They integrated various audio features such as Rhythm Patterns (RP), Statistical Spectrum Descriptors (SSD), and Rhythms Histograms (RH). Their research also analyzed lyrical features, including rhyme patterns, the frequency of unique rhyme words, Bag-of-Words (BoW) using TF-IDF to represent word frequency in lyrics, and text statics that measure average word length, words per minute (WPM), and punctuation usage. The classifiers used in the study include SVM, k-NN, and Naive Bayes. The most effective model resulted from combining SSD audio features with text statistics and POS features achieving an accuracy of 68.72%. This study demonstrates the value of a multi-modal approach to music genre classification allowing us to incorporate these methods into our experiment to explore their potential further and help improve the accuracy of our genre classification model [6].

4. Methods: Our research explores the effect of implementing a pre-trained BERT model on song lyrics as a means of classifying songs into genres. Additionally, we examine the influence of different genre-mapping strategies on our dataset, as well as the integration of various numerical features that describe the songs. The core of our approach is predicated on the hypothesis that lyrical content of songs encompasses distinct stylistic and thematic characteristics of different musical genres.

Data: To conduct this study with substantial genre representation, we use two datasets: Music4all [7] and Mendeley Data - Music Dataset: Lyrics and Metadata from 1950 to 2019 [8]. These datasets are concatenated with duplicates removed, and similar features are kept: {*release_year*, *danceability*, *energy*, *valence*, *lyrics*, *track_name*, *artist_name*, *genre*}. The *genre* will be the model label, *lyrics* will be the primary feature, and *release_year*, *danceability*, *energy*, *valence*, and *lyrics_length* (a new feature representing the number of words in the song lyrics) will be the additional numerical features. The resulting dataset has 74,070 entries. Given that there is substantial

research on classifying a small set of music genres, we focus on the impact of broadening this scope by working with both 8-genre and 11-genre mappings to represent a wider genre variety.

The datasets are then divided into train, validation, and test sets at 80%, 10%, and 10% of the entire dataset respectively. Labels are then encoded as integers and text is cleaned. Tables 1 and 3 show the genre distribution of the 8-genre and 11-genre classification datasets, while Tables 2 and 4 detail the count of genre labels specifically within the training sets. While we do not completely balance the training set, genres are set to a maximum of 7000 songs to avoid model favoritism towards over-represented genres.

	genre	count
0	Rock	15998
1	Pop	15047
2	Jazz and R&B	11216
3	Country and Folk	10364
4	Punk and Metal	10059
5	Electronic	5406
6	Hip Hop & Rap	3371
7	Reggae	2609

Table 1: Genre distribution for 8 genre datasets.

	genre	count
0	Pop	7000
1	Jazz and R&B	7000
2	Rock	7000
3	Country and Folk	7000
4	Punk and Metal	7000
5	Electronic	4324
6	Hip Hop & Rap	2693
7	Reggae	2060

Table 2: Genre distribution in training set for 8 genre datasets.

	genre	count
0	Rock	15998
1	Pop	15047
2	Rhythm & Blues	6864
3	Country	6282
4	Electronic	5406
5	Metal	5145
6	Punk	4914
7	Jazz	4352
8	Folk	4082
9	Hip Hop & Rap	3371
10	Reggae	2609

Table 3: Genre distribution for 11 genre datasets.

	genre	count
0	Pop	7000
1	Rock	7000
2	Rhythm & Blues	5505
3	Country	5058
4	Electronic	4324
5	Metal	4092
6	Punk	3927
7	Jazz	3469
8	Folk	3279
9	Hip Hop & Rap	2693
10	Reggae	2060

Table 4: Genre distribution in training set for 11 genre dataset.

Measuring Success:

The primary metric to measure the success of our model will mainly rely on the accuracy score which quantifies the proportion of correct predictions out of the total number of predictions. For our BERT models, we also incorporate precision, recall, and F1 scores to provide a more nuanced evaluation of the model's performance across different labels. These metrics will help us assess how well each genre label is predicted. Furthermore, confusion matrices are used to visualize the performance of the model, highlighting areas of strength and identifying specific genres that may require further refinement.

