# Transformer-Based Song Genre Classification

University of Pennsylvania
CIS 5300: Natural Language Processing

Henghak Kun, Sandeep Alankar, Sophie Luo, Tanvi Gupta

## Abstract

In this paper, we explore the task of predicting a song's genre based on its lyrical content, employing transformer-based natural language processing models. A baseline Long Short-Term Memory (LSTM) model is established using GloVe embeddings and subsequent transformer-based architectures including DistilBERT, BERT, and RoBERTa are implemented to enhance predictive performance. Notably, the incorporation of sentiment analysis, particularly with the DistilBERT model, yields superior results. Leveraging sophisticated language representations, our findings underscore the efficacy of transformer models in capturing nuanced relationships between lyrics and music genres, thereby exploring the limits of transformer-based models in genre classification within the domain of song analysis.

## 1 Introduction

Music, an art form synonymous with human expression, often transcends the auditory experience to reflect cultures, historical narratives, and emotional landscapes. In the domain of Natural Language Processing (NLP), the classification of music based on lyrics presents an intriguing challenge. The task involves deciphering the thematic and stylistic elements embedded within song lyrics to predict the genre - a multifaceted problem that intersects text analysis with music categorization.

At the heart of our endeavor lies the goal of enhancing music recommendation systems. Current streaming platforms, despite their sophistication, frequently overlook the granular genre distinctions that could facilitate deeper artistic exploration. Recognizing this gap, our project sets out to harness the capabilities of advanced language models to interpret song lyrics and assign accurate genre labels. This effort not only embodies an application of NLP in the entertainment industry but also promises to enrich the user experience by enabling more nuanced discovery and recommendation features.

Formally, the problem can be defined as follows: given a set of song lyrics, the task is to output a genre label from a predefined set of categories (10 genres in our project). The input is the textual content of a song, and the output is a genre classification such as Rock, Jazz, Hip-Hop, etc. This classification task can be framed as a supervised learning problem where the model learns from a labeled dataset of song lyrics associated with genre tags. This is also shown visually by the diagram below:

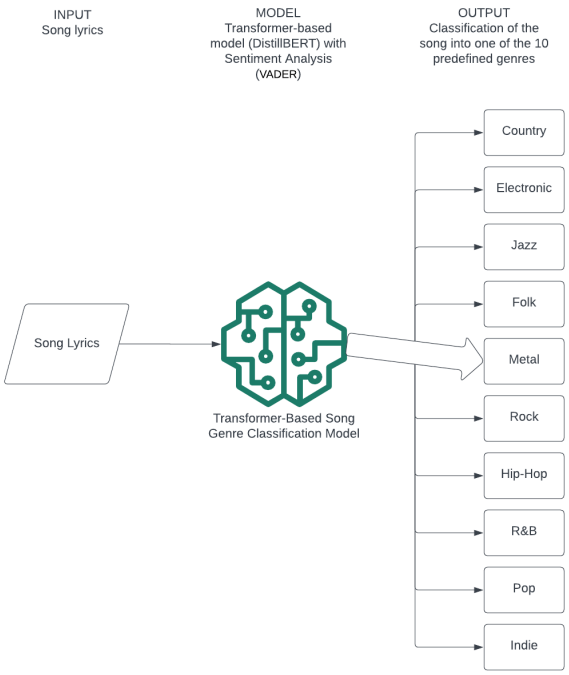


Figure 1: Illustration of Project

Our project is driven by a deep-seated passion for music and a strong foundation in language model expertise. Confronted with the limitations of existing music streaming platforms, particularly in genre-specific song recommendations, we were mo-

tivated to harness our collective experience with diverse musical genres and our technical skills in NLP. Our goal is to develop a model that not only accurately predicts music genres from lyrics, bridging the gap between text analysis and music classification, but also enhances the music discovery experience, enabling users to explore and connect with songs in a more meaningful and personalized way.

## 2 Literature Review

The endeavor of classifying music genres based on song lyrics intersects the domains of NLP, computational linguistics, and music information retrieval, challenging researchers to extract and interpret the nuanced linguistic features embedded within lyrics.

1. **MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training** (1): This study introduces MusicBERT, a transformer-based model tailored for symbolic music understanding. Employing OctupleMIDI encoding and a novel masking strategy, MusicBERT demonstrates exceptional performance in tasks including genre classification. The model's architecture and pre-training on the extensive MMD dataset significantly contribute to its capability to discern genre-specific features from symbolic music representations. This paper sets a precedent for using transformer models in music genre classification, aligning with our project's initial approach using RoBERTa.

