DS 4002-001 9/12/25

Project 1 - M1: Classifying Music Genre by Song Lyrics

Data Devils: Michael Hijduk (leader), Laura Espuna, Anthony Jiang

Goal Statement: We seek to develop a machine learning model that classifies song genres based solely on lyrics with an accuracy of 80% or higher.

Context: It is known that machine learning can be used to classify music by genre as well as offer insight into the sentiments and mood of the piece. Audio features have historically been most important in this pursuit but lyrics provide another valuable source of information. Cultural references, emotions, and linguistic patterns often correspond with genre conventions. This project seeks to explore that frontier, building on research that has shown measurable differences in linguistic style across genres [1], [2] and constructing a classification model that uses solely lyrics.

This research is important as it offers insight into a different approach towards music classification. Genre influences how listeners discover new music, how artists position themselves in the industry, and how streaming platforms design recommendation systems [3]. The business of algorithmic recommendations is critical to companies like Spotify, Apple Music, and Pandora, which rely on such systems to maintain their bottom line. Furthermore, many users stand to benefit from improved recommendation relevance or searchability.

Research Question: How well can song lyrics alone be used to classify song genre?

Modeling Approach: We plan to aggregate song lyrics using the LyricsGenius API, which collects data from the popular Genius platform. We will apply TF-IDF encoding and, following the approach of Fell and Sporleder [2], will also capture bigrams and potentially trigrams. N-grams allow us to capture not only word frequency and uniqueness from TF-IDF but also word order, offering more analytical power. Our problem is one of classification and we will employ a logistic regression model. Logistic regression has various beneficial features that make it a strong starting point for our modeling process. The interpretability of feature weights allows us to see which words are pulling the model towards certain genres. Additionally, logistic regression is robust to correlated features like “baby” and “love,” and its linear nature makes interpreting the decision boundaries simple.

[1] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *Proc. 9th Int. Conf. Music Information Retrieval (ISMIR)*, Philadelphia, USA, 2008, pp. 337–342. [Online]. Available: https://www.researchgate.net/publication/220723682_Rhyme_and_Style_Features_for_Musical_Genre_Classification_by_Song_Lyrics. [Accessed Sept. 11, 2025].

[2] M. Fell, C. Sporleder, “Lyrics-based Analysis and Classification of Music,” in Proc. 25th Int. Conf. Computational Linguistics, Aug. 2014, pp. 620-631. [Online]. Available: https://aclanthology.org/C14-1059.pdf. [Accessed Sept. 11, 2025].

[3] M. Schedl, H. Zamani, C. Chen, Y. Deldjoo, M. Elahi, “Current challenges and visions in music recommender systems research,” *International Journal of Multimedia Information Retrieval,* vol. 7, pp. 95-116, [Online]. Available: https://link.springer.com/article/10.1007/s13735-018-0154-2. [Accessed Sept. 11, 2025].