Classifying Music Genres Using Song Lyrics with Transformer-based Models

1. Description

Music genres are often defined by various elements such as rhythm, instrumentation, and lyrical content. Lyrical themes and language patterns can provide a valuable indication of a song's genre. This project aims to classify songs into genres using only their lyrical content. By analyzing large datasets of song lyrics, we intend to build and train machine learning models capable of accurately predicting a song's genre based on its lyrics. Classifying music based on lyrics alone could benefit recommendation systems, enhance music search and organization, and offer insights into how genres can be defined by lyrical themes. This approach could provide a deeper understanding of the relationship between lyrical content and genre, offering new ways to organize and explore music collections.

2. Approach to Solution

Dataset and Preprocessing

We will collect song lyrics from publicly available datasets like the Million Song Dataset or through APIs such as Genius or Lyrics.ovh. The dataset will be curated to include labeled songs from a diverse range of genres, such as Pop, Rock, Hip-Hop, and Country, to ensure variety in classification. Once collected, the lyrics will undergo several preprocessing steps. This will involve tokenizing the lyrics into individual words or tokens, removing stop-words that do not contribute significantly to genre distinction, and applying lemmatization or stemming to reduce words to their base form for consistency. Additionally, we will clean the text by eliminating special characters and punctuation while converting all words to lowercase for uniformity.

Modeling Approach

We will experiment with transformer-based models like BERT and RoBERTa due to their ability to capture contextual relationships within text effectively. We will also explore other text classification architectures like LSTM and CNN for comparison. As a baseline, we will implement traditional approaches by combining TF-IDF with machine learning classifiers such as SVM and Naive Bayes. For feature extraction, we aim to leverage pre-trained word embeddings and attention mechanisms to enhance the models' understanding of contextual information within the lyrics.

Evaluation Strategy

The performance of our models will be evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Additionally, confusion matrices will be employed to analyze misclassifications, offering insights into similarities or overlaps between genres based on lyrical content.

3. Assessment Methodology

Performance Metrics:

We will evaluate model performance using classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, confusion matrices will be used to analyze misclassifications, providing insights into genre similarities based on lyrical content.

Cross-Validation Strategy:

A 5-fold cross-validation strategy will be used to ensure the robustness of our models and assess the consistency of classification performance across different data splits.

Ablation Study:

We will conduct ablation studies to analyze the impact of different preprocessing methods (like stop-word removal or lemmatization) and feature sets on model performance. This will help identify the most influential factors in improving classification accuracy.

4. Timeline

Week 9: Data collection and preprocessing.

Week 10: Implement and train models (BERT, LSTM, CNN).

Week 11: Evaluate models, perform cross-validation, and compare with baseline approaches.

Week 12: Conduct ablation studies, finalize results, and prepare visualizations.

Week 13: Write the final report and prepare the project presentation.

5. Responsibilities

I am solely responsible for this project. My responsibilities include collecting and preprocessing the data, which involves cleaning, tokenization, and feature extraction. I will also be implementing and training transformer-based models like BERT and RoBERTa. Additionally, I will focus on developing baseline models, conducting cross-validation, and evaluating model performance using various classification metrics. I will handle ablation studies independently and be responsible for preparing the final report and presentation.