Music Genre Classification

Kevin Zhuo
Williamstown, MA 01267
kjz1@williams.edu

Keel Brissett
Williamstown, MA 01267
kmb9@williams.edu

Abstract

Traditionally, genre classification models are only based upon the audio features of songs. In this project, we seek to examine the impact of incorporating lyrics into these models, and whether or not it yields a greater accuracy in classification. We evaluate two approaches: a lyric-only model using BERT embeddings in a multiclass logistic regression, and a multimodal model where acoustic features are appended to BERT embeddings and fed into a multiclass logistic regression. Our experimental results show that the multimodal model outperforms the lyric-only model, with both approaches achieving significantly better results when compared to a majority classifier baseline.

1 Introduction

Music has transformed over the modern era, and the digitization of sounds has paved the way for the large-scale production and distribution of songs. As a result, music streaming platforms strive to enhance user experience by curating genre-aligned playlists. However, popular platforms like Spotify predominantly rely on acoustic features such as pitch, melody, tempo, and frequency to match songs with genres. While these acoustic features can be easily quantified and graphically represented, they fail to account for the lyrics, which may also convey valuable information. This limitation necessitates a deeper exploration of incorporating lyrics into genre classification, which we hypothesize will improve the overall accuracy of these models. We hope that by improving the accuracy of these models, users on streaming platforms are able to have a better experience when exploring songs that they enjoy. Furthermore, there is potential that improving genre classification may lead to other improvements in downstream tasks such as music generation (Agostinelli et al., 2023). Since prompts into music generation often include the genre as keywords, being able to label music to genre automatically can greatly improve the data used for this task. There are two primary objectives that we plan on exploring in this paper. Firstly, we examine the performance of BERT embeddings in a text-only genre classification approach. When it comes to lyrics, existing studies in the field have predominantly utilized static embeddings for genre classification, which fail to capture the nuanced semantic meanings inherent in the lyrics. By leveraging BERT embeddings, we hope to enhance the accuracy and effectiveness of text-only genre classification models. Secondly, we delve into the exploration of a multimodal approach with fusion. This approach involves combining the BERT embeddings with acoustic features to create a unified representation, which is subsequently utilized in Multiclass Logistic Regression for genre classification. Due to the inherent complexity of music, both the acoustic features and lyrics play a significant role in the overall formation of a song. As a result, we believe that this multimodal fusion approach will yield superior classification outcomes.

2 Related Work

One important related work is BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), which is employed to get our embeddings. In order to better learn the bidirectional representation of words and to capture semantic relationships between sentences, BERT introduces masked language modeling and next sentence prediction as two pre-training tasks. BERT's effectiveness comes from the fact that it is able to jointly condition on both left and right context in its layers. Unlike older causal transformers that only go left to right, BERT excels at sequence classification due to the fact that its self-attention mechanism ranges over the entire input. As a result, its output embedding is contextualized over the entire sequence. This is also an advantage over static embeddings

such as GloVe or Word2Vec, as those embeddings do not take into account the context that a word appears in, and only base their representations on co-occurence statistics. One other benefit of using BERT is that it can easily be used for downstream applications through fine-tuning. For the purpose of classification, BERT includes a "sentence embedding" which is located in the [CLS] token.

Some other published music genre classification approaches have dealt with only lyrics using a hierarchical attention network (Tsaptsinos, 2017). This paper uses the fact that lyrics exhibit a hierarchical layer structure, which can be seen through the fact that words form lines, lines form segments, and segments form songs. For their vector representation of the words, they used 100-dimensional GloVe embeddings. In the hierarchical attention network used in the paper, every layer is run through a bidirectional gated recurrent unit to get a vector with a weighted sum, which is then passed on to the next layer. The gates help determine how important the previous hidden state is to the next hidden state and the creation of the next memory. Furthermore, to capture which words and segments are more important, the hierarchal attention network included a layer at the word level and a layer at the segment level. Each layer had a separate GRU weight matrix, attention weight matrix, and relevance vector. Compared to some older models, the hierarchical attention network is able to achieve better accuracy as well as distinguish the words that were most strongly associated with certain genres.

Other methods of genre classification take into account text and audio using Convolutional Neural Networks (CNNs) (Oramas et al., 2018) in a multimodal approach. In that paper, the authors propose a system that predicts music genre labels given different modalities. First, a neural network is trained on the classification task for all the modalities. Afterwards, the representations are extracted and combined into a single vector. Since the dataset contained pure audio, they utilized Mel-Spectrograms to convert the audio into an image. Then, they fed the image through a CNN and retrieved the activation of the last hidden layer as the audio representation. The text in the dataset was fed through a Vector Space Model with a feed-forward network containing two dense layers, where the last hidden layer became the text representation. When the representations of the two different modalities are acquired, they get embedded in a new multimodal space that optimizes their similarity. A deep learning approach is then used to minimize the loss of the shared embedding, while negative sampling is also utilized to avoid mapping all vectors to a single point. When the final embedding of audio and text is obtained, it is fed through a multiclass logistic regression model for the predicted genre label.

