# Punk or Funk: Understanding the Performance of RoBERTa on Music Genre Classification

Stanford CS224N Custom Project

**Andrew Bempong**
Department of Computer Science
Stanford University
bempong@stanford.edu

**Deveen Harischandra**
Department of Computer Science
Stanford University
deveen@stanford.edu

## Abstract

Assigning a genre to a piece of music has always been a topic of interest to consumers and music industry tycoons alike. Traditionally, genre classification has been a task for human experts, who rely on their extensive knowledge and interpretation of musical elements. Humans excel in this task due to their ability to understand context, emotion, and subtle nuances. However, as the volume of music data grows, automated, scalable solutions are increasingly needed. AI offers the potential to process and classify vast amounts of music data quickly and consistently. While AI lacks the intuitive grasp of music that humans possess, it compensates by identifying patterns and correlations that may not be immediately obvious to human listeners.

This project aims to explore the limitations of transformer-based language models concerning their performance in music genre classification based on a variety of musical characteristics. Specifically, we evaluate the contributions of a model inspired by RoBERTa to perform this task via investigative feature ablation. Motivated by the desire to demystify neural networks, we aim to improve their interpretability through an ablation study, assessing if removing specific features, neurons, and layers enhances classification accuracy. We plan to ablate RoBERTa layers, position embeddings, and attention heads, comparing performance with other models. Key metrics for our findings include accuracy, recall, and F1 score.

## 1 Key Information to include

- Mentor: Olivia Lee
- External Collaborators (if you have any): No
- Sharing project: No

## 2 Introduction

Classifying music genres based on various musical characteristics (such as tempo, valence, lyrics and loudness) is a challenging task due to the complex and nuanced nature of music, which vary greatly in style, language, and content. Traditional methods often rely on audio features, overlooking the rich textual information in the music and lyrics itself. This project explores the limitations and potential of transformer-based language models, specifically models such as RoBERTa, for this task. Our motivation for this project stems from our passion for music, as we are amateur musicians ourselves. By understanding how these models interpret and classify music, we can gain valuable insights into the elements that define different genres. This knowledge not only fuels our academic curiosity but also helps us improve as musicians, enabling us to create more nuanced and genre-defying compositions. Furthermore, exploring the intersection of music and advanced machine learning techniques allows us to innovate and push the boundaries of how music is understood and categorized.

While transformer models like RoBERTa have achieved state-of-the-art performance in various NLP tasks, their application to music genre classification remains underexplored. Additionally, these models often act as black boxes, offering little insight into their decision-making processes, which is a significant drawback.

Our project conducts an investigative feature ablation study on our model which is based on RoBERTa, systematically removing specific features, neurons, and layers to understand their contributions to the model's performance and also explores the ways in which such models can be improved. This approach aims to enhance interpretability and potentially improve classification accuracy by identifying and eliminating redundant components.

The baseline for our study is a comprehensive implementation of a model inspired by RoBERTa with Spotify lyric embeddings, showing a development accuracy of 0.380. We are introducing new characteristics of music such as 'energy', 'tempo', and 'instrumentation' into our adjusted version of RoBERTa to help improve the accuracy of genre classification. We ablate features such as RoBERTa layers, position embeddings, and attention heads, and compare the performance of these modified models to the baseline using metrics like accuracy, recall, and F1 score.

By introducing new musical characteristics as features into the model and advancing the interpretability of neural networks through the ablation study, we aim to achieve better results of within the sphere of genre classification and provide valuable insights into the workings of transformer models in music genre classification. This understanding can lead to the development of more efficient and effective models, benefiting the broader field of NLP and its applications in music information retrieval.

## 3 Related Work

Our project draws inspiration from the multi-layer feature ablation work in BERT models, particularly as explored by Zhao et al. in their study on stock trend prediction. Zhao et al. address the limitations of the BERT model, which primarily uses features from its last layer, by employing a multi-layer feature ablation approach. They extract and analyze features from each BERT layer to improve topic recognition accuracy in stock comments.