2. **BECMER: A Fusion Model Using BERT and CNN for Music Emotion Recognition** (2): In this research, Sung et al. explore the amalgamation of audio and lyric analysis for music emotion recognition. The study emphasizes the impact of lyrics on emotion classification, employing NLP models such as BERT and ALBERT for textual analysis. The fusion model proposed showcases how combining different modalities enhances music classification, a concept that resonates with our project's aim to leverage textual information for genre prediction.

3. **Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network** (3): This paper explores the use of a Hierarchical Attention Network (HAN) for genre classification from song lyrics. The HAN's unique design mirrors the layered structure of lyrics and incorporates attention mechanisms at multiple levels, making it adept at identifying genre-defining linguistic elements. This approach of using hierarchical structures and attention mechanisms is particularly relevant to our project as it aligns with our aim to capture the contextual and stylistic nuances in lyrics for genre classification.

4. **Predicting song genre with deep learning** (4): This recent study explores various machine and deep learning models for song genre classification, including transformer-based models like BERT and XLM-RoBERTa. The findings indicate that deep learning models, when trained on comprehensive data sets, can significantly outperform traditional machine learning approaches in classifying music genres.

The mentioned studies collectively highlight the evolving landscape of music genre classification, underscoring the potential of leveraging NLP techniques for this task. As our project transitions from using transformer models to implementing LSTM with GloVe embeddings, insights from these papers, particularly regarding the importance of context and hierarchical structures in text, remain pertinent.

## 3 Experimental Design

### 3.1 Data

We used data from a Kaggle dataset Multilingual Lyrics for Genre Classification (5). This dataset is composed of three other Kaggle datasets: 150K Lyrics Labeled with Spotify Valence, dataset lyrics musics, and AZLyrics song lyrics. The training set is composed of 290,000 entries that contain song names, artists, genres, lyrics, and languages. Filtering down to include songs only in English, we have 212,000 entries and we extract each song's genre and lyrics. There are 10 unique genres and the length of the lyrics vary from 2 words to over 1500 words. The genre distribution is heavily imbalanced, with Rock and Pop being over three-quarters of the entire training set. For the sake of better generalization, we only used 10,000 entries for training, with roughly 1,000 entries of each genre. Our validation set consists of 1,800 entries in total, with roughly 180 entries per genre. The dataset also comes with a testing set that contains 7935 entries

and has an imbalanced genre distribution like the training set. We used this dataset to assess how well our model would perform against an imbalanced sample.

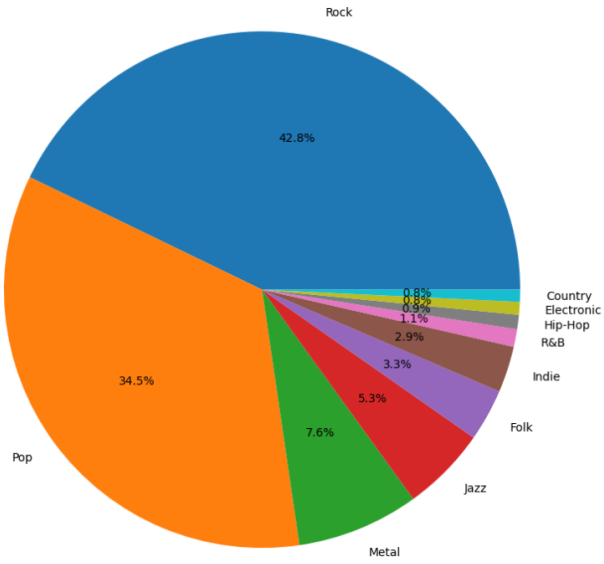


Figure 2: Data Genre Distribution

### 3.2 Evaluation Metrics

For evaluation, we used four metrics: accuracy, precision, recall, and F1 score. Accuracy is the total number of labels the model got correct over the total number of predicted labels. For each of the 10 possible genres, we keep track of the true positives, true negatives, and false positives. Then we sum the true positives (tp), true negatives (tn), and false positives (fp) for all classes and compute the precision, recall, and F1 score for each genre. In other words, for each genre, we compute