4. 1 Experiments

Naive Bayes: For our baseline model we implement a Multinomial Naive Bayes classifier. Given the possibility that varying genres feature distinct standard vocabularies, this feels like a strong place to start for a lyric-based music genre classifier and has been used as a baseline in similar research [6]. For this implementation, we use a Term Frequency-Inverse Document Frequency (TF-IDF) model. Song lyrics are transformed with the TF-IDF vectorizer with a maximum of 10,000 features and the Multinomial Naive Bayes model is trained using our training set.

Logistic Regression: A logistic regression model was selected as an additional baseline for this classification task as a secondary method for analyzing the TF-IDF vectorized data. To convert the text data into numerical features, we again employ a TF-IDF vectorizer with a maximum of 10,000 features. These features are then input into the regression model, which uses 'liblinear' as the solver. This initial experiment was conducted using 8 and 11 labels to compare the model's performance across different numbers of genre classes. We observed that the logistic regression model exhibited substantial overfitting, performing well on training data but showing a considerable drop in performance on unseen data. This finding indicates the need for further refinement, leading us to incorporate a BERT model.

BERT: The previous experiment (Logistic Regression) relied on static, pre-defined features such as TF-IDF representation, resulting in features that do not dynamically change based on the context of the words or specific text input. Lyrics often contain context-dependent words and ambiguous phrases with multiple meanings [9]. Efficiently predicting the genre of songs based solely on their lyrical content requires understanding these intricate layers of meaning, which poses challenges for traditional machine learning models. BERT addresses these challenges by reading text bi-directionally, considering both the preceding and following words to understand the overall context. This allows BERT to disambiguate words with multiple meanings and adjust interpretations based on context, thus improving genre classification accuracy.

To prepare our data, we encoded the genre labels using a 'LabelEncoder' and tokenized the lyrics with a BERT tokenizer, converting them into token IDs, token type IDs, and attention masks for input into the BERT model. Our initial model architecture utilized a pre-trained BERT base model, allowing for fine-tuning on our specific task. The tokenized inputs were processed by BERT, with the pooler output passed through a dense hidden layer with ReLU activation and dropout for regularization. A softmax output layer then classifies the lyrics into predefined genres. The model is compiled using an Adam optimizer, sparse categorical cross-entropy loss, and accuracy as the evaluation metric. However, this model architecture led to a high runtime, with a single epoch exceeding 30 minutes, prompting the need for a more time-efficient structure without compromising accuracy.

In an effort to address the long runtime, we modified the BERT model by freezing most layers except for the last transformer block and pooler layer, preserving pre-trained knowledge while enabling task-specific fine-tuning. We extracted the CLS token, which aggregates the entire sequence's representation, to capture the overall meaning of the lyrics. This token was then passed through a dense hidden layer with the same ReLU activation as before and dropout for regularization. Finally, a softmax layer predicted the genre across different classes.

Due to our class imbalance, we incorporated sample weights to prevent the model from becoming biased towards predicting the more common class. Initially, each sample is assigned a weight of 1. We then calculate weights for each class based on their frequency, assigning higher weights to less frequency classes and lower weights to classes with more samples. These adjusted weights are applied to the training sample as the model trained for 6 epochs with a batch size of 8, using a hidden size of 100, a learning rate of 2e-5, and a dropout rate of 0.3.

BERT & Numerical Features: Inspired by the research conducted by Mayer, Neumayer, and Raber, which combined audio numerical features with lyrical text to improve genre classification, we decided to experiment with a multi-modal approach. Our objective was to evaluate if incorporating additional numerical features could enhance the accuracy of our genre classification model. In this final experiment, we extended the BERT model used in our text-only approach by adding an input layer for non-textual features, which includes five numerical attributes (release, danceability, energy, valence, and lyrics length) associated with each song. The BERT model's output, represented by the CLS token, was concatenated with these numerical features combining text and numerical data. This combined input was then passed through a series of dense layers with ReLU activation and dropout for regularization. Output layer and model compilation were consistent with those used in the text-only BERT model.

