The Power of Lyrics

Analyzing the relationship between lyrical complexity and song popularity

Abstract

The rise of digital music platforms has provided unprecedented access to vast amounts of song metadata and lyrics, enabling new forms of musical analysis. This project explores the relationship between lyrical complexity and song popularity across 60,000 songs from Spotify. By leveraging natural language processing techniques, we quantify lyrical complexity using readability metrics such as the Flesch-Kincaid Index. Through visualizations such as scatter plots, heatmaps, and word clouds, we investigate patterns of lyrical complexity across different genres, time periods, and artists, as well as how it correlates with popularity metrics like stream counts. This research aims to uncover whether simpler or more complex lyrics drive higher popularity in music and whether this varies by genre or over time. Additionally, we provide insights into the most common lyrical themes in popular music through word cloud visualizations. The findings contribute to a deeper understanding of modern music consumption and the role of lyrical content in a song's commercial success.

Introduction

The increased accessibility of music to the public has resulted in an annual surge of investments (approximately $4.1 billion) from record labels seeking new talent[2]. With the exponential growth of available music tracks[3], there is a pressing need to develop innovative methods for efficiently searching and retrieving musical content. This demand has led to the emergence of Music Information Retrieval (MIR), a field within machine learning that focuses on classifying various aspects of music, such as emotions, instruments, artists, and genres, as well as identifying similar musical content[4]. By comparing different attributes of songs, MIR facilitates predictions regarding songs that share specific characteristics.

In this study, we aim to explore whether lyrical complexity plays a role in a song's popularity on streaming platforms, specifically focusing on Spotify. While several studies have analyzed the impact of genres, tempo, or artist popularity on song success, the relationship between lyrical depth and listener engagement remains underexplored.

This project will attempt to uncover patterns between the semantic richness of song lyrics and their popularity metrics on Spotify. Understanding these correlations may inform musicians and record labels on how to craft songs that resonate more deeply with audiences.

Motivation

Our project seeks to fill a critical gap in existing research. While much attention has been placed on musical features such as melody, beat, and tempo, lyrical analysis is an area with tremendous potential. We hypothesize that listeners may subconsciously gravitate toward songs with more emotionally engaging or complex lyrics, influencing their overall popularity.

Several industry studies have explored factors such as song tempo or streaming frequency, but none have rigorously examined the link between textual complexity and song success. Furthermore, cognitive studies have shown that humans find complex linguistic structures more intellectually stimulating. Therefore, songs with intricate lyrics could foster higher engagement levels and repeat listens. This project will analyze whether that hypothesis holds true by correlating lyrical complexity with Spotify popularity data, introducing a novel intersection of music analytics and cognitive linguistics.

3.1 Existing Research on Relevant Work and Visualizations

Prediction of the popularity of pop music by using a large dataset and focusing on a combination of unique lyrical features and common text-mining features has revealed unsatisfactory results[1]. The confusion matrices used for visualizing these results showed that none of the models used were competent enough to identify the highly popular songs.

The confusion matrix in *Fig. 1.* provides a visualization of the KNN classifier's classification results. With a baseline accuracy of 0.46, labeling all data points as the majority class would still yield 50% accuracy, suggesting inadequate model performance. The heatmap used for visualizing the KNN model's results is unsuitable due to its complexity, making it difficult to interpret the underlying data patterns.

