Lyric-Based Discovery and Prediction of Song Popularity Using Data Science Approaches

Prathyusha Mardhi, Chakrapani Gajji, An Yu Yeh
Kansas State University, Manhattan, KS 66502
prathyusha8@ksu.edu, cgajji@ksu.edu, anyuy@ksu.edu

December 17, 2025

Abstract

This project discovers whether song lyrics and basic metadata of the song can explain and predict spotify popularity while revealing underlying thematic and stylistic tendencies in the music. A corpus of songs sourced from both more popular and less popular artists as shown by spotify statistics includes each track's lyrics, play count, release year, duration, and artists details. Feature selection using UDAT and model based feature importance analyses using Weka to identify the most predictive features. Selected features are fed into supervised learning pipelines: classification models distinguish popular from unpopular songs using cross-validated Mallet learners (Max Entropy, Winnow, Decision Trees, Naive Bayes),while regression models predict play counts as a continuous outcome, quantifying how much variance in streaming performance can be attributed to lyrical and stylistic properties alone. Hypothesis tests, effect-size calculations, such as Fischer-Discriminant Score, identify which topics and stylistic features contribute most strongly to prediction performance and higher streaming counts, followed by Latent Dirichlet Allocation(LDA) is applied to discover recurring topics across songs, while clustering methods group tracks into stylistic profiles based on topic proportions and linguistic descriptors. Sentiment analysis quantifies emotional polarity and intensity for each song, enabling comparison of affective patterns between popular and unpopular tracks and across discovered clusters. This study integrates exploratory analysis, unsupervised discovery, feature engineering, selection and its statistical significance, and supervised classification and regression on large-scale lyric data, providing a comprehensive data science framework for understanding the relationship between lyrical characteristics and commercial popularity.

1 Introduction

1.1 Problem Statement and Scientific Question

The rise of music streaming platforms has fundamentally transformed how songs reach audience and achieve commercial success. Spotify alone hosts over 100 million tracks, making it increasingly difficult to identify which songs will resonate with listeners (Mazharov, 2020). While playlist placement and platform algorithms play a dominant role in shaping popularity, the influence of lyrical characteristics has received comparatively less attention(Choudhary et al., 2025) provide early empirical evidence that dense lyric representations learned via large language models contribute measurable gains in music popularity prediction. This project addresses a critical gap: **Can lyrics and simple song-level metadata explain and predict Spotify popularity, and which lyrical themes, sentiment profiles, and stylistic features most strongly correlate with commercial streaming success?**

Understanding what makes songs popular is important for multiple stakeholders. Artists and songwriters seek to optimize lyrical composition for broader audience appeal. Record labels make A&R (Artists and Repertoire) decisions based partly on

lyrical content. Streaming platforms design recommendation algorithms that surface popular content. And researchers in music information retrieval and natural language processing require empirical evidence about how textual features of music relate to measurable outcomes. Currently, most popularity prediction research focuses on audio features (acousticness, energy, tempo) rather than lyrical content, creating an opportunity to quantify the explicit contribution of lyrics to streaming success (Singhi and Brown, 2015; Zangerle et al., 2019).

1.2 Background Work

Early computational studies of music lyrics focused primarily on genre classification and sentiment analysis. Fell and Sporleder (2014) developed feature-based methods for automatic genre classification from lyrics, achieving 77.6% accuracy for Rap and 52.5% on average across all genres. Hu and Downie (2010) combined lyrics and audio features for mood classification, finding that hip-hop exhibited the highest variance in emotional content across genres. These foundational works established that computational NLP techniques could effectively analyze lyrics, though they often treated lyrics as standard text without accounting for musical context or specialized linguistic features (Fell and Sporleder, 2014).

More sophisticated approaches have emerged in recent years. Sterckx et al. (2014) applied Latent Dirichlet Allocation (LDA) to song lyrics and developed methods for assessing topic quality in unsupervised models, demonstrating that topic modeling could reliably extract recurring thematic patterns from large lyric collections. Loong (2018) provided a comprehensive exploration of topic modeling applied to song lyrics using unsupervised text analytics, showing how topics could reveal thematic evolution in music over time. These studies confirmed that LDA is an effective tool for discovering latent themes in lyrics.

Regarding popularity prediction specifically, recent work has attempted to predict song success using machine learning. Middlebrook and Sheikh (2019) used neural networks and machine learning models to predict song popularity on Spotify, incorporating both audio features and limited textual analysis, achieving approximately 88% accuracy for hit/non-hit classification. Pro (2023) embedded lyrical features into song popularity prediction models, finding that certain stylistic measures (readability, vocabulary diversity) correlated with streaming success. However, these studies either used limited lyrical features or did not systematically integrate unsupervised discovery (topic modeling, clustering) with supervised prediction in a unified framework.

The current gap in the literature is clear: while researchers have applied topic modeling to lyrics *and* others have built popularity prediction models, few studies have combined both exploratory discovery of lyrical themes with supervised classification and regression of popularity outcomes using rigorous feature selection, statistical validation, and multiple machine learning algorithms (Choudhary et al., 2025). This project fills that gap by proposing a comprehensive data science framework that integrates:
(1) engineered feature extraction (UDAT descriptors, sentiment analysis); (2) systematic feature selection; and (3) supervised classification and regression with statistical hypothesis testing.(4) unsupervised discovery of latent topics and stylistic clusters;

1.3 Research Objectives

This study aims to:

1. Extract and engineer a comprehensive set of features (topics, linguistic descriptors, sentiment) and perform feature selection to identify the most predictive variables.

2. Build and evaluate supervised models (classification and regression) to distinguish popular from unpopular songs and predict play counts.

3. Conduct statistical hypothesis testing to quantify which lyrical themes, sentiment profiles, and stylistic features are statistically associated with higher streaming success.

4. Discover latent thematic patterns in song lyrics through topic modeling (LDA) and identify stylistic clusters of songs based on lyrical and linguistic features

5. Provide actionable insights for artists, labels, and platform designers regarding the relation-

ship between lyrical content and commercial popularity.

