Unisia
2022, Vol. 40, No. 1, 1-26
https://doi.org/10.20885/unisia.vol40.iss1.art1

Check for updates

Data Science Approach to Compare the Lyrics of Popular Music Artists

Caleb Rosebaugh & Lior Shamir

Department of Computer Science, Kansas State University, Manhattan, Kansas State, United States

ABSTRACT

Popular music lyrics exhibit clear differences between songwriters. This study describes a quantitative approach to the analysis of popular music lyrics. The method uses explainable measurements of the lyrics and therefore allows the use of quantitative measurements for consequent qualitative analyses. This study applies the automatic quantitative text analytics to 18,577 songs from 89 popular music artists. The analysis quantifies different elements of the lyrics that might be impractical to measure manually. The analysis includes basic supervised machine learning, and the explainable nature of the measurements also allows to identify specific differences between the artists. For instance, the sentiments expressed in the lyrics, the diversity in the selection of words, the frequency of gender-related words, and the distribution of the sounds of the words show differences between popular music artists. The analysis also shows a correlation between the easiness of readability and the positivity of the sentiments expressed in the lyrics. The analysis can be used as a new approach to studying popular music lyrics. The software developed for the study is publicly available and can be used for future studies of popular music lyrics.

Keywords

automatic quantitative text analytics; basic supervised machine learning; popular music lyrics

Citation: Rosebaugh, C., & Shamir, L. (2022). Data science approach to compare the lyrics of popular music artists. *Unisia*, *40*(1), 1–26. https://doi.org/10.20885/unisia.vol40.iss1.art1

Article History

Received: February 14, 2022
Revised: April 18, 2022
Accepted: April 29, 2022
Published: July 3, 2022

Publisher's Note: Universitas Islam Indonesia stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Contact: Lior Shamir ✉ lshamir@k-state.edu

INTRODUCTION

The style and topics of popular music lyrics have clearly changed over time. Common topics and lyrics style used in the early days of popular music are substantially different from the lyrics of modern popular music, and different popular music artists express themselves by using their unique lyrics styles. For instance, while notable early songwriters focused mostly on love and romance (Ruth, 2019; Yoo et al., 2017), preeminent songwriters during the 1960s such as Bob Dylan or the Beatles often focused their lyrics on topics related to social or political issues (Lammer, 2016). It has also been shown that popular music songwriters are often influenced from events outside of the world of popular music such as political events (Fox & Williams, 1974; Gosa, 2017; Nielson, 2009).

Lyrics can exhibit substantial differences between popular music genres (Ballard et al., 1999; Condit-Schultz & Huron, 2015; Tsaptsinos, 2017). For instance, it had been shown that gender differences are expressed differently through lyrics in different pop music genres (Flynn et al., 2016; Freudiger & Almquist, 1978), and that rap lyrics can be linked to mental conditions (Kresovich et al., 2021). It has also been shown that lyrics can be different across songwriters, and certain songwriters can have signature characteristics in their lyrics style. For instance, analysis of the lyrics of Bob Dylan showed different aspects that characterize Dylan's songs, and can be linked to the period during which his songs were written (Davies, 1990; Dunlap, 2006). Stylistic and linguistic analyses of the lyrics of the Beatles also showed repetitive patterns and concepts common in the Beatles lyrics (Petrie et al., 2008; West & Martindale, 1996).

With the digitization of data and the availability of lyrics in digital formats, several studies used computer analysis for the purpose of analyzing song lyrics. For instance, in the computer science community certain efforts have been invested in automatic classification of popular music lyrics (Fell & Sporleder, 2014; Tsaptsinos, 2017), and showed that automatic classification of lyrics into genres can be done in accuracy significantly better than mere chance (Logan et al., 2004). Another growing related direction of research is identification of popular music songs by their emotions (An et al., 2017; Yeh et al., 2014).