In their study, Zhao et al. use a sliding window technique to divide long texts into shorter segments, ensuring comprehensive information capture and reducing overfitting. They then perform a multi-layer feature ablation study, fusing features from various layers to identify the most effective combinations for text classification. Their results show significant performance improvements in BERT's topic recognition capability when utilizing a more comprehensive feature set from multiple layers.

In our project, we adopt a similar ablation strategy to investigate the contributions of different components within our RoBERTa inspired model for music genre classification based on various musical characteristics. Zhao et al.'s study serves as a foundational reference, demonstrating the importance of multi-layer feature analysis in improving transformer-based models' text classification performance. We extend this approach to music genre classification, leveraging these insights to refine our model for better accuracy and interpretability.

Focusing on music genre classification, our project builds on the work by Kostrzewa et al., which investigates various deep learning models for music genre classification using the Free Music Archive (FMA) dataset. They evaluated the performance of Convolutional Neural Networks (CNNs), Convolutional Recurrent Neural Networks (CRNNs) with 1D and 2D convolutions, and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) cells. By proposing ensemble methods to combine different architectures, they achieved state-of-the-art results.

Kostrzewa et al. focused on classifying music tracks using deep learning models, utilizing mel-spectrograms and MFCCs as input features. Their work demonstrated the effectiveness of ensemble methods in enhancing classification accuracy.

In our project, we extend this approach to transformer-based models like RoBERTa, incorporating musical characteristics and other varied nuances and focusing on interpretability through an ablation study. Inspired by Kostrzewa et al.'s use of deep learning architectures and ensemble methods, we aim to refine and optimize our model for better music genre classification.

# 4 Approach

The preprocessing step aimed at preparing the dataset for training a machine learning model to classify music tracks into their respective genres. The dataset is initially loaded from a CSV file into a pandas DataFrame, ensuring it is structured in a tabular format for easy manipulation and analysis. Relevant features such as 'popularity', 'danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', and 'tempo' are selected for the classification task, with 'track genre' identified as the target variable. The 'track genre' labels are then encoded into numerical values using 'LabelEncoder', making them suitable for machine learning models that require numerical input.

Next, the pandas DataFrame is converted into a Hugging Face Dataset, which supports efficient data manipulation and processing. A preprocessing function is defined to structure the dataset for model input, separating features and labels and creating a dictionary where the features are mapped to their respective values, and the labels are assigned to the 'labels' key. This function is applied to the dataset in a batched manner, transforming the entire dataset accordingly. The transformed dataset is then split into training and testing sets with an 80-20 split, ensuring that the model can be trained on one subset of the data and evaluated on another to assess its performance and generalization capability. Finally, the preprocessed training and testing datasets are returned, ready for the subsequent model training phase.

Afterward, we aimed to incorporate this preprocessed training dataset into a reconfigured RoBERTa baseline implementation with a sequence classification head. This model was chosen since it's quite a bit more robust than BERT with regard to text embedding within the context of a transformer architecture. We first tokenized the text features from our dataset to be compatible inputs for our RoBERTa baseline. Then, we added a classification head to Hugging Face's 'roberta-base' model to output the likelihoods for each possible track genre and thus find the highest. We trained this model with Hugging Face's 'Trainer' class, which streamlined the process of iterative training, optimization, and evaluation, which was highly beneficial given how repetitive this process became for different ablations/variations of the original baseline. During a fine-tuning stage, we configured a few hyperparameters (i.e. number of epochs, batch size, and learning rate) through a random search process.

Finally, we performed a study to optimize model performance from the baseline where we'd chiefly involve ablation of features of interest with some room for augmentation/feature addition. Originally an ablation study, we'd go on to methodically excise several features—first at random, then varying the hidden layer size of noticeably high-impact features, and noticing how the absence of their hidden layers entirely affected model performance. However, we noticed there was still an opportunity to improve the model via augmentation, so we also pursued cycling through different activation functions, as well as adding additional features (e.g. track length, explicit content, etc.) to enhance the model's general comprehension of the task at hand. From here, we'd train each ablated/augmented model and compare the evaluated genre classification performance of ones of interest, which were those that significantly deviated from the baseline.