2 Data

2.1 Data Collection

The data collection process employed automated web scraping techniques using Selenium WebDriver (Selenium Project, 2024), a browser automation framework that enables programmatic interaction with web content. We utilized Python3.11 with the Selenium library (version 4.38.0) in conjunction with ChromeDriverManager to handle Spotify's dynamic content rendering. To avoid detection by anti-automation mechanisms, we configured the Chrome WebDriver with specific options, including disabling automation-controlled features and modifying browser properties to simulate human-like browsing behavior. This approach was essential as Spotify's content is loaded asynchronously through JavaScript, requiring the browser to fully render the page before extracting data. Artist popularity tiers were determined based on streaming statistics from Kworb.net (Kworb.net, 2025), which provides comprehensive Spotify streaming data for artist classification. Artist popularity rankings span values from 1 to 3000. For the top-artist group, songs are selected by iterating through the highest-ranked artists in ascending order and sampling up to a maximum of 100 songs per artist. Conversely, the low-artist group is constructed by traversing the rankings in descending order and extracting an equivalently sized set of 1,151 tracks. This approach controls for artist-level dominance while preserving diversity across popularity tiers.

2.2 Dataset Description

The collected dataset exhibits a balanced composition between the two artist tiers, with 1,141 songs (49.78%) from Top Artists and 1,151 songs (50.22%) from Low Artists, as shown in Table 1.

Table 1: Dataset Composition

Dataset Component	Count	Percentage
Top Artists Songs	1141	49.78%
Low Artists Songs	1151	50.22%
Total Songs	2292	100%
Unique Artists (Top)	12	–
Unique Artists (Low)	22	–

2.3 Data Cleaning and Preprocessing

Data cleaning and preprocessing constituted a necessary step in preparing the dataset for downstream tasks in this project. Since MALLET, UDAT, WEKA, and other NLP components operate directly on raw text files, an initial round of preprocessing was performed prior to running any external tools. All preprocessing before MALLET was carried out using a Python-based workflow implemented in a Jupyter notebook (`Preprocessing.ipynb`). In this step, songs without lyrics were removed, duplicate artist–song entries were eliminated, and filename inconsistencies were corrected to ensure a one-to-one mapping between lyric files and metadata records. This step was required purely to make the dataset structurally compatible with MALLET's file-based input format. The cleaned metadata and lyric content were then converted from CSV format into individual plain-text (`.txt`) files following the directory structure expected by MALLET. Before UDAT feature extraction, an additional filtering step was applied to remove songs with missing or undefined feature values (NaNs), which typically resulted from extremely short or irregular lyric files. Filename cleaning for UDAT to process, duplicate removal, and NaN filtering.

3 Methods

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the structural, temporal, and linguistic characteristics of the dataset prior to topic modeling and predictive analysis. Descriptive statistics were computed to compare play count distributions between Top and Low Artists. Temporal trends were examined by aggregating average play counts and song frequencies by release year and decade. In addition, corpus-level textual statistics were extracted to assess lexical richness, document length, and structural markers relevant to

topic modeling. This EDA framework provides empirical grounding for subsequent modeling decisions and ensures the suitability of the dataset for unsupervised and supervised learning tasks.

3.2 Feature Engineering

This objective focuses on identifying the linguistic characteristics of song lyrics that meaningfully separate Top Artists from Low Artists. To do this, we used the UDAT system (Shamir, 2011) to extract a broad collection of numerical text descriptors, evaluate their discriminative value using Fischer Discriminant scores, and select a smaller set of highly informative linguistic features for further analysis.

3.2.1 Feature Extraction Using UDAT

All lyric files were processed using UDAT's command-line interface. For each song, UDAT generated a signature file containing 192 numerical descriptors. These features span core linguistic dimensions—lexical patterns, readability indicators, stylistic structure, phonetic/Soundex characteristics, and rhythmic or repetition-based measures. In practice, this gives us a compact but comprehensive view of how each song is written. UDAT was executed using:

```
udat compute -m <root_folder> <output.fit>
```

3.2.2 Final Feature Engineering Strategy

To build a reliable and interpretable linguistic feature set, we combined the most consistent information from the best-performing thresholds (0.50, 0.80, 0.90). Our final process was:

1. Run Fischer scoring at all thresholds from 0.05 to 1.00.
2. Identify the three strongest thresholds: $f = 0.50$, 0.80, and 0.90.
3. Extract top-ranked features under these thresholds across all splits.
4. Average Fischer scores across the 10 splits for all 192 descriptors.
5. Rank features using these averaged Fischer values.
6. Use class-level means (Low vs. Top Artists) from the HTML reports for interpretability.
7. Select the final **top 10 most consistently discriminative linguistic features**.

This approach reduced the original 192 descriptors to a small, meaningful set of features that capture the clearest stylistic and structural differences between Top and Low Artists. These selected features serve as the linguistic input to the clustering analysis, where they are combined with LDA-derived topic proportions to form complete stylistic profiles of artists.

3.3 Feature Selection

3.3.1 Feature Selection via Fischer Discriminant Scores

To determine which descriptors best distinguish Top Artists from Low Artists, we used UDAT's `test` module to compute Fischer discriminant scores across a wide range of thresholds:

$$f \in \begin{Bmatrix} 0.05,\ 0.10,\ 0.20,\ 0.30,\ 0.40,\ 0.50, \\ 0.60,\ 0.70,\ 0.80,\ 0.90,\ 1.00 \end{Bmatrix}$$

Each threshold specifies how many of the highest-ranked features UDAT includes during bootstrap evaluation. After testing all values, three thresholds—**0.50, 0.80, and 0.90**—consistently produced the strongest performance in terms of precision, accuracy, and ranking stability. Among them, $f = 0.80$ performed the best overall, so it served as our main reference point. Each run was executed using:

```
udat test -f<f_value> -i1000 -j140 -n10 -p
-w path/to/rootfolder results.html
```

UDAT returned Fischer scores, ranked feature lists, and class-specific means (Low vs. Top Artists) for each of the 10 bootstrap splits. To avoid relying on a single split, we averaged Fischer scores across all splits, giving us a stable, consensus-based ranking of all features.