The ability to analyze lyrics by using quantitative methods has enabled new paradigms of studying popular music that were not possible in the pre-information era. For instance, the ability to quantify sentiments expressed in thousands of songs has shown the change of sentiments expressed in popular music lyrics over time (Napier & Shamir, 2018). Statistical analysis and text mining enabled the identification of changes in patterns of words used in Korean pop music lyrics (Yoo et al., 2017). North et al. (2021) used computer analysis to identify links between music and lyrics in popular music songs. Yang (2020) used computational analysis to study long-term patterns in popular music such as loudness, repetitiveness, and simplicity. Quantitative analysis was also used to profile cover song relationship between popular music artists (Ortega, 2021). Other applications of text analysis related to popular music includes non-lyrics tasks such as analyzing on-line discussions about popular music artists (de Boise, 2020).

Popular music is one of the most influential forms of art, with impact not merely on leisure, but also on politics, society, and culture. For instance, the dominant societal movements during the 1960s were largely enabled by popular music, and popular music artists played impactful social leadership role during that time (Kutschke, 2016). As music was pivotal in these movements, the cultural and social changes they accomplished could have been different in the absence of popular music (Vandagriff, 2015). The Vietnam war era was also heavily impacted by popular music, demonstrating the impact of popular music on political issues (Bindas & Houston, 1989). On the opposite side of that scale, popular music has been widely used by fascist movements (Richardson, 2017), using music as a mean of communication. Music was used by these movements as a tool of communication, coordination, and recruitment of new members, especially before the emergence of the Internet. The music used for these purposes was also characterized by unique lyrics style. As popular music is the most popular forms of art, and popular music artists continue to have paramount societal impact (Martin, 2006), it is important to study popular music by using all available analysis tool and paradigms.

This study uses a data science approach to identify quantitative stylistic elements that can differentiate between songwriters and provide cues and new knowledge about the unique lyrics styles that characterize different songwriters. That is done by first computing a large collection of quantitative text measurements from each song. These quantitative elements can be used to analyze differences in lyrics written by different songwriters. The computational approach allows to measure stylistic elements used by songwriters, and its automatic nature allows the analysis of very large datasets of songs to extract new knowledge about popular music.

As briefly described above, substantial work has been focused on the ability to classify lyrics automatically, or to measure specific lyrical elements. The approach proposed here makes use of a large number of text measurements applied in concert, therefore allowing to identify text elements that exhibit differences between songwriters in a data-driven manner. That is, the user does not need to make a specific hypothesis or focus on a specific text measurement but can rely on the method to analyze and identify specific text elements. Such measurements can show differences between different songs, albums, songwriters, genres, and more. Instead of using measurements that can discriminate between the songwriters and associate a song with its creating artist automatically, the analysis here is based on using intuitive measurements that can be interpreted manually and can therefore expand the knowledge regarding popular music. Although these text measurements are intuitive to understand, they are difficult to analyze by manual reading of the lyrics. Therefore, the proposed approach provides a new way to analyze and study lyrics that was not practical in the pre-information era.

METHOD

Data

The dataset used in this study contained text files of lyrics of 89 artists, such that each artist in the dataset is represented by at least 100 songs. The total number of songs was 18,577, with an average of 209 songs per artist. The artist with the lowest number of songs was Guns N' Roses with 101 songs, and the artists with the highest number of songs was Neil Young with 576 songs. The distribution of the songs of the different artists as well as the year in which the songs were released are shown in Table 1.

The lyrics files were retrieved from AZlyrics.com. The lyrics files were originally in HTML format and were converted to plain text (ascii). The files were formatted such that the files only included the lyrics, and no other information such as the name of the artist, album, or year of release that were included in the file. Naturally, it is impractical to analyze all popular music artists. Artists that were selected in this study are all well-known influential popular music artists that also had lyrics for at least 100 songs. The artists were also selected such that they represent multiple genres and eras of popular music. Genres included rock, pop, hip-hop, soul, R&B, and heavy metal, as primary genres of popular music. The vast majority of the musicians were mostly active as influential songwriters during the 1960s through the 1990s. The list of musicians excludes more recent artists, as scholarly work has not yet been done by the scientific community to fully analyze and characterize these artists, making it more difficult to associate the lyrics elements with qualitative analysis of their lyrics style. For that reason, most artists used in this study were active in previous decades, while artists that were mostly active in the third millennium are not widely represented in the dataset. The full list of artists and the years during which they were active is provided in Appendix A.