# 5 Experiments

## 5.1 Data

The dataset used in this project, the 'spotify-tracks-dataset', is a comprehensive collection of music tracks, each described by various features that are crucial for understanding the track's characteristics made publicly available through Kaggle. The primary task associated with this dataset is to classify each track into its respective genre based on these features. Each entry in the dataset includes a unique identifier for the track ('track id'), the name of the artist or artists who performed the track ('artists'), the name of the album the track is part of ('album name'), and the title of the track ('track name'). Additionally, the dataset provides a popularity score ('popularity'), which is a numerical value representing how well-liked the track is.

The dataset also includes detailed audio features such as the duration of the track in milliseconds, and an 'explicit' flag indicating whether the track contains explicit content. Key musical characteristics are represented by 'danceability', which measures how suitable a track is for dancing, and 'energy', which gauges the intensity and activity of the track. The 'key' and 'mode' provide information

about the musical key and modality (major or minor) of the track, while ‘loudness‘ gives the overall loudness level in decibels.

Further features include ‘speechiness‘, which indicates the presence of spoken words in the track, ‘acousticness‘, reflecting the acoustic nature of the track, and ‘instrumentalness‘, which estimates the likelihood that the track is instrumental. The ‘liveness‘ feature captures the presence of a live audience in the recording, while ‘valence‘ describes the musical positiveness conveyed by the track. Finally, ‘tempo‘ measures the speed or pace of the track, and ‘time signature‘ indicates the track’s time signature.

The target variable in this dataset is ‘track genre‘, a categorical variable representing the genre of each track. The objective is to train a machine learning model to predict the genre of a track based on the provided features. The input to the model consists of the aforementioned numerical and categorical features, while the output is the predicted genre. By learning from the patterns and characteristics in the training data, the model aims to accurately classify new tracks into their respective genres. This task involves understanding the intricate relationships between the audio features and the genres, enabling the model to generalize well to unseen data.

### 5.2 Evaluation method

For each version of the reconfigured RoBERTa model—ablated, augmented, or not—we found the raw accuracy as well as the F1 score via looping through their post-training outputs. For the F1 score, this was computed via the precision and recall by counting off true positives and false negatives/positives and found their harmonic mean. Specifically:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

to calculate prediction performance such that:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

We took advantage of the Hugging Face "Trainer" class in order to efficiently calculate these values with a custom metrics function.

### 5.3 Experimental details

We performed our experiments using the Pytorch, scikit-learn, and the Hugging Face ’transformers’ libraries. To initialize our dataset, we loaded in the CSV files we preprocessed from the ’spotify-tracks-dataset’ as pandas DataFrames for ease of manipulation. We then utilized scikit-learn’s StandardScaler and LabelEncoder to standardize features of interest and encode track genres as target classifying variables, respectively. Furthermore, track features were tokenized and made into usable text descriptions via the RobertaTokenizer.

To configure our implementation of RoBERTa for the performance task of outputting genre likelihoods for tracks, we modified the Hugging Face library’s pre-trained RobertaForSequenceClassification with a classification head customized for our ablation trials. The classification head is essentially a ReLU activation function layer in between two connected layers, a hidden layer and an output layer for the genre class probabilities. The ’transformers’ library allotted us the opportunity to not only leverage this pre-trained RoBERTa model, but also to make use of the Adam optimization algorithm and cross-entropy loss since we aimed to optimize classifications between several "classes" (in our case, genres).

Generally, the Adam update rule was defined as:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{4}$$

where $\theta_t$ are the model parameters at iteration $t$, $\eta$ is the learning rate, $\hat{m}_t$ is the bias-corrected first moment estimate, and $\hat{v}_t$ is the bias-corrected second moment estimate. These are defined as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{5}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{6}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{7}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{8}$$

where $g_t$ is the gradient at iteration $t$, $\beta_1$ and $\beta_2$ are the decay rates for the moment estimates, typically set to 0.9 and 0.999, respectively.