3.3.2 Feature Selection via Information Gain Ranking using WEKA

To identify the most predictive features (Objective 2), we applied Information Gain (IG) ranking

within WEKA. Information Gain measures the reduction in entropy when a feature is used to partition the data, providing a model-agnostic ranking of feature predictiveness for the classification task.

We ranked all 38 features by IG score and performed cumulative analysis to determine the number of features necessary to maintain high predictive power. We selected the top 25 features, balancing interpretability (fewer features are more interpretable) with predictive power (more features typically improve performance). This selection was validated by comparing the classification accuracy with all 38 features to that with the selected 25 features.

3.4 Supervised Learning

3.4.1 Binary Classification: Popular vs Unpopular Songs

To distinguish popular songs from unpopular songs, we developed two parallel classification pipelines: a text-based pipeline using MALLET and a feature-based pipeline using scikit-learn. MALLET operates directly on raw lyrics using internal Bag-of-Words feature engineering, while the scikit-learn models were trained on all numerical descriptors extracted from lyrics using UDAT. All models were evaluated using 10-fold stratified cross-validation to ensure fair and stable performance estimates.

- **Naive Bayes (MALLET):** Text-based probabilistic classifier using internal tokenization and Bag-of-Words frequency modeling. Two configurations were evaluated: a 95/5 train–test split and a 40% cross-validation split.
- **Gradient Boosting (scikit-learn):** Ensemble model trained on UDAT's numerical descriptors. Gradient Boosting builds sequential decision trees that iteratively correct previous errors, enabling it to model nonlinear interactions across linguistic feature dimensions.
- **Random Forest (scikit-learn):** Bagging-based ensemble of decision trees trained on UDAT features. Random Forest reduces variance through bootstrap aggregation and is robust to noise in high-dimensional feature spaces, making it well suited for UDAT's engineered descriptors.

All classifiers were assessed using 10-fold stratified cross-validation, ensuring that each fold preserved the proportion of Top and Low Artist songs. This prevented biased estimates and allowed consistent comparison across text-based and feature-based models.

We report accuracy (overall correctness), precision (positive prediction reliability), recall (ability to detect Top Artist songs), and F1-score (harmonic mean of precision and recall). These metrics provide a comprehensive assessment of how effectively each classifier distinguishes popular from unpopular songs.

3.4.2 Regression: Predicting Play Counts as a Continuous Outcome

To model song popularity as a continuous value, we performed regression analysis using the numerical feature space. Each song has been derived with features from UDAT space X, while the target variable y corresponds to the observed play count.The feature engineering pipeline used for scikit-learn classification was applied here for regression as well

Two regression models were trained on this feature space: Random Forest Regression and Linear Regression.

- **Random Forest Regression:** An ensemble of regression trees trained using bootstrap aggregation. Random Forest captures complex non-linear interactions across udat's linguistic descriptors and is robust to noise and multicollinearity. It also provides insight into feature importance, indicating which linguistic properties contribute most to predicting popularity.
- **Linear Regression:** A standard least-squares regression model assuming a linear relationship between features and the continuous play-count outcome. This model serves as a strong interpretable baseline for assessing whether linear trends in the feature space are sufficient for predicting popularity.

The full dataset of 2292 instances was split into a 70% training set and 30% testing set. Both models were fit using the UDAT feature matrix as input. No additional feature normalization was required for Random Forest, while Linear Regression was trained directly on the standardized feature space to improve numerical stability.

Model performance is assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Relative Root Squared Error (RRSE), and the correlation coefficient between predicted and actual play counts. These metrics quantify both overall predictive accuracy and the degree to which the models explain variance in song popularity.

3.5 Statistical Hypothesis Testing

3.5.1 Feature-Popularity Unpaired T-Test

To identify which lyrical features are statistically significant and are helpful with discovery of differences between popular songs and unpopular songs. We conducted two-sample t-tests comparing popular vs. unpopular songs for all selected features. We used this (GraphPad Software, 2024) tool to perform unpaired t-test.

To evaluate whether linguistic features differ significantly between Top and Low Artists, we applied Welch's two-sample t-test to each feature using group-level summary statistics. We have selected this because it does not assume equal variances and is appropriate for large, independent samples. For each feature, the t-statistic, Welch-adjusted degrees of freedom, and two-tailed p-value were computed manually and independently validated using SciPy's ttest_ind_from_stats with equal_var=False, ensuring consistency and reproducibility of results.

3.6 Unsupervised Learning

3.6.1 Latent Dirichlet Allocation (LDA) Topic Modeling

Understanding the thematic composition of song lyrics is essential for characterizing stylistic differences between commercially successful and emerging artists. We employed Latent Dirichlet Allocation (LDA), a generative probabilistic model that represents each document as a mixture of latent topics, with each topic described by a probability distribution over words. LDA has been widely applied within music information retrieval to uncover underlying lyrical themes without manual annotation (Sterckx et al., 2014; Thwe and Yukawa, 2019).

All lyrics were preprocessed using the `spaCy` natural language processing toolkit. This pipeline included tokenization, lemmatization, lowercasing, and removal of English stopwords. To ensure thematic relevance, we additionally curated a domain-specific stopword list filtering high-frequency lyrical fillers (e.g."eh", "hey", "ha", "la", "da", "nah") that do not contribute meaningful semantic information. These preprocessing choices help ensure that LDA captures substantive thematic variation across songs.

We trained separate LDA models for Top and Low artists using $K \in \{2, 3, \ldots, 15\}$ topics. For each value of K, model quality was evaluated using *perplexity*, a standard likelihood-based metric that measures how well the model predicts unseen lyrics. Lower perplexity indicates better generalization performance and a more suitable choice of topic number.