$$Precision_i = \frac{tp_i}{tp_i + fp_i}$$

$$Recall_i = \frac{tp_i}{tp_i + fn_i}$$

$$F1_i = \frac{2 * precision_i * recall_i}{precision_i + recall_i}$$

In addition, we also compute the weighted-average and macro-average precision, recall, and F1 score for all genres. This is

$$Precision_{macro} = \frac{tp_{total}}{tp_{total} + fp_{total}}$$

$$Recall_{total} = \frac{tp_{total}}{tp_{total} + fn_{total}}$$

$$F1_i = \frac{2 * precision_{total} * recall_{total}}{precision_{total} + recall_{total}}$$

$$Precision_{weighted} = \frac{\sum_i count_i * Precision_i}{count_{total}}$$

$$Recall_{weighted} = \frac{\sum_i count_i * Recall_i}{count_{total}}$$

$$F1_{weighted} = \frac{\sum_i count_i * F1_i}{count_{total}}$$

### 3.3 Simple Baseline

We wanted to start with the simplest baseline possible. Our simple baseline approach for classifying song genres was a simple majority class method, identifying the most prevalent genre in the dataset as the predicted genre for all songs. Analysis of genre distribution revealed "Rock" as the dominant genre, comprising approximately 17.77% of our dataset.

We assigned the "Rock" label to every song across the training, validation, and test datasets. This method served as a fundamental benchmark, reflecting the proportion of "Rock" songs in the dataset.

This simple baseline yielded a result of roughly 10% in both the training and dev set as we balanced the dataset and it yielded a 17.77% in test accuracy (for the full results, consult figure 3 and 4). This low accuracy highlights the limitations of the majority class baseline methods in handling genre diversity and underscores the necessity for more advanced models capable of nuanced genre classification.

| Genre | Count | Precision | Recall | F1_score |
|---|---|---|---|---|
| Indie | 169 | 0 | 0 | 0 |
| Folk | 175 | 0 | 0 | 0 |
| R&B | 186 | 0 | 0 | 0 |
| Pop | 201 | 0 | 0 | 0 |
| Rock | 184 | 0.102222 | 1 | 0.0927419 |
| Metal | 168 | 0 | 0 | 0 |
| Country | 183 | 0 | 0 | 0 |
| Hip-Hop | 163 | 0 | 0 | 0 |
| Jazz | 184 | 0 | 0 | 0 |
| Electronic | 187 | 0 | 0 | 0 |

Overall
Accuracy: 0.10222222222222223
Weighted Precision: 0.010449382716049383
Weighted Recall: 0.10222222222222223
Weighted F1 Score: 0.009480286738351256
Macro Precision: 0.10222222222222223
Macro Recall: 0.10222222222222223
Macro F1 Score: 0.05111111111111111

Figure 3: Majority Dev-Set Results

| Genre | Count | Precision | Recall | F1_score |
|---|---|---|---|---|
| Indie | 510 | 0 | 0 | 0 |
| Folk | 495 | 0 | 0 | 0 |
| R&B | 510 | 0 | 0 | 0 |
| Pop | 1110 | 0 | 0 | 0 |
| Rock | 1410 | 0.177694 | 1 | 0.150883 |
| Metal | 810 | 0 | 0 | 0 |
| Country | 810 | 0 | 0 | 0 |
| Hip-Hop | 960 | 0 | 0 | 0 |
| Jazz | 660 | 0 | 0 | 0 |
| Electronic | 660 | 0 | 0 | 0 |

Overall
Accuracy: 0.1776937618147448
Weighted Precision: 0.031575072987875256
Weighted Recall: 0.1776937618147448
Weighted F1 Score: 0.026810936774616392
Macro Precision: 0.1776937618147448
Macro Recall: 0.1776937618147448
Macro F1 Score: 0.08884688090737239

Figure 4: Majority Test-Set Results

# 4 Experimental Results

## 4.1 Strong Baseline

One of the published models that was used to do genre classification is the Long Short-Term Memory (LSTM) model. The reason why LSTM might perform well is due to its ability to remember and form connections between a string of sequential words. For our strong baseline, we implemented an LSTM model with a 50-dimensional GloVe Embedding from the Huggingface StanfordNLP repository. Our model consists of an embedding layer (roughly 36k parameters), followed by the LSTM layer from Pytorch, and then finally a linear layer to map into 10 outputs for the genre classification. This model gave us a validation accuracy of 23.61% and a testing accuracy of 23.62% (for the full metrics, consult figure 5 and 6). In the paper by Bagić Babac, they were able to get an accuracy of about 54% with LSTM with GloVe embeddings. One reason is the fact that we use different datasets. The dataset that was used in the paper consists of 5 genres to classify, while in our dataset, there are 10 genres. The difference in the number of genres makes it hard to compare results.