For our multi-class classification use of cross-entropy loss, we defined the equation as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{9}$$

where $N$ is the number of track genres, $y_i$ is the true label (1 if the class is correct, 0 otherwise), and $\hat{y}_i$ is the predicted probability for genre $i$.

Training and fine-tuning itself leveraged the Hugging Face library's 'trainer' class for its ease of consolidating iterative training/optimization and evaluation methods. Our hyperparameters (chiefly the number of epochs and learning rate) were tuned by the use of a random search process via the hyperparameter capabilities of scikit-learn. As a result of this random search we conducted, the baseline RoBERTa model's configuration essentially consisted of 20 epochs, a batch size of 16, a learning rate of 0.001, and a weight decay of 0.01,

Then, we essentially performed an ablation study to understand the impact of each feature, layer, and parameter. Each one of these was methodically excised and then we'd retrain/re-evaluate the ablated model to compare performance. We additionally augmented the baseline model on a few trials to evaluate the addition of features and its impact on performance. The swapping of activation functions from ReLU was also taken into account to gain a bird-eye-view of the non-linearity of the model.

### 5.4 Results

Table 1: Evaluation Results

| Ablation | Accuracy (%) | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Baseline | 62.69 | 0.64 | 0.68 | 0.66 |
| Without popularity | 48.69 | 0.51 | 0.59 | 0.55 |
| Without danceability | 51.81 | 0.50 | 0.55 | 0.52 |
| Without energy | 50.64 | 0.51 | 0.56 | 0.53 |
| Without loudness | 57.25 | 0.52 | 0.58 | 0.55 |
| Without speechiness | 57.10 | 0.59 | 0.65 | 0.62 |
| Without acousticness | 57.64 | 0.57 | 0.63 | 0.60 |
| Without instrumentalness | 57.29 | 0.60 | 0.66 | 0.63 |
| Without liveness | 58.50 | 0.53 | 0.59 | 0.56 |
| Without valence | 57.29 | 0.57 | 0.62 | 0.59 |
| Without tempo | 58.24 | 0.57 | 0.61 | 0.59 |
| With explicit lyrics | 57.84 | 0.61 | 0.66 | 0.63 |
| With track duration | 57.67 | 0.56 | 0.62 | 0.59 |
| Hidden Dim 64 | 57.33 | 0.57 | 0.63 | 0.60 |
| Hidden Dim 32 | 56.06 | 0.52 | 0.57 | 0.54 |
| Hidden Dim 16 | 52.84 | 0.53 | 0.59 | 0.56 |
| No Hidden Layer | 47.02 | 0.48 | 0.53 | 0.51 |
| Activation Tanh() | 53.11 | 0.54 | 0.60 | 0.57 |
| Activation Sigmoid() | 52.50 | 0.59 | 0.65 | 0.62 |

Detailed in Table 1, our findings resulted in a comparison of different altered versions of our RoBERTa-based model. Specifically, each model is either feature ablated, feature augmented, and