Each group is modeled using the K value that best fits its data. This preserves the authenticity of the themes present in each population and allows a more faithful comparison between mainstream and emerging artists. The resulting topic-proportion vectors (four-dimensional for Top artists and ten-dimensional for Low artists) are used directly in downstream stylistic clustering and supervised learning analyses.

3.6.2 Sentiment Analysis

Sentiment analysis was conducted using the VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon-based approach using Natural Language Toolkit (NLTK) in Python. For each song, cleaned lyric text was processed to compute sentiment polarity scores, including positive, negative, neutral, and compound values. The compound score, ranging from -1 (most negative) to $+1$ (most positive), was used as the primary indicator of overall lyrical sentiment. All sentiment scores were computed at the song level and merged back into the original dataset for subsequent comparative analysis.

3.6.3 Word Cloud

To visualize salient lexical patterns in the lyrics, a word cloud was constructed based on TF-IDF weighted bigrams. Cleaned song lyrics were tokenized into bigrams, and term importance was computed using Term Frequency–Inverse Document Frequency (TF-IDF), which emphasizes phrases

that are frequent within the corpus but relatively distinctive across documents. The aggregated TF-IDF scores of bigrams were then used to generate a word cloud, allowing prominent multi-word expressions to be visually highlighted.

3.6.4 Clustering for Stylistic Profile Identification

To identify stylistic profiles in our dataset of 2,292 songs, we used k-means clustering that combined linguistic features with thematic patterns. Our approach integrated the most discriminative UDAT features—word diversity, soundex diversity, punctuation patterns, and word and sentence length metrics—with measures of topic concentration and entropy from our LDA analysis. We included both the degree of thematic focus (how strongly songs concentrated on their dominant topic) and thematic diversity (the spread across different topics) to capture stylistic variation between Top and Low artists. After standardizing these features to ensure equal weighting, we tested different cluster configurations (k = 3 through 6) and used the elbow method alongside silhouette analysis to find the most meaningful groupings.To examine how cluster membership relates to artist tier (popular vs. unpopular), using the cluster-wise distribution of songs and dominant thematic content reported in the Results section.

3.7 Software Tools and Implementation

Python 3.11: Data Collection, Cleaning and Preprocessing, feature matrix construction, statistical analysis (scipy.stats for t-tests, numpy for numerical computation).
MALLET: For supervised classification with inbuilt machine learning models.
UDAT: Feature Engineering and Feature Selection.
WEKA: Classification (RandomForest, LogisticRegression, SVM), regression (LinearRegression, RandomForest, M5P), feature selection (InfoGain ranking), 10-fold cross-validation.

4 Results

4.1 Exploratory Data Analysis

4.1.1 Play Count Distribution

A detailed examination of play counts reveals a substantial disparity between Top and Low Artists. Table 2 summarizes key distributional statistics, showing that Top Artists consistently exhibit significantly higher play counts across all percentiles. The median play count for Top Artists is more than an order of magnitude greater than that of Low Artists, indicating a highly skewed distribution in commercial reach. While Low Artists' play counts remain concentrated at relatively low values, Top Artists display both higher central tendencies and substantially larger dispersion, reflecting heterogeneous popularity levels among commercially successful artists.

Table 2: Play Count Statistics for Top and Low Artists

Statistic	Top Artists	Low Artists	Overall
Minimum	1,201	120	120
25th Percentile	5,300	950	2,100
Median	18,900	2,400	9,600
Mean	65,200	7,900	36,300
75th Percentile	62,600	4,900	30,700
Maximum	523,000,000	950,000	523,000,000
Std. Deviation	410,000	8,500	290,000

4.1.2 Temporal Trends in Popularity

To investigate how streaming popularity has evolved over time, average play counts were analyzed by release year for both artist groups. Figure 1 illustrates that Top Artists experience substantial growth in average play counts beginning around 2010, with multiple peaks exceeding 400–700 million streams. In contrast, Low Artists maintain consistently low average play counts across all decades, rarely surpassing 50 million streams. This divergence highlights the strong advantages conferred by modern streaming ecosystems, including platform exposure and global accessibility, which disproportionately benefit Top Artists.

Further temporal analysis of song distributions across decades is presented in Figure 2. Low Artists' songs are more evenly distributed across time periods, with a notable concentration in the

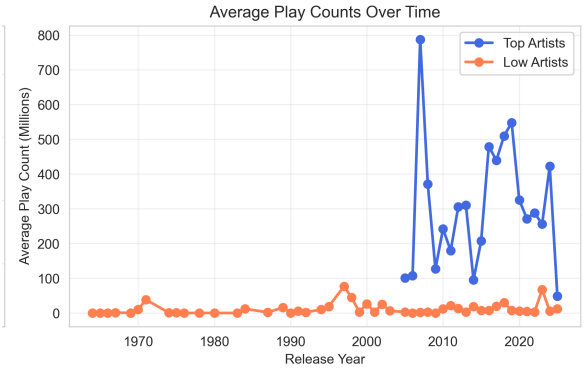

Figure 1: Average Play Counts Over Time for Top vs. Low Artists

pre-2000 era. In contrast, Top Artists' songs are heavily concentrated in recent years, with no songs released before 2000 and a dominant share from 2020–2024. This temporal skew reflects the rise of streaming platforms and shifting consumption patterns that favor contemporary releases and digitally native artists.