5. Results: After fine-tuning our models using the validation set, we tested the four approaches on both the 8-genre mapped and 11-genre mapped test sets, garnering in Table 5:

Test Accuracy Score		
Model	**8-Genre Mapping (%)**	**11-Genre Mapping (%)**
Naive Bayes (TF-IDF)	41	37
Logistic Regression (TF-IDF)	43	40
BERT	48	44
BERT & Numerical Features	**51**	**47**

Table 5: Test accuracy score for 8 and 11 genre mapping.

Two things are clear from these results: 1. Mapping songs into fewer and broader genres makes classification easier, and 2. Genre classification from song lyrics performs better with more complex models, as well as the addition of other song-related features. Ultimately, we are able to predict a set of 8 genres with 51% accuracy and a set of 11 genres with 47% accuracy using a BERT model combined with number-based song features. In both of these instances, we achieve a 10% increase in accuracy compared to the baseline Naive Bayes model

F1 Score		
Genre	**Text-Only BERT**	**Multi-Modal BERT**
Country	**0.53**	**0.56**
Electronic	0.30	0.38
Folk	0.33	0.48
Hip-Hop & Rap	**0.78**	**0.76**
Jazz	0.35	0.35
Metal	**0.62**	**0.66**
Pop	0.47	0.45
Punk	0.40	0.44
Reggae	0.43	0.48
Rhythm & Blues	0.33	0.36
Rock	0.37	0.45

Table 6: F1 scores for Text-Only BERT and Multi-Modal BERT.

In Table 6, we see the F1 Scores for the 11-genre mapped data using our BERT Model with and without the additional numerical features. As noted, the overall accuracy of our model increased with the added number-based features. With that in mind, we notice that this is not the case for all label-level metrics. Though most of the label level F1 scores increase with the multi-modal model, Hip-Hop & Rap and Pop perform worse, suggesting that these genres actually share more similarities with other genres musically, as represented in this instance by the numerical features. Especially considering how well Hip-Hop & Rap perform with a lyric-based classifier, it makes sense that much of what distinguishes this genre is its lyrical content.

Moreover, we see that Hip-Hop & Rap, Metal, and Country have the overall highest F1 scores when predicted with these models, suggesting that these genres might actually be more defined by their lyrics than others. Notably, the 2018 Kumar, Rajpal, and Rathore study that reaches 74% classification accuracy only compares these three genres along with Christian music.

6. Discussion: In our error analysis we noticed that both text-only and multi-modal models exhibit a bias towards the higher-class genre, pop and rock, often predicting these genres even when the actual genre is different. While we applied sample weights during the training phase to address the class imbalance, additional techniques such as data augmentation, specifically by extracting POS and incorporating them as textual features into our model, may help it differentiate and identify POS patterns that are characteristic of certain genres.

We also employed Optuna, an automatic hyperparameter optimization software, to identify the best parameter for our model. Utilizing Bayesian optimization, Optuna conducts multiple trials to find the best hyperparameters [10]. Initially, we conducted a broad search to determine the best values for learning rate, hidden dimension, batch size, and dropout rate. However, after observing no significant improvement in accuracy scores, we refined our search to focus on more specific values. According to 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding' [11], increasing the hidden size from 200 to 800 can help, but values beyond 1000 do not yield further accuracy improvements. Consequently, we experimented with the hidden sizes of 400, 600, and 800. For the learning rate, we tested 5e-5, 4e-5, and 3e-5 and kept the dropout to 0.1 as suggested in the article. After running these experiments on both text-only and multi-modal models, there were no changes in the accuracy scores indicating that adjusting the hyperparameters was not sufficient to improve accuracy or F1 scores. Due to the long duration per epoch and limited computational resources, we only ran 5 trials.