modified activation functions. We hypothesized that the majority of features may not have as much impact as select few, as we suspected that the most important features would be 'energy', 'loudness', 'instrumentalness', and 'liveness'—since these consist of some of the most stark differences between genres. For instance, rock music has much more energy and loudness than RnB.
Although the general concept of a select few salient features seemed to survive the rigors of evaluation, the specific features we hypothesized as being highly impactful weren't exactly what we thought. To our surprise, it was 'energy' and 'danceability' that seemed to be the most important features in terms of performance. This can be attributed to the considerable drop in performance of the model exhibited by the removal of either feature, with the 'without energy' F1 score being 0.53 and 'without danceability' being 0.52, which are some of the worst performances when compared to the baseline 0.66 F1 score.
In terms of layers, we hypothesized that certain components of the hidden layer were more important than others—and thus, the more hidden layer dimensions we removed, at some point performance would peak after a certain number of ablations and would steadily decline afterwards. It seems as though performance peaked with the hidden layer in-tact as it steadily declined with smaller and smaller dimensions, decreasing in term of 16s to depict the steep decline. In fact, the lowest F1 score achieved was when the hidden layer was completely ablated.
We also believed that the addition of other assumed salient features would cause a noticeable boost in model performance. For instance, we hypothesized that the augmentation of the model via the addition of other features such as explicit lyrical content and the length of the track would result in higher F1 scores than the baseline. However, these did not seem to make much of an impact, as they resulted in F1 scores 0.63 and 0.59 respectively, which were not too far off from the baseline. These were assumed to be high impact features due to the same line of reasoning as the ablated features. Additionally, the swapping of activation functions seemed to only bolster the fact that ReLU performed the best for this multi-genre classification task, which is about what we expected.
Overall, I believe these findings may mean our approach generally had a grasp on what sort of concepts and features contribute to the contrasting elements of different musical genres, as well as a comprehensive grasp of the capabilities of our RoBERTa base model to identify such. However, I believe our approach could have with more time as we could have delved more into the high impact features from different perspectives (for instance, evaluating in relation to other models).

## 6 Analysis

In terms of qualitative evaluation, this is highly influenced by the fact that we took advantage of a pre-trained version of RoBERTa from Hugging Face and how we're stressing its ability to handle multi-class identification. Given our F1 Score, our model is definitely able to pick up on several trends and patterns within the data, but there is still opportunity to boost the model's performance.

Given the salient features of our model 'energy' and 'danceability', it seems as though this model is at its best when classifying between genres that are starkly different with undeniably defining characteristics. For instance, hip-hop and electronic dance music were often correctly identified most likely as they are distinct in their lack of live instrumentation when compared with other genres of music. Although, it became increasingly obvious that genres with vague definitions or broad catelogs seemed to be victim to frequent misclassification. For instance, many experimental/alternative genres were confused with each other and broader genres such as pop and rock were often mixed up with themselves (and sometimes with other genres). We believe that this is only worsened by the fact that many of the features could have cross-over ('energy', 'loudness', and 'valence' could be argued to mean the same thing) and the fact that audio features in general can be vague and lossy.

The ablation study only seems to underscore these capabilities/inefficiencies between feature salience and outputs. As per the results, ablating important features such as 'energy' or 'danceability' decreases model performance, but this could be because the majority of our testing data had to do with electronic music (dance, rap, hip hop, afrobeats, etc.) so features like 'acousticness' and 'instrumentalness' may have been better for testing sets where classical genres were dominant. The model's high complexity is also necessary, as exhibited by the need for ReLU activation function and the drop in performance with the ablated hidden layer. Perhaps the combination of the ambiguous nature of audio feature descriptions, the structure of the testing data, and the model's strength led to

these qualattitive results that could serve as room for improvement for potential future trials and studies in the research space.

## 7 Conclusion

In this project, we aimed to explore the potential of transformer-based language models, specifically RoBERTa, for music genre classification based on various musical characteristics such as 'energy' and 'tempo' and also lyrical content. Through an extensive ablation study, we assessed the impact of various features, layers, and activation functions on the model's performance. Our findings highlighted the critical importance of features like 'energy' and 'danceability' for distinguishing between genres, especially those with starkly different characteristics such as hip-hop, electronic dance music, and afro-beats.

Our baseline model achieved an F1 score of 0.66, demonstrating its capability to identify several trends and patterns within the data. However, the removal of key features like 'energy' and 'danceability' led to significant drops in performance, with F1 scores plummeting to 0.53 and 0.52 respectively. This underscores the model's reliance on these features for accurate classification. Conversely, features like 'acousticness' and 'instrumentalness' appeared less impactful, likely due to the nature of our testing dataset, which was dominated by electronic music genres.

The ablation of hidden layers further revealed the model's dependency on its complexity. Performance degraded steadily as hidden layer dimensions were reduced, with the lowest F1 score observed when the hidden layer was completely ablated. This confirmed the necessity of maintaining a certain level of model complexity to preserve performance.