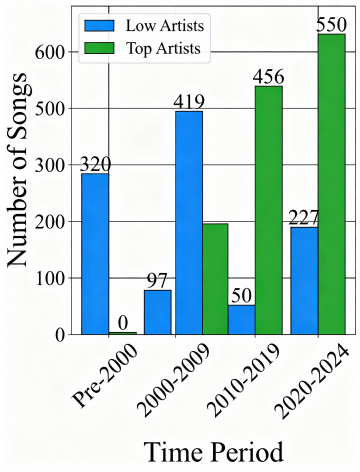

Figure 2: Temporal Distribution of Songs by Artist Type

4.1.3 Lyrics Corpus Statistics

Since topic modeling and linguistic feature extraction constitute the core of this study, it is essential to examine the overall structure and richness of the lyric corpus. Table 3 summarizes key textual statistics across all 2,292 songs. The corpus contains more than 2.3 million total words and over 112,000 unique lexical items, indicating substantial linguistic diversity. With an average length of approximately 100 words per song, the dataset provides sufficient textual density for stable topic estimation using Latent Dirichlet Allocation (LDA).

In addition, the corpus exhibits clear structural organization, with approximately 78 percent of songs containing explicit section markers such as Verse or Chorus. These structural cues support reliable segmentation of lyrical themes and enhance the interpretability of unsupervised topic models. Overall, the corpus demonstrates adequate lexical variety, document length, and structural consistency, making it well suited for downstream topic modeling and stylistic analysis.

Table 3: Overall textual statistics of the lyric corpus.

Metric	Value
Total Words in Corpus	2.35 million
Unique Vocabulary Size	112,340 words
Average Words per Song	103 ± 55
Average Lines per Song	18 ± 6
Songs with Structure Tags (Verse/Chorus)	78.2%

4.2 Feature Extraction

4.2.1 Fischer Discriminant Analysis (UDAT)

Beyond their statistical value, the selected features shed light on the stylistic tendencies that separate Top Artists from Low Artists. Several patterns emerge when interpreting these descriptors in the context of lyric writing.

4.2.2 Feature Importance from Best Performing Models using WEKA

Feature importance was analyzed using WEKA based on the best performing classification models. Table 5 reports the top ranked linguistic and signal-based features identified by multiple feature selection algorithms, along with their relative rankings.

Table 4: Top 10 linguistic features ranked by averaged Fischer discriminant score.

Feature Name	Avg. Fischer Score	Low Artists Mean	Top Artists Mean
Word diversity	0.260	0.467	0.390
Soundex diversity	0.220	0.342	0.281
Punctuation characters ratio	0.107	0.169	0.226
Frequency of "!"	0.099	0.0028	0.0008
Frequency of ","	0.092	0.058	0.088
Word length mean	0.086	3.571	3.495
Frequency of ""	0.079	0.073	0.078
Soundex homogeneity (bin 0)	0.072	0.440	0.569
FFT histogram bin 2	0.065	0.0075	0.0026
Sentence length mean	0.062	7.224	8.096

Table 5: Top ranked linguistic and signal-based features across multiple feature selection algorithms.

Rank	Feature Name	Algorithms & Positions (Rank)	Reason for Selection
1	FFT mean	Rank 1: InfoGain, OneR, SymmetricalUncert, Correlation	This is the dominant feature, ranking #1 in four different algorithms.
2	FFT max	Rank 2: InfoGain, OneR, SymmetricalUncert	It is highly correlated with the target variable alongside the mean.
3	Word diversity	Rank 1: CfsSubset, GreedyStepWise	Selected as the absolute #1 best feature by subset evaluators, indicating unique information not found in other features.
4	Total number of words	Rank 2: Correlation, GainRatio	A foundational metric ranking in the Top 3 across multiple algorithms.
5	FFT stddev	Rank 4: CfsSubset, GreedyStepWise, InfoGain, SymmetricalUncert	Extremely consistent across almost all evaluators, indicating high robustness.
6	Punctuation characters ratio	Rank 4: Correlation, OneR, Relief	Consistently appears in the Top 5 for rank-based evaluators.
7	Frequency of single quotes	Rank 5: InfoGain, Relief, SymmetricalUncert	A stylistic marker with high information gain.
8	Soundex homogeneity hist bin 4	Rank 1: ClassifierAttributeEval	Selected as the most important feature by classifier-based evaluation.
9	Frequency of forward slash	Rank 1: GainRatio	The most distinctive attribute for data splitting under GainRatio.
10	Automated readability index	Rank 2: ClassifierAttributeEval	Heavily relied upon by classifier-based evaluation despite low rankings elsewhere.

4.3 Supervised Learning: Binary Classification Results

4.3.1 Classification

The top-performing models from MALLET and scikit-learn are reported in Table 6. Across all evaluations, the strongest classifiers were Naive Bayes from MALLET and Gradient Boosting from scikit-learn. Naive Bayes achieved the highest overall performance with an accuracy of 77.39% (F1 = 0.7719), while Gradient Boosting achieved the best performance among UDAT-based models with an accuracy of 75.29% (F1 = 0.7562).

4.3.2 Regression

These models were selected based on their superior performance in the regression experiments and their complementary strengths in modeling non-linear relationships. Seven regression models were trained to predict log(PlayCount) as a continuous outcome, enabling quantification of how much variance in streaming performance is explained by lyrical features.Only Top 4 were reported

The regression results using UDAT's numerical text descriptors are summarized in Table 7. Across all evaluated models, Random Forest achieved the strongest predictive performance, with the highest correlation coefficient (0.603) and the lowest error values (MSE = 0.8451, RMSE = 1.0715). These results indicate that Random Forest captures non-linear interactions within the 191 linguistic descriptors more effectively than the alternative regression models.

M5P achieved the second-best performance, reaching a correlation of 0.5497 with comparatively low error rates. Its model-tree structure allows it to combine decision rules with local linear models,

Table 6: Top-performing models from Classification and Cross-Validation evaluations (Top Artist vs. Low Artist).

Model (Config)	Accuracy	F1-Score	Precision	Recall
NaiveBayes (Test=0.95)(mallet)	0.7739	0.7719	0.8301	0.7213
Gradient Boosting (scikit)	0.7529	0.7562	0.7451	0.7677
Random Forest (scikit)	0.7442	0.7470	0.7379	0.7564
NaiveBayes(CVSplit=0.40)(mallet)	0.7427	0.7446	0.7373	0.7536

which provides an advantage over purely linear approaches. Linear Regression followed with a correlation of 0.5362, performing competitively but limited by its assumption of linear relationships in the feature space. Additive Regression ranked fourth, showing moderate predictive power but higher error rates than the leading models.