| Genre | Count | Precision | Recall | F1_score |
|---|---|---|---|---|
| Hip-Hop | 164 | 0.552795 | 0.542683 | 0.273846 |
| Rock | 186 | 0.110345 | 0.172043 | 0.0672269 |
| Indie | 185 | 0.155039 | 0.108108 | 0.0636943 |
| Pop | 185 | 0.22 | 0.118919 | 0.077193 |
| Metal | 184 | 0.274611 | 0.288043 | 0.140584 |
| R&B | 167 | 0.221311 | 0.161677 | 0.0934256 |
| Country | 183 | 0.2 | 0.224044 | 0.10567 |
| Jazz | 190 | 0.348548 | 0.442105 | 0.194896 |
| Electronic | 189 | 0.183206 | 0.126984 | 0.075 |
| Folk | 167 | 0.144737 | 0.197605 | 0.0835443 |

Overall
Accuracy: 0.2361111111111111
Weighted Precision: 0.23870737810538922
Weighted Recall: 0.2361111111111111
Weighted F1 Score: 0.11635736238849749
Macro Precision: 0.2361111111111111
Macro Recall: 0.2361111111111111
Macro F1 Score: 0.11805555555555555

Figure 5: LSTM Dev-Set Results

| Genre | Count | Precision | Recall | F1_score |
|---|---|---|---|---|
| Hip-Hop | 960 | 0.664642 | 0.569792 | 0.306786 |
| Rock | 1410 | 0.211915 | 0.176596 | 0.096325 |
| Indie | 510 | 0.0723684 | 0.0862745 | 0.039356 |
| Pop | 1110 | 0.211864 | 0.0900901 | 0.0632111 |
| Metal | 810 | 0.301994 | 0.392593 | 0.170692 |
| R&B | 510 | 0.124444 | 0.109804 | 0.0583333 |
| Country | 810 | 0.202105 | 0.237037 | 0.109091 |
| Jazz | 660 | 0.220794 | 0.286364 | 0.12467 |
| Electronic | 660 | 0.128165 | 0.122727 | 0.0626935 |
| Folk | 495 | 0.106987 | 0.19798 | 0.0694543 |

Overall
Accuracy: 0.23616887208569629
Weighted Precision: 0.247510020884755
Weighted Recall: 0.23616887208569629
Weighted F1 Score: 0.11783029097960623
Macro Precision: 0.23616887208569629
Macro Recall: 0.23616887208569629
Macro F1 Score: 0.11808443604284814

Figure 6: LSTM Test-Set Results

## 4.2 Extensions

For our first extension, we decided to use the RoBERTa (Robustly optimized BERT approach) model due to its performance and efficiency advantages over vanilla BERT and DistilBERT. Modifications like removing the next sentence prediction task and use of larger mini-batches lead to faster convergence during training. RoBERTa's improved contextual understanding and pretrained models can capture nuanced language patterns and aid in deciphering genre-specific features. It can accommodate longer input sequences, which, in theory, would make it a compelling choice for genre classification from song lyrics.

While there was a significant improvement in dev set accuracy, we observed little to no improvement in test set performance when compared to our LSTM baseline. We hypothesize this to be due to overfitting, a common challenge in machine learning particularly with complex models like RoBERTa. We reduced our original training dataset from 212k lyrics to 10k to allow for faster training times. Using a model like RoBERTa on this dataset, with 125M parameters as compared to DistilBERT's 66M, would theoretically increase risk of overfitting since we are using a larger model on a smaller training set. This would explain why even though RoBERTa had a higher development set accuracy when compared to DistilBERT, DistilBERT had a much higher test set accuracy.

Additionally, in an attempt to improve the model's performance, we decided to add a TF-IDF feature. However, while our dev set accuracy was 53.3%, we observed a slight drop in the test set performance of RoBERTa with TF-IDF. We attribute this to the fact that RoBERTa itself captures rich semantic representations of words and phrases through its contextual embeddings. If the TF-IDF features overlap significantly with the information already captured by RoBERTa, it may introduce redundancy or even noise, explaining the observed decrease in performance.