Despite these achievements, our study also uncovered several limitations. Genres with vague definitions or broad catalogs, such as experimental or alternative genres, frequently suffered from misclassification. This was likely exacerbated by overlapping features like 'energy', 'loudness', and 'valence', which can be ambiguous and lossy in nature. Additionally, the pre-trained RoBERTa model from Hugging Face, while powerful, showed varying levels of accuracy and efficiency depending on the specific characteristics of the test data.

In summary, our project successfully demonstrated the application of our RoBERTa-inspired model for music genre classification and provided valuable insights into the importance of specific features and model complexity. However, future work could focus on refining feature selection, incorporating a more diverse dataset, and exploring alternative models or ensemble methods like we saw in the project conducted by Kostrzewa et al, to enhance classification accuracy. Further research might also delve into improving the interpretability of the model, potentially through more sophisticated ablation studies or the integration of explainable AI techniques.

## 8 Ethics Statement

One significant ethical challenge that could be potentially related to this project is the possibility for bias in the dataset, which can lead to unfair or inaccurate genre classifications. If certain genres, artists, or types of music are underrepresented, the model may not perform well for those categories, reinforcing existing stereotypes and cultural biases. This may particularly be the case when it comes to international music where music is heavily influenced by culture and tradition. For example, genres such as Indian Bollywood music or Afro-Beats music maybe inaccurately classified by this model due to insufficient data and representation of it during the training process. Further, some of the features and characteristics we are using in our model to classify genres such as 'tempo' and 'valence' may not translate across various types of music and therefore may lead to inaccurate classification as well. Since, music is deeply tied to cultural identities, incorrect classification can lead to cultural insensitivity. Misclassification of genres can also impact artists' reputations and listeners' perceptions, potentially affecting how music is marketed and received. Finally, there are concerns grounded in privacy and data usage; it's crucial to ensure that the data used complies with privacy regulations and is collected with proper consent.

To mitigate these risks, using a diverse and representative dataset that includes a wide range of genres, artists, and cultural backgrounds is essential. Ensuring that we regularly audit the dataset for balance can help address bias. Implementing transparency and explainability in the model's decision-making

process can identify and correct biases.

Further, prioritizing ethical data collection practices, including obtaining proper consent and anonymizing data where possible, is critical for privacy compliance. Continuous monitoring of the model's performance and gathering user feedback can help refine the model and ensure it remains fair and accurate. We can also implement educational seminars or workshops to educate the development team on cultural sensitivity, which can guide informed decision making during dataset curation and model development. Conducting ablation studies to systematically understand the impact of various features or data points on the model's predictions can further enhance the model's robustness and fairness. These strategies collectively aim to develop a fairer, more accurate, and culturally sensitive music genre classification model, aligning the project with current ethical standards and societal expectations.

## References

## A Appendix (optional)

## References


[1] Huang, Zihan, Low, Charles, Teng, Mengqiu, Zhang, Hongyi, Ho, Daniel, Krass, Mark, and Grabmair, Matthias. (2021). Context-Aware Legal Citation Recommendation using Deep Learning. In *Association for Computational Linguistics (ACL)*.

[2] Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Delangue, Clement, Moi, Anthony, Cistac, Pierric, Rault, Tim, Louf, Rémi, Funtowicz, Morgan, Davison, Joe, Shleifer, Sam, von Platen, Patrick, Ma, Clara, Jernite, Yacine, Plu, Julien, Xu, Canwen, Le Scao, Teven, Gugger, Sylvain, Drame, Mariama, Lhoest, Quentin, and Rush, Alexander M. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

[3] Kostrzewa, Daniel, Piotr Kaminski, and Robert Brzeski. "Music genre classification: looking for the Perfect Network?" Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland. ICCS Camera Ready Version 2021. DOI: 10.1007/978-3-030-77961-0_6.

[4] Zhao, Feng, et al. "Multi-layer Features Ablation of BERT Model and Its Application in Stock Trend Prediction." Expert Systems With Applications, vol. 207, 2022, p. 117958. DOI: 10.1016/j.eswa.2022.117958.