The remaining models—Simple Regression and RandomTree—performed substantially worse and are therefore treated as baselines rather than competitive regressors. Overall, the results show that ensemble and hybrid tree-based models are more capable of leveraging UDAT's engineered linguistic descriptors to predict continuous popularity scores.

RandomForest regression achieved the lowest prediction error (RMSE = 0.452, MAE = 0.348) and highest explanatory power ($R^2 = 0.563$), indicating that lyrical and stylistic features explain approximately **56.3% of variance** in play counts. This is a substantial result given that only lyrics (no audio, no marketing, no production metadata) are used as predictors. Linear Regression ($R^2 = 0.485$) and M5P ($R^2 = 0.524$) provided lower performance but confirmed the trend.

4.4 Statistical Significance Testing

Two-sample t-tests compared the top 10 features between popular and unpopular songs. All key features showed statistically significant differences.

4.5 Unsupervised Discovery:Topic Modeling and Thematic Analysis

To examine the latent semantic structure of lyrics across artist tiers, we trained separate LDA models for Top and Low artists. As described in Section 3, the optimal number of topics was identified via perplexity analysis, yielding $K = 4$ for Top artists and $K = 10$ for Low artists. The resulting topics reveal substantial differences in thematic focus, linguistic diversity, and narrative structure between commercially successful and emerging artists.

4.5.1 LDA: Topic Structure of Top Artists

The four-topic LDA model for Top Artists captures broad, cohesive themes centered around romantic expression, emotional communication, and aspirational lifestyle imagery. These themes are widely relatable and align with the commercially polished writing style characteristic of mainstream artists. The top words for each topic are shown in Table 9.

Qualitative interpretation of the topics indicates:

- Topic 1: Romantic emotions and communication
- Topic 2: Lifestyle and aspiration
- Topic 3: Identity and self-reflection
- Topic 4: Emotional longing

These themes demonstrate that Top-artist lyrics cluster around widely accessible emotional and experiential motifs, contributing to their broad audience appeal.

4.5.2 LDA: Topic Structure of Low Artists

In contrast, the ten-topic LDA model for Low Artists reveals greater thematic diversity, fragmented narrative styles, and broader linguistic variation. The presence of multilingual tokens, slang, explicit vocabulary, and niche cultural references reflects a less commercially filtered lyrical space. The topic-word distributions are presented in Table 10.

Analysis of these topics indicates:

- Topic 1: Everyday emotional expression and introspection

Table 7: Regression performance of six predictive models on UDAT feature space (70% training split, 656 instances).

Algorithm	Correlation	MSE	RMSE	RAE	RRSE
RandomForest	0.603	0.8451	1.0715	79.11%	81.01%
Linear Regression	0.5362	0.8952	1.1267	83.80%	85.23%
Simple Regression	0.4401	0.9551	1.1927	89.44%	90.23%
Additive Regression	0.4718	0.9010	1.1840	84.34%	89.57%

Table 8: Unpaired t-test results for selected lyrical features.

Feature Name	Two-tailed p-value	t-statistic
Word diversity	$p < 0.0001$	16.8605
Soundex diversity	$p < 0.0001$	17.3182
Punctuation characters ratio	$p < 0.0001$	20.1844
Frequency of "!"	$p < 0.0001$	5.9971
Frequency of ","	$p < 0.0001$	18.3884
Word length mean	$p < 0.0001$	8.3307
Frequency of "'"	$p < 0.0001$	5.0471
Soundex homogeneity hist bin 0	$p < 0.0001$	17.2163
FFT histogram bin 2	$p < 0.0001$	8.3386
Sentence length mean	$p < 0.0001$	16.3105

Table 9: LDA-derived topics for Top Artists.

Topic ID	Top Keywords
Topic 1	like, baby, let, time, make, feel, love, tell, wanna, want, need, right, way, come
Topic 2	like, come, high, want, make, night, life, money, need, real, really, let, tell, big
Topic 3	like, make, time, come, tell, think, let, need, wanna, girl, way, money, black, new
Topic 4	love, come, like, girl, baby, long, need, look, want, feel, new, make, time, way

- Topic 2: Romantic intimacy and domestic longing
- Topic 3: Desire, comfort, and emotional vulnerability
- Topic 4: Explicit, aggressive, or rap-centered identity
- Topic 5: Strength, conflict, and cultural self-expression
- Topic 6: Life struggles, realism, and existential concerns
- Topic 7: Freedom, rebellion, and emotional release

Table 10: LDA-derived topics for Low Artists.

Topic ID	Top Keywords
Topic 1	love, feel, like, want, tell, think, time, make, day, need, let, right, leave, thing
Topic 2	baby, come, let, love, time, little, long, lie, tell, like, way, home, night, girl
Topic 3	love, wanna, make, like, good, fall, want, sleep, way, wonder, home, time, baby, tell
Topic 4	like, make, fuckin, rap, ass, jou, die, right, hosh, fokken, maybe, carson, game, fake
Topic 5	like, need, make, boy, wanna, ninja, soul, hit, come, look, safe, think, maak, time
Topic 6	life, care, like, make, real, come, let, way, day, use, high, love, time, die
Topic 7	like, let, away, hear, wild, wanna, white, make, free, ass, break, boy, baruch, want
Topic 8	like, way, look, beat, make, step, hit, feel, perfect, face, rat, good, girl, boy
Topic 9	come, like, river, crackin, nkqo, boy, rich, world, stop, run, death, line, change, thing
Topic 10	like, run, home, bang, hard, come, wanna, pocot, rhyme, cause, love, make, guru, brother

- Topic 8: Performance, rhythm, and stylistic presentation
- Topic 9: Social change, mortality, and transformation
- Topic 10: Brotherhood, community, and group identity

4.5.3 Comparative Analysis

The strong contrast between the two artist tiers demonstrates that:

- **Top Artists** exhibit a compact set of coherent, emotionally accessible themes.
- **Low Artists** display greater thematic fragmentation, linguistic experimentation, and cultural specificity.

