NLP Project Proposal

Hein Huijskes (s2386836)
Tibet Tugay (s2489384)
Group 42

September 2025

Project Description

This project will use different NLP methods for classifying the music genre of songs based on their song lyrics.

If time permits, we would prefer building a model that can generate song lyrics based on the datasets, however this likely seems out of scope for now. Therefore the focus of the project will be on classifying songs.

Therefore our research question is as follows: ”Which combination of models and Natural Language Processing techniques yields the highest accuracy for classifying music genres based on song lyrics?”

Relevant Literature

We will attempt to implement some of the methodology used by Fell and Sporleder [1]. They propose a multitude of sophisticated features that can be extracted from song lyrics, combined with an n-gram model, that make for an accurate model. We also consider the 2024 paper by Green et al., surveying 560 publications in music genre recognition [2]. They provide insight into methodologies and datasets that have been used over the years and their success.

Potential Methods

There are many potential methods to choose from based on the provided literature and our classes. We would like to pursue the effects of the following methods:

- n-gram model
- Sophisticated features [1]
  - Vocabulary
  - Style

  - Semantics
  - Orientation
  - Song structure

- Data preparation
  - Stop word removal
  - Rare word removal
  - Balancing the dataset
  - Normalization
  - Lemmatization or Stemming
  - Vectorization

- Different models
  - Linear regression
  - Logistic regression
  - SVM
  - Naive Bayes
  - Decision Tree

Dataset

We will be using the following datasets:

- song_lyrics dataset: this is a dataset that consists of the tag (genre), the lyrics, and the title of a song. The dataset has 4.39 million entries and that include 6 genres, which are pop, rap, rock, misc, rb, and country with 42.1, 27.9, 18.4, 4.6, 4.4, and 2.5 percent, respectively. We plan to remove the misc entries and train an instance of the model with balanced genre percentages to account for the huge discrepancy between the number of entries in each genre, reducing the dataset to approximately 300 to 400 thousand entries.

- GTZAN dataset: This is a dataset with 1000 entries and 10 genres. The GTZAN dataset is a rather famous dataset in music classification. We also intend to train our model with this dataset.

Evaluation Metrics

The main evaluation metrics used will be the confusion matrix, accuracy and F1 score since some variation of our data will have high imbalance.

References

[1] Michael Fell and Caroline Sporleder. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 620–631, 2014.

[2] Owen Green, Bob Sturm, Georgina Born, and Melanie Wald-Fuhrmann. A critical survey of research in music genre recognition. *International Society for Music Information Retrieval*, 2024.