Keeping in mind the results of our first extension, and the hypothesis we had about overfitting, we decided to perform our second extension on DistilBERT instead of RoBERTa. We hypothesized that DistillBERT's 66M parameters would reduce the risk of overfitting and give us a more stable model. Moreover, we decided to implement sentiment analysis with our DistilBERT model. This decision was rooted in the understanding that differ-

ent music genres often convey distinct emotional tones and themes, which sentiment analysis can effectively capture. For instance, genres like metal or certain subgenres of rock might exhibit more intense or somber lyrics, while pop music often features upbeat, positive themes.

We incorporated a sentiment analysis tool called VADER (Valence Aware Dictionary and sEntiment Reasoner), a pre-trained, lexicon-based sentiment analysis tool. We used the SentimentIntensityAnalyzer class from the VADER library to analyze the sentiment of each lyric and output a compound sentiment score, which is a normalized, weighted sum of the valence scores of the individual words in the lyrics. It takes into account both positively and negatively connotated words as well as their relative intensities to output a single value ranging from -1 to 1 that describes the overall sentiment of a given lyric. Adding this feature to our model resulted in a test accuracy improvement of around 7% compared to the base DistilBERT model and an improvement of 13.6% over the base RoBERTa model. Our extension results are summarized in Table 1.

| Model | Test acc | Dev acc | Weighted F1-score (Dev,Test) |
|---|---|---|---|
| LSTM Base | 23.7 | 23.6 | 0.118, 0.116 |
| RoBERTa | 24.3 | 53.2 | 0.093, 0.270 |
| BERT | 32.2 | 44.2 | 0.156, 0.217 |
| DistilBERT | 33.3 | 43.7 | 0.164, 0.212 |
| RoBERTA (tf-idf) | 23.8 | 53.3 | 0.100, 0.270 |
| DistilBERT (Sentiment) | 40.0 | 43.2 | 0.193, 0.212 |

Table 1: Model Performance

From Table 1, we see that the DistilBERT model paired with sentiment analysis performed the best in test accuracy. For the full results of this model, consult figures 7 and 8.

### 4.3 Error Analysis

The categorization of misclassifications by our final model, distilBERT with sentiment analysis, are shown in Table 2.

These misclassifications by the model display a pattern and are suggestive of a few key insights:

- **Genre Overlap Confusion**: there appears to be a significant overlap in lyrical themes between genres like Rock, Indie, Pop, Country and Folk. This overlap might be causing confusion for the model.

- **Mood and Style Confusion**: The misclassification between genres like Metal and Electronic or Jazz and Country might be due to similarities in mood or style that the lyrics convey, which the model is picking up on.

- **Ambiguity in Rock**: Rock genre sees to have the highest number of misclassifications, possibly due to its broad spectrum and overlap with several other genres.

Specific Examples:

1. **"Country" misclassified as "Folk"**: in the song titled *"Song Of The Ocean Lyrics by Charlie Landsborough"*, the lyrical content of this song share several common features with Folk music, such as a focus on nature, a narrative style, and a refelective tone. This likely led the model to classify it as Folk instead of Country. This example illustrates how the thematic overlap and stylistic similarities between genres can lead to misclassifications,

```
Validation Accuracy: 0.43222222222222223
Genre         Count    Precision    Recall    F1_score
----------  -------  -----------  --------  ----------
Folk            174     0.323077  0.362069   0.170732
Hip-Hop         176     0.648889  0.829545   0.36409
Indie           193     0.322222  0.300518   0.155496
Pop             177     0.243816  0.389831   0.15
Electronic      192     0.341463  0.291667   0.157303
Metal           173     0.596774  0.641618   0.309192
R&B             172     0.377622  0.313953   0.171429
Country         184     0.586667  0.478261   0.263473
Jazz            188     0.554455  0.595745   0.287179
Rock            171     0.291667  0.122807   0.0864198

Overall
Accuracy: 0.43222222222222223
Weighted Precision: 0.4286541363634661
Weighted Recall: 0.43222222222222223
Weighted F1 Score: 0.21154034271628167
Macro Precision: 0.43222222222222223
Macro Recall: 0.43222222222222223
Macro F1 Score: 0.21611111111111111
```