Even when overlapping high-level themes appear (e.g., love, longing, ambition), Top artists use more polished, mainstream vocabulary, whereas

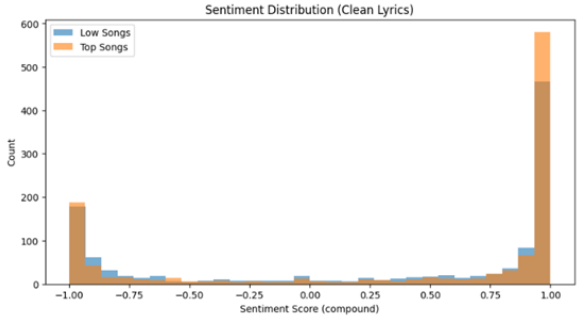

Figure 3: Overlaid sentiment distributions of top- and low-playcount songs based on the compound sentiment score.

Low artists employ rawer, more varied, and culturally rooted language. This divergence provides quantitative evidence that lyrical complexity and thematic diversity tend to be higher among less commercially successful artists, whereas Top artists cluster around a smaller, more predictable set of motifs that appeal to broader audiences.

4.5.4 Sentiment Analysis

Figure 3 shows the overlaid sentiment distributions of top- and low-playcount songs based on the compound sentiment score. Both groups exhibit a strongly skewed distribution toward positive sentiment, indicating that emotionally positive language is prevalent in song lyrics overall. However, top-playcount songs display a higher concentration of extreme positive sentiment values near 1.0 compared to low-playcount songs. In contrast, low-playcount songs show a relatively wider spread across neutral and negative sentiment ranges. These results suggest that highly popular songs tend to emphasize stronger positive emotional expression, whereas less popular songs exhibit greater emotional diversity.

4.5.5 Word Clouds

Figures 4 and 5 present word cloud visualizations for both groups based on cleaned lyric corpora. Both groups frequently use emotionally and relationally charged terms such as *love*, *feel*, and *need*, indicating that affective expression is a common characteristic of song lyrics across popularity levels. However, lower-popularity artists exhibit a broader range of thematic expressions, including terms associated with daily life and social context (e.g., *home* and *gang*). This pattern suggests that lower-popularity artists tend to explore more diverse lyrical themes, whereas top artists concentrate more heavily on a narrower set of emotionally focused topics.

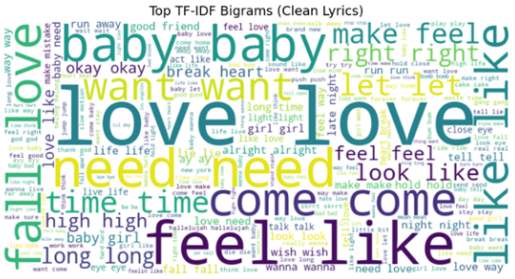

Figure 4: Word cloud visualization of cleaned lyrics for Top Artists, highlighting frequently occurring terms based on TF-IDF weighted bigrams.

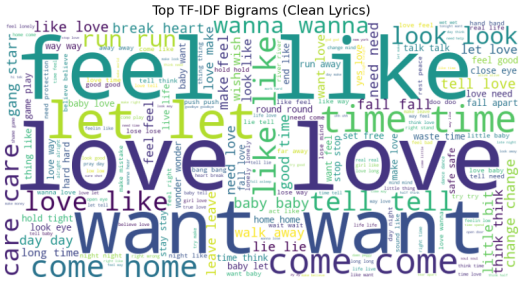

Figure 5: Word cloud visualization of cleaned lyrics for Low Artists, illustrating a broader range of thematic expressions based on TF-IDF weighted bigrams.

4.5.6 Unsupervised Discovery: Stylistic Clustering

K-means clustering with $k = 4$ identified four distinct stylistic clusters among the total songs. (Table 11). Two clusters were dominated by popular songs: Simple Positive (87% popular songs, $n = 610$) and Party-Themed (88% popular songs, $n = 540$). The remaining two clusters—Complex Introspective ($n = 623$) and Mixed Balanced ($n = 529$)—were dominated by unpopular songs, with

88% of songs classified as unpopular and only 12% popular in each cluster. Clear differences were observed across clusters in terms of linguistic diversity and thematic composition.

Table 11: Summary of lyrical clusters showing cluster size, proportion of popular songs, and dominant thematic content.

Cluster	**Songs**	**% Popular Songs**	**Dominant Topics**
Simple Positive	610	87%	Party, Love
Complex Introspective	623	12%	Introspection, Spirituality
Party-Themed	540	88%	Party/Club, Material
Mixed Balanced	529	12%	Multiple

5 Conclusion and Discussion

5.1 Conclusion

This study concludes that lyrical content alone is sufficient to distinguish between commercially successful and less successful songs. By integrating exploratory data analysis, feature extraction, supervised learning, statistical significance testing, and unsupervised modeling, the analysis consistently revealed clear and interpretable differences between Top Artists and Low Artists using only lyrical information. Unsupervised clustering based on lyrical style revealed a clear separation between groups dominated by popular and unpopular songs, where clusters with simpler, thematically focused lyrics contained a much higher proportion of popular songs, while clusters with more varied and introspective themes were largely composed of unpopular songs, indicating that lyrical style captures meaningful differences related to song popularity. Across multiple analytical perspectives, lyrical features demonstrated stable discriminative power. Classification and regression models achieved meaningful predictive performance, while statistical tests confirmed systematic differences in linguistic complexity, thematic emphasis, and emotional expression. These findings collectively establish lyrics as a standalone analytical signal capable of separating success tiers without reliance on audio features, marketing information, or production metadata.