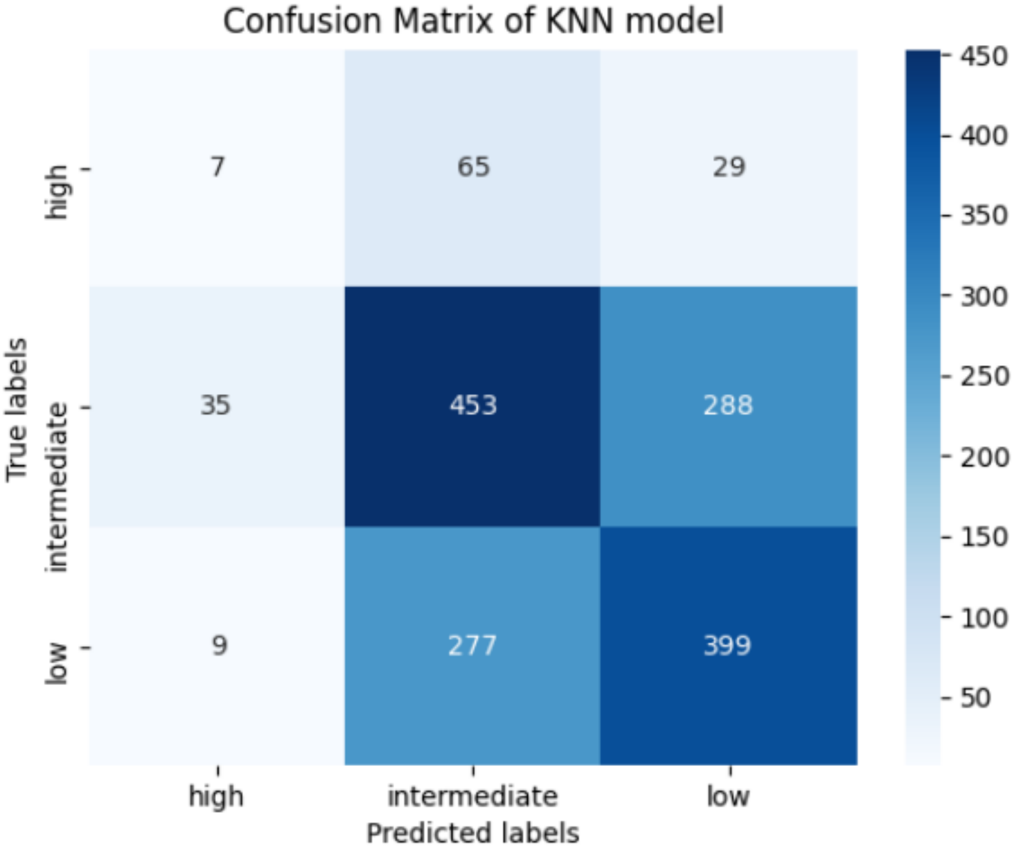

Fig. 1. Confusion matrix for KNN model trained on the mixed feature set [1]

The SVM model achieved the highest accuracy of 0.58 among all evaluated models, with precision and F1-scores of 0.59 and 0.60, respectively. However, these scores indicate that fewer than two-thirds of the data points were correctly classified. While the SVM improved intermediate track classification, it struggled with distinguishing between intermediate and low popularity, similar to the KNN model.