Analysis method

The lyrics were analyzed using Udat (Shamir, 2021) which is a tool that extracts a comprehensive set of numerical text content descriptors from text samples. Unlike some common document classifiers, Udat does not attempt to identify certain patterns of words that differentiate between different text classes for the purpose of associating a test text sample with a text class. Instead, it analyzes text elements that reflect patterns in the way the text is written (Shamir, 2021). For that purpose, the measurements are explainable and can be interpreted in a follow-up qualitative manual analysis. The text elements computed by Udat are described in detail in (Shamir, 2021). In summary, they include the following:

1. Readability. Automated Readability Index (Smith & Senter, 1967), and the Coleman-Liau index (Coleman & Liau, 1975) are established methods for estimating the level of difficulty of reading the text. Both methods are based on the length of words and length of sentences, which are expected to provide an indication of the level of reading difficulty. The Automated Readability Index (ARI) is computed by the average word length (AWL) and average sentence length (ASL). The formula (4.71 x AWL + 0.5 x ASL - 21.43) is skewed to prefer a higher AWL. The ARI is similar to the Coleman-Liau Index calculation. According to the Coleman-

Liau , the ASL is replaced by the sentences per word (SAW), which can be defined by 1/ASL. The Coleman-Liau formula (5.88 x AWL - 29.6 x SAW - 15.8) is also skewed to favor a higher AWL. Automated Readability Index and the Coleman-Liau Index are two established readability indices used to quantify the level of reading difficulty. A high score means that the text is more difficult to read, and a low (also negative) score is relatively easy to read.

2. Average sentiments. Sentiments expressed in the lyrics, as computed by automatic sentiment analysis (Socher et al., 2013). Each song is assigned with a sentiment value from 1 to 5, such that 1 reflects very negative sentiment, 3 is neutral, and 5 reflects very positive. The sentiment analysis works by using 215,154 labeled phrases and 11,855 sentence that make a parse tree analyzed by a deep recurrent neural network (RNN). While the analysis process is not necessarily intuitive, the output is a simple score that reflects the estimated sentiment expressed in the text.
3. Differences in sentiments. Sentiment difference is measured by the difference between the most positive sentence in the lyrics and the most negative sentence in the lyrics. That measurement reflects the variations in the sentiments expressed in the lyrics. Songs that include negative and positive sentences have a higher sentiment difference. Songs where the different sentences are more consistent in the sentiments, they express have a lower sentiment difference.
4. Punctuation characters. The method measures the frequency of the use of different punctuation characters. That is the fraction of characters such as '?', '!', ':', etc. in the total number of characters in the lyrics.
5. Word length. Word length statistics includes the word length mean, word length standard deviation, and histogram of the length of the different words in the text. That measurement reflects the use of words of different lengths in the lyrics.
6. Lyrics length. Statistics of the distribution of the length of the lyrics, including the mean and standard deviation of the number of words in the song.
7. Sounds. Diversity of sounds using the Soundex algorithm (Odell, 1956) which encodes each word into sounds regardless of the word spelling. The diversity of the sounds is computed by the number of different sounds used in the lyrics divided by the total number of sounds.
8. Word diversity. Diversity of the words appearing in the lyrics. Determined by the number of unique words divided by the total number of words. A song that uses the same words repetitively will have a lower word diversity score.
9. Parts of speech. Frequency of different parts of speech. The parts of speech are tagged automatically by using the CoreNLP Natural Language Processing library (Manning et al., 2014) part of speech tagger. Then, the frequency of each part of speech is computed by dividing the number of occurrences of a certain part of speech by the number of words in the lyrics.
10. Topics. Frequency of different topics mentioned in the lyrics. The topics are determined by a predefined thesaurus that covers a range of different topics and words that identify them.

The complete list is provided in Shamir (2021). Among these topics, Udat also measures the frequency of terms related to men and terms related to women, which can be used to test gender reference differences in popular music lyrics. A more detailed description of these text content descriptors is available in Shamir (2021) and the code is available for free download (Shamir, 2017).