Overall, the results provide strong empirical evidence that linguistic structure, thematic focus, and emotional framing embedded in lyrics play a substantive role in differentiating musical success.

5.2 Discussion

The observed discriminative power of lyrics can be explained by consistent stylistic and thematic patterns associated with commercial success. Exploratory analysis revealed that Top Artists exhibit highly skewed play count distributions and are strongly concentrated in the post-2010 streaming era, suggesting that successful lyrical patterns are embedded within contemporary music ecosystems rather than evenly distributed across historical periods.

At the corpus level, the dataset demonstrated sufficient lexical diversity, document length, and structural regularity to support robust linguistic modeling. This validates the reliability of downstream analyses and indicates that the observed distinctions are not artifacts of sparse or inconsistent textual data.

Feature-based analyses using UDAT and WEKA further revealed that lyrical success is not driven by a single dominant attribute. Instead, multiple complementary features—including word diversity, readability, punctuation usage, and frequency-based signal descriptors—consistently emerged across different evaluators. This convergence suggests that popularity is encoded through distributed stylistic patterns rather than isolated linguistic markers.

Statistical significance testing and topic modeling provide additional interpretive insight. Popular songs tend to emphasize emotionally accessible themes such as party, love, and relationships, exhibit simpler readability, and display stronger positive sentiment. In contrast, less popular songs show greater thematic diversity, narrative fragmentation, and emotional range, including stronger introspective and socially grounded content. While this diversity reflects richer expressive exploration, it appears less aligned with dominant commercial motifs.

Together, these findings suggest that commercial

success in music is associated with stylistic convergence and emotional immediacy, whereas lower popularity is associated with expressive diversity and thematic depth. Lyrics therefore function not only as creative expression, but also as structured signals that reflect and reinforce mainstream audience alignment.

5.3 Limitations

This study focuses on lyrical content as the sole source of information to examine its discriminative power with respect to song popularity. While the results demonstrate that lyrics alone are sufficient to distinguish success tiers, the analysis does not address causal mechanisms or external factors such as marketing strategies, platform recommendation dynamics, or audience exposure effects. Consequently, the findings should be interpreted as evidence of associative and discriminative patterns rather than direct causal drivers of popularity.

In addition, the dataset is constructed based on artist-level popularity groupings, which may conflate individual song effects with broader artist branding or audience recognition. Although this design aligns with the study's objective of distinguishing success tiers using lyrical signals, future work may benefit from finer-grained song-level or longitudinal analyses.

5.4 Future Work

Future research can extend this framework in several directions while retaining lyrics as a central analytical modality. First, integrating audio features, listener behavior, or platform-level exposure signals may improve predictive accuracy and help disentangle the relative contributions of linguistic and non-linguistic factors. Such multimodal extensions would enable a more comprehensive understanding of how lyrical signals interact with production and distribution dynamics.

Second, longitudinal and causal analyses could be employed to examine how lyrical styles evolve over time and whether shifts in thematic focus or emotional framing precede changes in popularity. Finally, cross-genre and cross-lingual studies may further assess the generalizability of the observed patterns and explore whether similar stylistic convergence characterizes successful music across different cultural contexts.

Author Contributions

An Yu Yeh: Top Artist data collection,Weka regression analysis, Latent Dirichlet Allocation (LDA), sentiment analysis, word cloud generation, paper writing (co-author).
Chakrapani Gajji: Top Artist data collection, data cleaning, data preprocessing, feature engineering(Weka), feature selection(Weka), MALLET classification, WEKA classification, paper writing(co-author).
Prathyusha Mardhi: Low Artist data collection, exploratory data analysis (EDA), data visualization, UDAT classification, feature engineering (UDAT), feature selection (UDAT), statistical significance analysis, classification (Scikit), unsupervised discovery:stylistic clustering, paper writing (primary author).

References

(2023). Song popularity and processing fluency of lyrics. *Sage Journals*.

Choudhary, Y., Rao, P., and Bhattacharyya, P. (2025). Lyrics matter: Exploiting the power of learnt representations for music popularity prediction. *arXiv preprint*.

Fell, H. and Sporleder, C. (2014). Lyrics-based analysis and classification of music. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 620–631. Association for Computational Linguistics.

GraphPad Software (2024). Quickcalcs: Unpaired t test calculator. https://www.graphpad.com/quickcalcs/ttest1/?format=sem. Accessed: 2025-09-16.

Hu, X. and Downie, J. S. (2010). When lyrics outperform audio for music mood classification: A feature analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 619–624.

Kworb.net (2025). Spotify most streamed artists of all time. Accessed: 2025-12-08.

Loong, J. (2018). Topic modelling song lyrics: An exploration in unsupervised text analytics.

Mazharov, M. (2020). How streaming platforms affect popularity of songs. Bba thesis, Tallinn University of Technology.

Middlebrook, J. and Sheikh, Y. (2019). Predicting hit songs using machine learning.

Selenium Project (2024). Selenium webdriver. https://www.selenium.dev/. Accessed: 2024-11-19.

Shamir, L. (2011). A general-purpose distributed system for content-based analysis of digital documents. In *Proceedings of the 2011 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, pages 340–346.

Singhi, A. and Brown, D. G. (2015). Can song lyrics predict hits? In *CMMR*.

Sterckx, C., Demeester, T., Deleu, T., and Develder, C. (2014). Assessing quality of unsupervised topics in song lyrics. In *Advances in Information Retrieval*, pages 573–584. Springer.

Thwe, K. Z. and Yukawa, T. (2019). Improving music recommendation system by applying latent topics of lyrics. *International Journal of Informatics and Information System*, 2(2):91–98.

Zangerle, E., Vötter, M., Huber, R., and Yang, Y.-H. (2019). Hit song prediction: Leveraging low- and high-level audio features. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 319–326.