Beyond just applying BERT to natural language, there also exists an application for BERT when it comes to pure music. Called MusicBERT (Zeng et al., 2021), this pre-trained model hopes to gain understanding of music from symbolic data. Since music songs are more structural and complicated, directly applying BERT to music data often results in marginal gains. Furthermore, traditional ways of encoding music are too long, which makes them computationally impossible to train with BERT. To tackle this problem, the paper proposes a new music encoding method called OctupleMIDI, which is able to encode every note into a tuple that contains various acoustic features. Since every note contains features that pertain to the music genre, the OctupleMIDI encoding method is much simpler and shorter than previous encoders. Beyond just the music encoder, the paper also designs a masking strategy for MusicBERT, where elements in the same type in the same bar are regarded as the same unit and masked at the same time. This avoids information leakage and allows for better contextual representations. Similar to BERT, MusicBERT also has success with various downstream tasks such as melody completion, accompaniment suggestion, and genre classification.

3 Dataset

The dataset that we used for our model and analysis was "Audio features and lyrics of Spotify songs"[1]. This choice was motivated by the need to address copyright issues commonly associated with song lyrics. While most music-related models rely on count-based bag-of-words representations, word order is extremely important when it comes to music. This is because artists often align specific words with beats and match words with each other to create rhythm. Furthermore, our approach with BERT relies on the positional and contextual information of words. So, we opted for a dataset where the lyrics are preserved in their original order. The audio features and genre were originally collected

[1] https://www.kaggle.com/datasets/imuhammad/audio-features-and-lyrics-of-spotify-songs

from the third week of the TidyTeusday Project, [2] and the lyrics were added with the genius library in R. The dataset encompasses over 18,000 songs, providing comprehensive details such as artist, album, lyrics, audio features, and genre. The genres included are pop, rock, rap, R&B, and latin. While the audio features are all numerical vectors, the lyrics are in the form of space-separated words. In terms of the audio features, there were four that we thought were important: acousticness (confidence measure from 0-1 on how acoustic the track is), tempo (estimated tempo of the track in beats per minute), valence (measure from 0-1 describing the positivity conveyed from the track), energy (measure from 0-1 that represents the intensity), and danceability (how suitable the track is for dancing). Although any individual audio feature may be useless in isolation, we believe that combining multiple together can provide deeper information about the overall audio of a song.

4 Methods

Figure 1 shows the overall workflow of our project

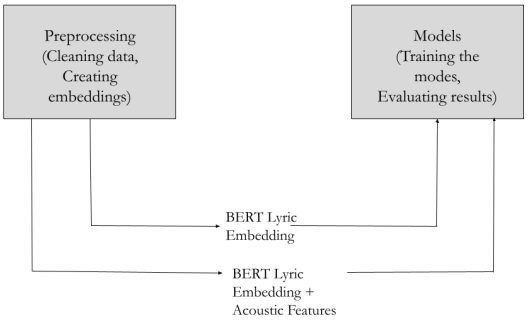

Figure 1: Pipeline for the project

4.1 Preprocessing

The first step that we undertook was preprocessing the data. After the dataset was imported in, we removed all songs that were not in English. This is because the BERT model that we utilize was not made to handle multiple languages, and we wanted to avoid the scenario where certain words are spelled the same, but contain varying semantic meanings in different languages. We then subset the data to only contain the columns "lyrics", "playlist_genre", "acousticness", "tempo", "valence", "danceability", and "energy" since those were the only features that mattered for our project. There were also some rows with empty values in the dataset that we dropped. Since the genres were strings, we used the LabelEncoder from sklearn to transform the genres into corresponding numbers which we concatenated to the dataframe. Due to computational limits, we had to randomly sample 200 rows to use as our dataset so the program would not use too much memory and be able to run in a reasonable amount of time. We then created a random split of 70% for the training data and 30% for the test data.

4.2 Creating the Embeddings

For the embeddings, there were numerous methods that we considered. However, due to the lyrics in the dataset being preserved in their original order, we believed that the BERT embeddings would be the best since it takes into account the contextual meaning of words as well as their semantic relationships. For all the tasks relating to creating the embeddings, we used the GPU instead of the CPU to speed up capabilities (Buber and Diri, 2018).