The numerical text content descriptors allow the identification of specific differences between the way different songwriters express themselves through their lyrics. Additionally, it allows certain machine learning tasks by applying machine learning algorithms to the numerical content descriptors computed from the text. Udat implements the Weighted Nearest Distance (WND) algorithm (Shamir et al., 2008) for classification. WND is an instance-based machine learning algorithm that makes use of weighted harmonic multi-dimensional distances from all samples in the dataset to make a classification, and its main advantage is that it can provide the similarities between the classes and between individual text samples (Shamir, 2021; Shamir et al., 2008).

According to WND, the association between a song and a musician is determined by the musician *m* with the shortest distance *d(s,m)* to the given song *s*, as defined by:

$$d(s,m) = \frac{\sum_{t \in T_m} [\sum_{f \in F} W_f (s_f - t_f)^2]^p}{|T_m|}$$

where T is the set of songs used for training, T_m is the training set of the songs from musician *m*, *t* is a feature vector from T_m, and S_f is the feature vector of song *s*. The feature vectors are the quantitative text elements described above, and also described in Shamir (2021). The exponent p is set to -5, as determined experimentally and fully explained Orlov et al. (2008). The weight W_f of each feature *f* is computed by

$$W_f = \frac{\sum_{m=1}^{N} \left(\overline{T_f} - \overline{T_{f,m}}\right)^2}{\sum_{m=1}^{N} \sigma_{f,m}^2} \cdot \frac{N}{N-1}$$

where W_f is the weight assigned to feature *f*, T_f is the mean of all values of feature f in the training set, and $T_{f,m}$ is the mean of all values of feature f in songs in the training set written by musician m. The variance $\sigma^2_{f,m}$ is the variance of the values of feature f in the songs of musician m. N is the total number of musicians. The use of the weights is explained with detailed empirical results in Orlov et al. (2008), Shamir et al. (2008), and Shamir et al. (2010).

The main goal of Udat is not necessarily to perform automatic classification, but to extract knowledge from the data. The WND algorithm is used for its ability to estimate the similarity between different classes (Shamir et al., 2010). Unlike some document classifiers, the main purpose of Udat is not to provide a "black box" association between a song and a songwriter automatically, but to identify certain intuitive interpretable text measurements that reflect the differences between different songwriters. Such lyrics elements that are consistent in the lyrics of one artist but are different when measured from the lyrics of other songwriters can help define the unique lyrics styles and help identify signature lyrics styles in a quantifiable manner. Due to the large number of

songs and large number of measurements, that analysis is impractical to perform by manual analysis.

RESULTS AND DISCUSSION

The Udat method described above can perform several data science tasks related to analysis of text data, including supervised and unsupervised learning, profiling similarities between text classes, outlier detection, feature selection, and identification of specific text elements that can differentiate between text classes (Shamir, 2021). In this study, automatic classification of the lyrics is followed by feature selection and analysis of specific text elements that can identify between songwriters and genres.

Automatic classification of songs by their lyrics

While numerous previous studies focused on automatic classification of lyrics by their musician or songwriter (Fell & Sporleder, 2014; Logan et al., 2004; Tsaptsinos, 2017), the data science approach aims at turning data into new knowledge rather than providing automation of data annotation. However, a first step of observing the classification accuracy can provide an indication of whether the data contains information that can differentiate between the artists. That is, while automatic labeling of the data is not necessarily the primary goal of a data science approach, it can provide a general indication of whether the different classes can be separated, which indicates on the presence of differentiative signal in the data.

The machine learning algorithm described above was applied to associate between lyrics and artists automatically. That is, the songs of each artist are separated into training and test songs. Eighty songs of each artist were used for the machine learning algorithm to "learn" the style of the lyrics, and the remaining 20 songs of each artist were used to test whether the machine learning algorithm can associate the lyrics with the artist automatically. The machine learning system first "learns" from the 80 training songs of each artist. After the training, the machine learning algorithm attempts to associate each of the 20 test songs with their artist. The accuracy of the algorithm is measured by the number of songs that the algorithm was able to associate correctly with their artist, divided by the total number of songs that the algorithm attempts to associate with an artist, including the songs that were associated with the wrong artist.