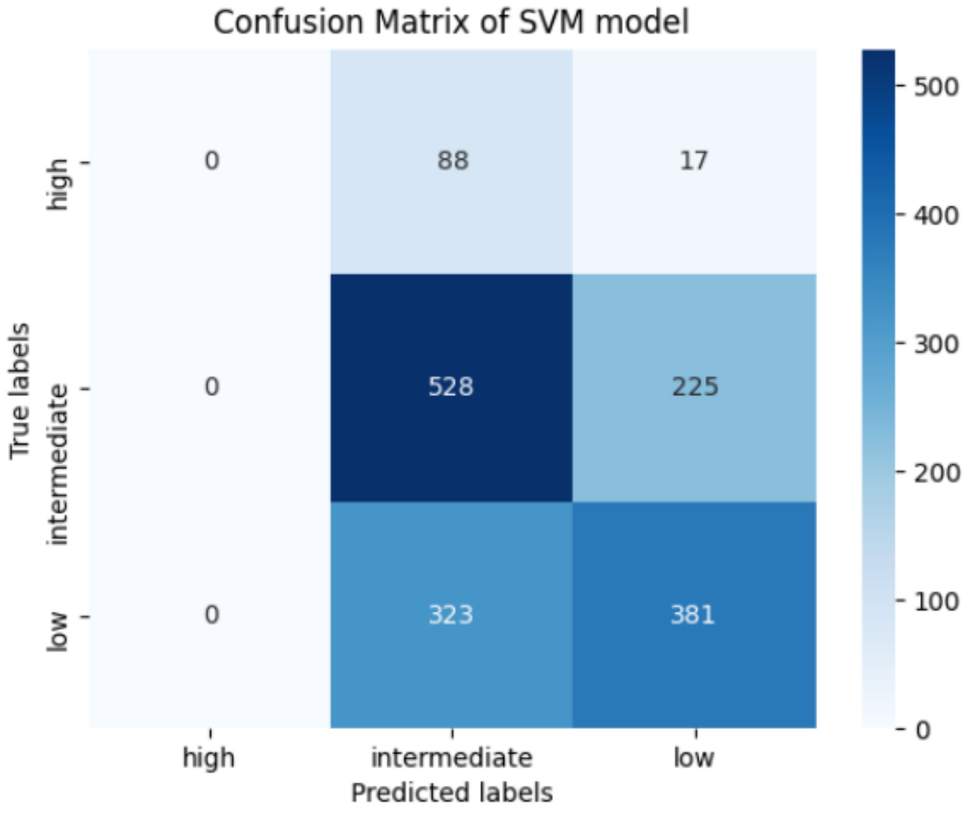

Fig. 2. Confusion matrix for SVM model with unbalanced class weights, trained on the tf-idf feature set [1]

On adding a parameter to balance class weights, the new balanced model did not achieve the original model's scores on the mixed and tf-idf feature sets. However, it outperformed other classifiers on the concatenated feature set. The confusion matrix indicated a slight improvement in correctly classifying minority class entries, suggesting further investigation could make this a viable option for us to explore in our research.

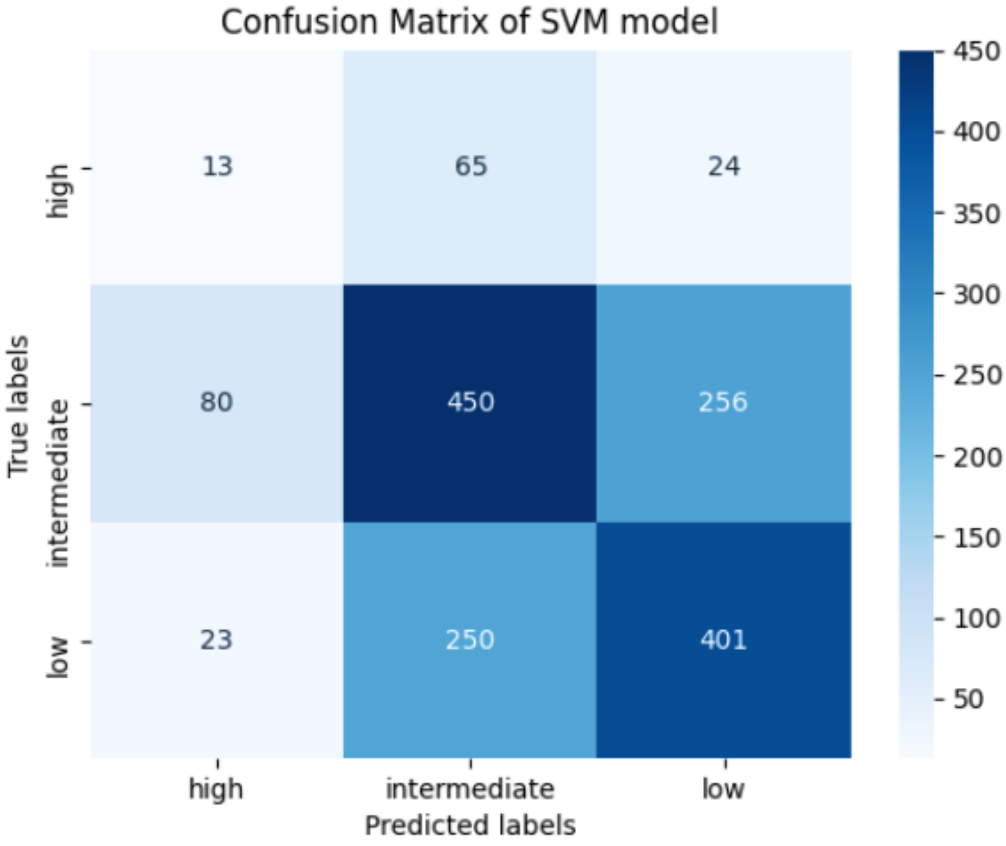

Fig. 3. Confusion matrix for SVM model with balanced class weights, trained on the concatenated feature set [1]

The Naive Bayes Classifier achieved the highest accuracy on the mixed feature set. Its recall score was relatively high compared to the baseline performance, indicating better classification of some classes. However, the overall performance suggests significant potential for further improvement in accurately predicting data points across all classes.

