Natural Language Processing

UNIVERSITY
OF TWENTE.

Project Proposal

Academic Year 2025–2026

Onat Akca s3521281

Soraya Charkas s3696340

September 2025

1 Project Description

Music is often described as a universal language, but genres reflect cultural and individual differences in taste. In this project, we aim to explore how natural language processing can uncover the patterns in music lyrics to determine the genre. We aim to answer the following questions :

- can we accurately classify language or genre of music based on its lyrics using NLP tools?
- what are the most important features that help us to differ between on genre or another?

2 Relevant Literature

For the literature review, we will use Lyrics-based Analysis and Classification by Fell and Sporleder (2014) [link], as it was one of the first works to propose and evaluate large-scale lyrics datasets for genre classification. We will also refer to the more recent Genius Lyrics with Language Information dataset article [link], since it provides a multilingual collection of over one million lyrics, making it especially suitable for exploring cross-linguistic comparisons and modern NLP methods.

3 Potential methods or models

Before starting with classification methods, we will preprocess large amounts of text. Then we will begin with baseline models using TF-IDF features and classical classifiers (Logistic Regression, SVM). Next, we plan to experiment with pretrained embeddings (e.g., Word2Vec, multilingual BERT) for improved representation of lyrics.

4 Datasets

We will use the Genius Song Lyrics with Language Information dataset [link], which contains over one million song lyrics in multiple languages, including genre labels and other metadata useful for classification tasks.

5 Evaluation Method

To evaluate our model, we will use supervised classification with common methods such as SVM and Logistic Regression. Evaluation will be done on a separate test set, and we will report metrics including precision, recall, F1-score, and accuracy.