The specific model and tokenizer utilized in our study is DistilBERT (Sanh et al., 2020), a compact variant of BERT that retains 97% of its language understanding capabilities while reducing its size by 40%. To prepare the song data for embedding generation, we employed the DistilBERT tokenizer. This process involves tokenizing the lyrics by breaking words into subword units, which are then mapped to their respective IDs. Additionally, special tokens such as [CLS] (representing the start of the sequence) were inserted at the beginning of the tokenized sequence and [SEP] (marking sentence boundaries) was inserted at the end of the sentences.

Given that our lyric data consisted solely of words, sentence distinctions were not crucial, and all segment embeddings of the tokens were set to 0. To capture positional information, position embeddings were also added. Due to hardware limitations, we imposed a maximum token length of 100, with automatic truncation and padding applied as necessary (Sun et al., 2020). The tokenized training and test inputs were then fed through the DistilBERT model, allowing us to extract the corresponding embeddings by capturing the representation associated with the [CLS] token from the final hidden

[2] https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-01-21

state.

4.3 Lyric-Only Model

In the genre classification model that exclusively focused on lyrics, the text embeddings served as the sole input. Once the embeddings were appropriately processed, they were fed into a Multiclass Logistic Regression model (Agarwal et al., 2021). To reduce the chance of model overfitting, we incorporated a L2 penalty term. During the training process, the model was trained with a maximum of 1000 iterations to converge, and updating its parameters every iteration. Given the presence of multiple classes, we utilized the multinomial logistic regression approach to handle the genre classification task.

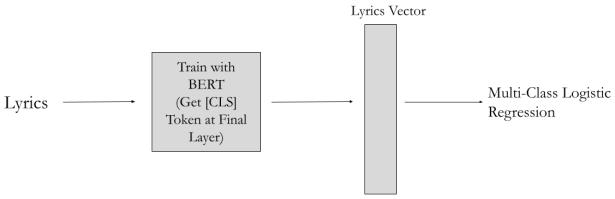

Figure 2: Lyric-Only Model with BERT embeddings

4.4 Lyric and Acoustic Features Multimodal Model

In order to align the different modalities within a joint embedding space, we employed a multimodal approach where the acoustic feature vectors were concatenated with the BERT embeddings for lyrics. Since the acoustic feature vectors were one dimension and the lyric embeddings were two dimensions, we had to unsqueeze the acoustic feature vectors in order to match the first dimension of the lyric embedding. This concatenation allowed for the integration of textual and acoustic information, enabling the fusion of modalities in a unified representation. When the joint embeddings was obtained, we fed it through a Multiclass Logistic Regression model. Similar to the Lyric-Only model, we employed an L2 penalty term with 1000 iterations.

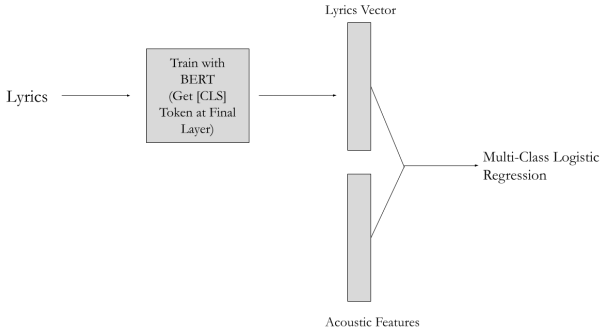

Figure 3: Late-fusion multimodal Model with Bert embeddings and Acoustic Features

5 Results and Evaluation

The results that we achieved on the test set can be observed in Table 1.

	Weighted F1 Score
Baseline (MC)	0.09
Audio	0.17
Lyrics	0.33
multimodal	0.37

Table 1: Genre Classification weighted F1 scores on the test set using Majority Classifer (MC), Acoustic Features Multiclass Logistic Regression (Audio), Lyrics-Only Multiclass Logistic Regression (Lyrics), and Lyrics + Acoustic Features Multimodal Model (Multimodal)

The outcome of our experiments reveal a substantial performance improvement of both models compared to the Baseline model, which acts like a majority classfier by predicting "pop" as the genre for every song. The model exclusively trained on BERT embeddings of song lyrics achieves a weighted F1 score of 0.33. Meanwhile, the multimodal model which incorporates BERT embeddings concatenated with acoustic feature vectors demonstrates superior performance with a weighted F1 score of 0.37. Both approaches were able to outperform the audio only model, which consisted of the acoustic feature vectors fed through a Multiclass Logistic Regression.