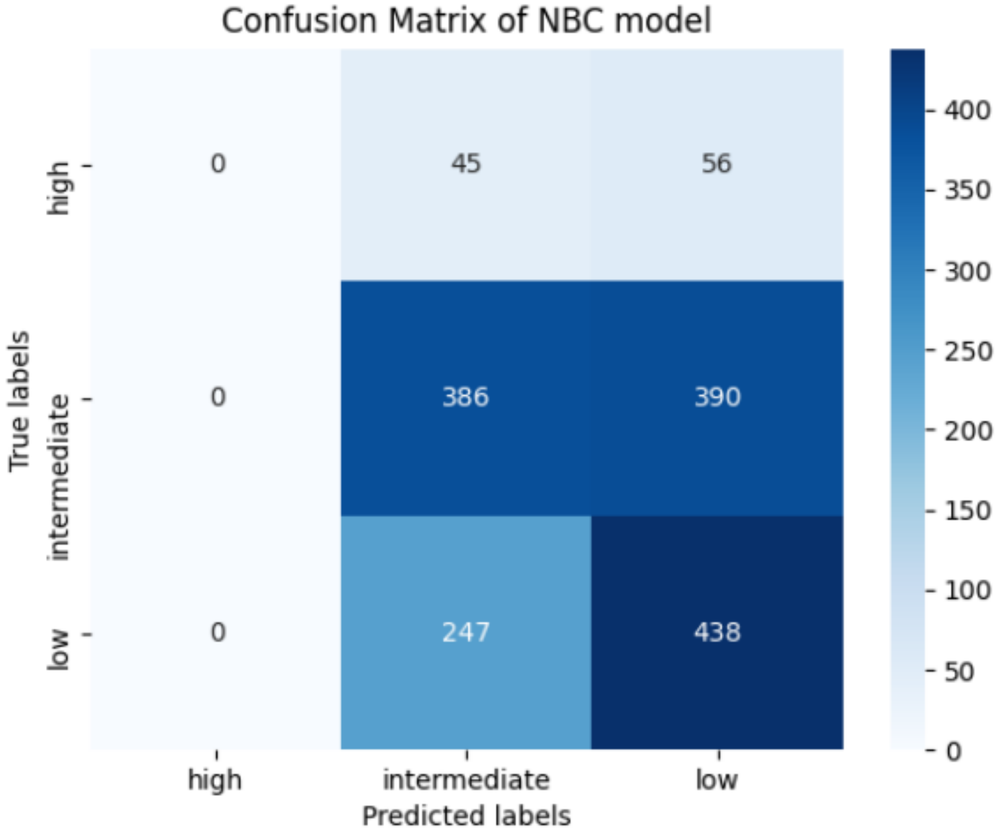

Fig. 4. Confusion matrix for NBC model trained on the mixed feature set [1]

Within the fields of hit song science and Music Information Retrieval (MIR), various studies have been conducted, some focusing explicitly on the lyrical aspects of songs, while others emphasize acoustic features. For instance, Herremans[4] utilized multiple classifiers to predict whether dance songs made it into the top 10 charts by analyzing acoustic features, metadata, and temporal data, yielding promising results from classifiers like Logistic Regression (LR) and Naive Bayes Classifier (NBC). Similarly, Interiano et al.[6] trained a random forests model on over 500,000 tracks, primarily using acoustic features, including a 'superstar' attribute for artists with previous chart appearances. Moreover, research on Turkish music[7] demonstrated the effectiveness of SVMs using a comprehensive feature set of audio and lyrical elements, achieving an impressive accuracy of over 99%. However, the absence of diverse performance metrics calls into question the reliability of their results.

Numerous previous studies have explored various factors influencing a song's success. Prior research has investigated the evolution of popular music through genre analysis[11], and examined how sentiment in lyrics changed over time. However, studies focusing on the specific relationship between lyrical complexity and listener engagement are scarce. Most research in the field of music analytics has examined broader, surface-level metrics such as genre trends or artist fame, rather than the content of the lyrics themselves.