The results of the experiment showed that 12.3% of the songs were associated by the algorithm with their correct artist. While that is not necessarily very high accuracy, with 89 artists in the dataset a random guess of the artist would lead to accuracy of 1/89, which is ~1.1%. That shows that even if the machine learning algorithm cannot associate between the lyrics and the artist in all cases, the fact that it makes the association in higher rate than random guess shows that there is information in the quantitative elements that can show patterns that are unique to a certain artist. That is done without identifying certain words or pattern of words that are used differently by different songwriters, but merely with the quantitative measurements. It is important to mention that associating lyrics with an artist is a complex cognitive task, and even

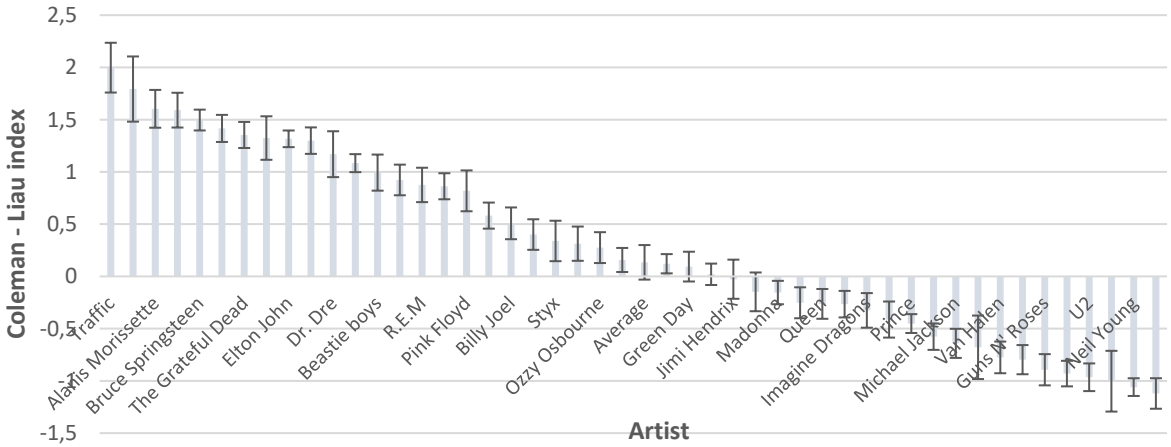
humans might find it difficult to associate lyrics with the relevant creating artist, unless they are already familiar the specific lyrics.

Differences in readability index

To understand the specific differences between the lyrics of popular music artists, the specific measurements were compared. Figures 1 and 2 show the Coleman-Liau readability index and the Automated Readability Index, respectively, of the lyrics of each artist. That is done by calculating the average readability index of all songs of each artist in the dataset. The figure shows the artists with the lowest and highest average readability scores, as well as artists that the average readability score of their lyrics are in between the easiest to read and the most difficult to read. The figures show substantial differences between the readability of the lyrics of the different popular music artist. The low values are the result of the short lines that are common in popular music lyrics and limit the complexity of sentences that can fit the tune. However, it provides a comparative scale by which we can rank and compare the different musicians.