The decision to utilize the weighted F1 score instead of pure accuracy stems from the imbalanced class distribution observed in our dataset. Certain genres, such as pop and rap, exhibit higher frequency compared to others. By considering the weighted F1 score, we account for this class imbalance and prioritize a comprehensive evaluation

metric that encompasses both precision and recall. Furthermore, our emphasis on penalizing false-positives in the evaluation metric aligns with our intention to optimize user experience in streaming services. We want to make sure that our evaluation metrics align closely with listener expectations. Inaccurate genre assignments can lead to suboptimal user experiences, which is why a weighted F1 score would be better than accuracy in this instance.

One thing to note in our results is that by changing the random seed used to subset the data, the results for the Audio-only model, Lyric-only model, and multimodal model vary greatly. However, even with the great fluctuations, they are all significantly above the baseline which hovers around a weighted F1 of 0.1. This can be attributed to the fact that we were only able to randomly sample 200 songs out of the 18000 present in the dataset due to hardware limitations. Even when running the model on the GPU and lowering the number of tokens down to 100, it would take hours to train the model to get the BERT embeddings. Since 200 songs is proportionally very small when compared to 18000, only being able to randomly sample 200 could lead to a skewed dataset that is not representative of all the songs.

6 Conclusion

In conclusion, our project highlights the value of incorporating lyrics in genre classification tasks. By leveraging BERT embeddings, we observed improvements in genre classification performance, indicating the relevance and usefulness of lyrics in genre categorization. This insight contributes to the broader field of music analysis and classification, highlighting the importance of considering textual information in addition to other modalities. However, even though we are able to state that lyrics are an important factor when it comes to genre classification, we are unable to conclusively say that it is more important than the audio features of the song.

Through working on the project and reading the literature, we gained a comprehensive understanding of how BERT works and how it can be used for a variety of different applications. Additionally, we recognized the significance of data cleaning and preprocessing steps in ensuring the quality and effectiveness of our models. The overall process of reading theoretical knowledge and then applying it by writing code has allowed us to gain a deeper understanding of the topics.

There were some limitations to our project that we hope we can tackle in the future. One big problem was the training time for the DistilBERT model. Even though the model is much lighter than BERT and we ran it on the GPU, the training time was too long to justify it being practical. When scaled up to the hundreds of thousands of songs, a very robust system would be needed to process all the data. Another limitation that we faced was the audio features being in the form of numerical vectors. Although this did streamline the process of obtaining the acoustic features, using the true audio file would be a more accurate representation and most likely yield better results.

Some ways in which we can expand this project is to include the MuLan embeddings (Huang et al., 2022) instead of BERT embeddings. While our method was late fusion where the embeddings were first trained and then appended onto the acoustic features vector, the MuLan embeddings are an early fusion method where contrastive loss is used to jointly train the shared embedding space between the music audio and text. This joint training approach enhances the embeddings' ability to capture the intricate semantic connections between music and lyrics. By incorporating these MuLan embeddings, the accuracy of genre classification could be greatly enhanced, as well as open up the avenue for various downstream applications.

Another way that we could expand this project is to include more genres in our dataset. The dataset that we used only contained five genres, which represents a considerably smaller subset compared to the actual multitude of genres existing in the music landscape. Expanding the dataset to encompass a larger number of genres may present challenges in terms of hardware capabilities, such as storage capacity and computational resources. However, by incorporating a broader range of genres, we can capture a more comprehensive representation of musical diversity and cater to individual preferences more effectively.

One real-world implication for the work that we have done is showing the importance of lyrics when it comes to music genre classification. Since using lyrics and the multimodal model result in nontrivial increases to the weighted F1 score, there is potential that streaming platforms such as Spotify also incorporate lyrics when creating their genre playlists. Since lyrics are an integral part of music, streaming platforms will be able to create more

tailored music recommendations that provide users with an immersive music discovery experience.

References

Naman Agarwal, Satyen Kale, and Julian Zimmert. 2021. Efficient methods for online multiclass logistic regression.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text.

Ebubekir Buber and Banu Diri. 2018. Performance analysis and cpu vs gpu comparison for deep learning. pages 1–6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. Mulan: A joint embedding of music audio and natural language.

Sergio Oramas, Francesco Barbieri, Oriol Nieto, and Xavier Serra. 2018. Multimodal deep learning for music genre classification. *Trans. Int. Soc. Music. Inf. Retr.*, 1:4–21.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Alexandros Tsaptsinos. 2017. Lyrics-based music genre classification using a hierarchical attention network.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. Musicbert: Symbolic music understanding with large-scale pre-training.