3.2 Proposed Enhancements

In terms of visualizations, existing studies have used heat maps to depict lyrical trends, but they rarely map out correlations between complexity and streaming metrics and are harder to comprehend. Our study plans to utilize more intuitive visualizations, such as scatter plots and line graphs to visually correlate metrics such as Flesch Reading Ease with popularity scores, and word clouds to visualize common lyrical themes. By introducing topic modeling and word embeddings, we will move beyond basic text-mining methods and offer a more sophisticated analysis of how lyrical themes influence a song's success.

Data Sources and Dataset Appropriateness

For this project, we will rely on two main data sources:

- **Spotify's Worldwide Daily Song Ranking Dataset[16]:**

This dataset contains the daily rankings of approximately 3.4 million songs across various countries, specifically focusing on the daily ranking of the 200 most listened to songs in 53 countries from 2017 and 2018 by Spotify users. It includes over 2 million rows, encompassing 6,629 artists and 18,598 songs, totaling an impressive 105 billion stream counts. The data spans from January 1, 2017, to January 9, 2018, collected from Spotify's regional chart data. This dataset provides crucial metadata, including song popularity metrics such as stream counts, user ratings, and listener engagement over time, allowing for a comprehensive analysis of patterns in song popularity across different geographical areas, offering insights into both local and international music trends. Both well-known and obscure songs are included, providing a diverse perspective on music consumption.

Column Name	Description	Data Type
Position	The daily ranking position of the song	Integer
Track Name	The title of the song	String
Artist	The performer or group of the song	String
Streams	The total number of streams the song received that day	Integer
URL	The Spotify URL link to the song	String (URL)
Date	The date of the ranking entry	Date
Region	The country or region where the ranking applies	String

Table 1. Schema for Spotify's Worldwide Daily Song Ranking Dataset

- **Song Lyrics Dataset[17]:**

The dataset we are working with contains around 60,000 song lyrics with fields for artist, song title, lyrics, and hyperlinks to the original text. This allows for in-depth text analysis in the ways in which lyrics impact a song's popularity or impact over time. An extensive summary of the corresponding columns and data types is provided below:

Column Name	Description	Data Type
Artist	The performer or group of the song	String
Song	The title of the song	String
Link	URL link to the original lyrics source	String (URL)
Text	The full lyrics of the song.	String (Text)

Table 2. Schema for Song Lyrics Dataset

The Spotify API will supplement this dataset with popularity metrics like play counts and likes, while also providing additional contextual data, such as release date, album information, and artist popularity. We will ensure the dataset is complete by checking for any missing or incomplete data.

Methodology

Our research methodology will focus on three key stages:

1. Preprocessing the Lyrics:

- Cleaning the lyrical text by removing stop words, punctuation, and unnecessary symbols.
- Perform tokenization and lemmatization to break down the lyrics into meaningful linguistic units.

2. Lyrical Complexity Analysis:

- Using readability indices like the Flesch-Kincaid Reading Ease and Gunning Fog Index, we will measure the readability and complexity of each song's lyrics.
- Computation of the lexical diversity (type-token ratio) to assess the breadth of vocabulary used in each song.

3. Popularity Correlation:

- For each song, we will extract popularity metrics from Spotify, including play counts, likes, and listener retention rates.
- Using Pearson's correlation coefficient and multiple linear regression, we will analyze the relationship between lyrical complexity and song popularity.
- We will also segment the analysis by genre, artist, and year of release to determine if complexity correlates differently across different categories.

4. Visualizations:

- We will employ scatter plots to visualize the correlation between complexity metrics and popularity scores.

- Heatmaps will display the relationship between lyrical diversity and engagement metrics across different genres and time periods.
- Word clouds and sentiment maps will provide an intuitive visualization of thematic trends and their potential impact on song success.

Expected Outcomes

We expect to uncover significant insights into how lyrical complexity influences song popularity:

- Songs with higher lexical diversity and thematic richness may show stronger listener engagement and popularity.
- Certain genres, such as pop and indie, may see stronger correlations between complexity and success compared to others like hip-hop or EDM.
- The project's predictive model will offer a potential tool for predicting a song's success based on its lyrics, benefiting songwriters and the music industry at large.