Despite our efforts to improve the model's accuracy through fine-tuning, we observed no significant changes in accuracy. This prompted us to look further into the word distribution to understand the most common words across different music genres, as each genre often has a unique set of vocabulary or slang that can help distinguish them from one another [12]. We found that there was significant overlap in frequently used words such as 'know', 'like', 'love', and 'oh' across genres like pop, rock, and jazz. These overlaps can make it challenging for the model to accurately differentiate between genres, especially when such words are prevalent in multiple categories. Even though we already utilized TF-IDF to help weight distinctive words more heavily, implementing bi-gram and tri-gram can provide additional benefits to help capture more context-specific phrases unique to each genre.

7. Conclusion: Our research emphasizes the importance of expanding genre classification efforts to include a broader range of genres while acknowledging that lyrics alone may not sufficiently distinguish between subgenres. This study highlights the nuanced roles of human input in defining genres and the extent to which lyrical content influences these divisions. We see this in the way that certain genres are more easily predicted based on their lyrics alone. Moreover, we identify the success of implementing an attention-based model such as BERT has on making predictions based on song lyrics, suggesting there is significance beyond the vocabulary differences that make up a genre. Additionally, the inclusion of numeric features enhances classification accuracy, suggesting the potential of multi-modal models. Based on the combined findings of the 2018 spectrogram study [3] and our current research, we recommend future studies adopt a multi-modal approach, incorporating both audio features, such as spectrograms, and textual features from lyrics, to develop more robust genre classification models that are able to predict a wide variety of music genres.

References:

1. H. Mukherjee, M. Marciano, A. Dhar and K. Roy, "A song emotion identification system from lyrics using heterogeneous ensemble learning," 2023 IEEE Silchar Subsection Conference (SILCON), Silchar, India, 2023, pp. 1-5, doi: 10.1109/SILCON59133.2023.10404341.
2. Naseri, S., Reddy, S., Correia, J., Karlgren, J., & Jones, R. (2022). The contribution of lyrics and acoustics to collaborative understanding of mood. In Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022).
3. Bahuleyan, Hareesh. (2018). Music Genre Classification using Machine Learning Techniques.
4. Teh Chao Ying, S. Doraisamy and Lili Nurliyana Abdullah, "Genre and mood classification using lyric features," 2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, 2012, pp. 260-263, doi: 10.1109/InfRKM.2012.6204985.
5. Kumar, Akshi & Rajpal, Arjun & Rathore, Dushyant. (2018). Genre Classification using Word Embeddings and Deep Learning. 2142-2146. 10.1109/ICACCI.2018.8554816.
6. Mayer, Rudolf & Neumayer, Robert & Rauber, Andreas. (2008). Combination of Audio and Lyrics Features for Genre Classification in Digital Audio Collections. MM'08 - Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops. 159-168. 10.1145/1459359.1459382.
7. Igor André Pegoraro Santana and Fabio Pinhelli and Juliano Donini and Leonardo Catharin and Rafael Biazus Mangolin and Yandre Maldonado e Gomes da Costa and Valéria Delisandra Feltrim and Marcos Aurélio Domingues. Music4All: A New Music Database and its Applications. In: 27th International Conference on Systems, Signals and Image Processing (IWSSIP 2020), 2020, Niterói, Brazil. p. 1-6.
8. Moura, Luan; Fontelles, Emanuel; Sampaio, Vinicius; França, Mardônio (2020), “Music Dataset: Lyrics and Metadata from 1950 to 2019”, Mendeley Data, V2, doi: 10.17632/3t9vbwxgr5.2
9. Ambrosch, Gerfried. 2020. The Poetry of Song: The Synergy of Music and Lyrics. *Culturico*.
10. *A hyperparameter optimization framework*. Optuna. (n.d.). https://optuna.readthedocs.io/en/stable/
11. Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805* (2019). https://doi.org/10.48550/arXiv.1810.04805.
12. Chesley P (2011) You Know What It Is: Learning Words through Listening to Hip-Hop. PLoS ONE 6(12): e28248. https://doi.org/10.1371/journal.pone.0028248