Figure 7: DistilBERT (Sentiment) Dev-Set Results

```
Test Accuracy: 0.4002520478890989
Genre         Count    Precision     Recall    F1_score
----------  -------  -----------  ---------  ----------
Folk            495     0.236383   0.438384   0.153574
Hip-Hop         960     0.786316   0.778125   0.391099
Indie           510     0.196104   0.296078   0.117969
Pop            1110     0.297368   0.305405   0.150667
Electronic      660     0.263302   0.292424   0.13855
Metal           810     0.534754   0.712346   0.305453
R&B             510     0.213277   0.296078   0.123974
Country         810     0.617778   0.514815   0.280808
Jazz            660     0.401408   0.431818   0.208029
Rock           1410     0.392857   0.0702128  0.0595668

Overall
Accuracy: 0.4002520478890989
Weighted Precision: 0.4205323614094045
Weighted Recall: 0.4002520478890989
Weighted F1 Score: 0.1927798116990907
Macro Precision: 0.4002520478890989
Macro Recall: 0.4002520478890989
Macro F1 Score: 0.2001260294454942
```

Figure 8: DistilBERT (Sentiment) Test-Set Results

| Genre Misclassification | Instances |
|---|---|
| Rock as Indie | 332 |
| Pop as R&B | 246 |
| Metal as Electronic | 231 |
| Rock as Country | 215 |
| Rock as Electronic | 206 |
| Pop as Electronic | 152 |
| Rock as Folk | 150 |
| Rock as Pop | 146 |
| Pop as Country | 145 |
| Jazz as Country | 132 |

Table 2: Categorization of the Misclassifications

especially when lyrics share elements common to multiple genres.

2. **"Metal" misclassified as "Electronic"**: in the song titled "Angel (come Walk With Me) by Conception", the lyrics and this song's thematic elements that can be found in both Metal and certain subgenres of Electronic music. The absence of more explicit Metal elements might have led the model to classify it as Electronic. This example shows the challenge of genre classification based on lyrics alone, especially when dealing with genres that can have overlapping thematic and stylistic elements.

3. **"Rock" misclassified as different genres**: Rock music has the highest number of misclassifications. A deeper dive into these examples suggested that most of these misclassification stem from a combination of the absence of distinct Rock elements and the presence of themes or styles more commonly associated with the genres they were misclassified into.

Drop in performance of RoBERTa after adding TD-IDF feature:
The inclusion of a TF-IDF (Term Frequency-Inverse Document Frequency) feature alongside RoBERTa may not always lead to improved performance, and several factors could contribute to the observed decrease in model performance. One of the reasons, as mentioned earlier, could be the redundancy introduced by the TD-IDF feature due to its overlap with the information already captured by RoBERTa. Another reson could be that the combined feature set might have increased the complexity of the model, potentially leading to overfitting. Instead of combining features directly, we could consider using ensemble methods to combine predictions from RoBERTa and a model trained solely on TF-IDF features in an attempt to mitigate issues associated with feature combination.

## 5 Conclusion

For this project, our objective was to explore the applications of transformer-based architectures in song genre classification. Notably, our most successful implementation featured DistilBERT coupled with sentiment analysis, yielding a test accuracy of 40%. While direct comparisons with prior studies are challenging due to variations in genre granularity and overall approach, our selection of 10 output genres distinguishes our work. For context, the Babac paper, which similarly explored the performance of BERT variants along with non-transformer-based models, adopted only five genres. Though our accuracy does not surpass the benchmarks set by established studies, our approach of leveraging sentiment analysis with transformer-based models contributed to a richer understanding of the intricate interplay between lyrical content and musical genres. The achieved accuracy reflects the complexities inherent in predicting genres with diverse and nuanced lyrical compositions, which we were able to explore thoroughly with this project.

## Acknowledgements

We would like to thank our TA mentor for this project, Anson Huang, for his guidance and helpful advice given during each step of our implementation. We would also like to thank Professor Yatskar for providing us with the strong natural language processing skills and foundation for us to be able to conduct this project.

## References

[1] Mingjing Zeng et al. MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training.

[2] Bo-Hsun Sung and Shih-Chieh Wei. BECMER: A Fusion Model Using BERT and CNN for Music Emotion Recognition.

[3] Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network

[4] Marijic and Bagic Babac, 2023. Predicting song genre with deep learning.

[5] Matei Bejan. Multi-Lingual Lyrics for Genre Classification (Kaggle), Jan 2021.