Challenges and Limitations

- Data Imbalance: Some songs may have incomplete popularity data (e.g., songs that are not widely played or liked). We will handle this by filtering out songs with incomplete metadata.
- Genre Bias: Certain genres may inherently have simpler or more complex lyrics, skewing the results. We will attempt to mitigate this by performing genre-specific analysis.
- Listener Subjectivity: The impact of lyrics on a song's success might vary based on cultural and individual listener preferences. Future research could consider listener demographics.

References

[1] Somme SVD, Sogancioglu G, Paperno D. Popularity of music tracks based on lyrics. Utrecht University; 2021.

[2] The Role of a Record Company. url: https://powering- the- musicecosystem.ifpi.org/.

[3] M. A. Casey et al. “Content-Based Music Information Retrieval: Current Directions and Future Challenges”. In: Proceedings of the IEEE 96.4 (2008), pp. 668–696. doi: 10.1109/JPROC.2008.916370.

[4] Dorien Herremans, David Martens, and Kenneth S¨orensen. “Dance Hit Song Prediction”. In: Journal of New Music Research 43.3 (2014), pp. 291– 302. doi: 10.1080/09298215.2014.881888. eprint: https://doi.org/ 10.1080/09298215.2014.881888. url: https://doi.org/10.1080/ 09298215.2014.881888.

[5] Sumiko Asai. “Factors Affecting Hits in Japanese Popular Music”. In: Journal of Media Economics 21.2 (June 2008), pp. 97–113. doi: 10.1080/ 08997760802069895.

[6] Myra Interiano et al. “Musical trends and predictability of success in contemporary songs in and out of the top charts”. In: Royal Society Open Science 5.5 (May 2018), p. 171274. doi: 10 . 1098 / rsos . 171274. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5990848/.

[7] Onder C¸ oban. “Turkish Music Genre Classification using Audio and Lyrics ¨ Features”. In: Suleyman Demirel University Journal of Natural and Applied Sciences (May 2017). doi: 10.19113/sdufbed.88303.

[8] Ricardo Malheiro et al. “Emotionally-Relevant Features for Classification and Regression of Music Lyrics”. In: IEEE Transactions on Affective Computing 9 (Jan. 2018), pp. 240–254. doi: 10.1109/TAFFC.2016.2598569.

[9] Markus Schedl. “Genre Differences of Song Lyrics and Artist Wikis: An Analysis of Popularity, Length, Repetitiveness, and Readability”. In: The World Wide Web Conference. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 3201–3207. isbn: 9781450366748. doi: 10 . 1145 / 3308558 . 3313604. url: https : / / doi . org / 10 . 1145 / 3308558.3313604.

[10] Kahyun Choi et al. “Music subject classification based on lyrics and user interpretations”. In: Proceedings of the Association for Information Science and Technology 53.1 (2016), pp. 1–10. doi: https://doi.org/10. 1002/pra2.2016.14505301041. eprint: https://asistdl.onlinelibrary. wiley.com/doi/pdf/10.1002/pra2.2016.14505301041. url: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2016. 14505301041.

[11] Mauch, M., et al. (2015). The evolution of popular music: USA 1960–2010. *Royal Society Open Science*.

[12] Arora, S., Rani, R. Soundtrack Success: Unveiling Song Popularity Patterns Using Machine Learning Implementation. *SN COMPUT. SCI.* **5**, 278 (2024). https://doi.org/10.1007/s42979-024-02619-5

[13] Yee YK, Raheem M. Predicting music popularity using spotify and youtube features. Indian J Sci Technol. 2022;15:1786–99. https://doi.org/10.17485/ijst/v15i36.2332.

[14] Brooks, A., et al. (2019). Sentiment Analysis of Lyrics in Popular Music. *Journal of Music Psycholgy*.

[15] Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*.

[16] https://www.kaggle.com/datasets/edumucelli/spotifys-worldwide-daily-song-ranking

[17] https://www.kaggle.com/datasets/joebeachcapital/57651-spotify-songs/data