Figure 1

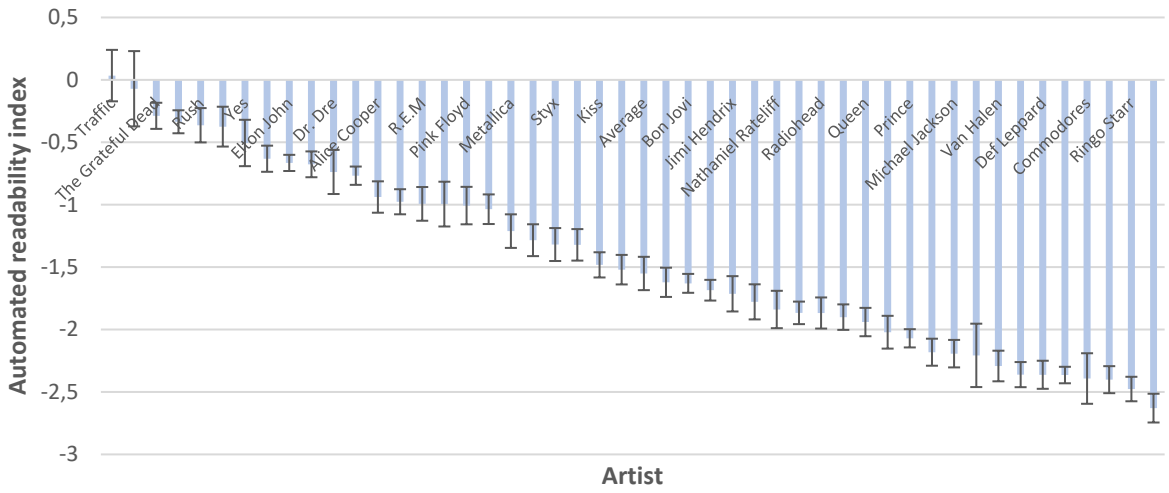
The Average Coleman-Liau Readability Index of The Lyrics of The Different Artists.



Note: The values are the mean of the readability index of the songs of each artist, and the error bars show the standard error of the mean.

Figure 2

Average Automated Readability Index (ARI) of The Lyrics of Different Artists.



The figures show that musician with higher readability index include Bob Dylan, Alanis Morrissette, Elton John (Bernard Taupin), Bruce Springsteen, and the bands Yes, Rush, and Traffic. Hip-hop artists such as Gill Scott Heron and the Beastie Boys are ranked with a high readability index, as well as the rapper Dr. Dre (Andre Romelle Young). Hip-hop and rap are musical styles that allow longer sentences, and consequently more complex lyrics. It has been proposed that rap allows to express deeper ideas regarding social issues or personal experiences (Edwards, 2002; Mise, 2020; Williams, 2020). The Beastie Boys often described fictional stories in their lyrics, making their lyrical style more complex (Hess, 2005). The rock band *Traffic*, which has a relatively high readability index, has a unique musical style the features long improvisation-like musical pieces, and could also allow to express more complex lyrics. The uniqueness of *Traffic* lyrics is also shown in several other unrelated measurements discussed later in this section.

Artists with lower readability index include Guns N' Roses, U2, Van Halen, and Ringo Starr. Like Traffic lyrics being characterized by different measurements that indicate on complexity of the lyrics, Starr's lyrics showed several unrelated measurements that indicate on simplicity. These measurements will be discussed later in this section. Among the popular music artists whose lyrics are the easiest to read is also Neil Young, who is considered a notable lyricist (Echard, 2005; Tomiyama, 2017). However, Neil Young's lyrics style is also unique in its simplicity. For instance, Echard (2005) argued that Young's songs were "seemingly simple", and that "Neil Young's lyrics have always been deceptively simple. It is this simple complexity which make Young's lyrics so intriguing" (Thrasher's Wheat, 2004).

Differences in sentiments expressed in lyrics

Figure 3 shows the average sentiment expressed in the lyrics of different artists. As expected, different artists express different positivity in their lyrics. Artists such as Ringo Starr, Van Halen, George Harrison, and Yes express, on average, more positive

sentiments in their lyrics. The artists that express more negative lyrics include Hip-Hop artists such as Gil Scott-Haron and Dr. Dre (Andre Romelle Young) as well as bands in the Rock genres such as Crosby, Stills, Nash & Young, Motorhead, and Bon Jovi.

It is interesting that some of the artists with the lower readability index are also some of the artists that their lyrics express the most positive sentiments. The correlation between the readability of the lyrics and the sentiment expressed in the lyrics was tested by correlating the readability index and the sentiment expressed in the 18,557 songs. That was done by using the Pearson correlation, which was 0.21. With 18,557 songs, the probability to have such correlation by chance is lower than 0.00001. That provides quantitative evidence that songs that are less positive also tend to use longer words and sentences, making them somewhat more difficult to read.

Figure 3

Average Sentiment Expressed in The Lyrics of The Different Artists

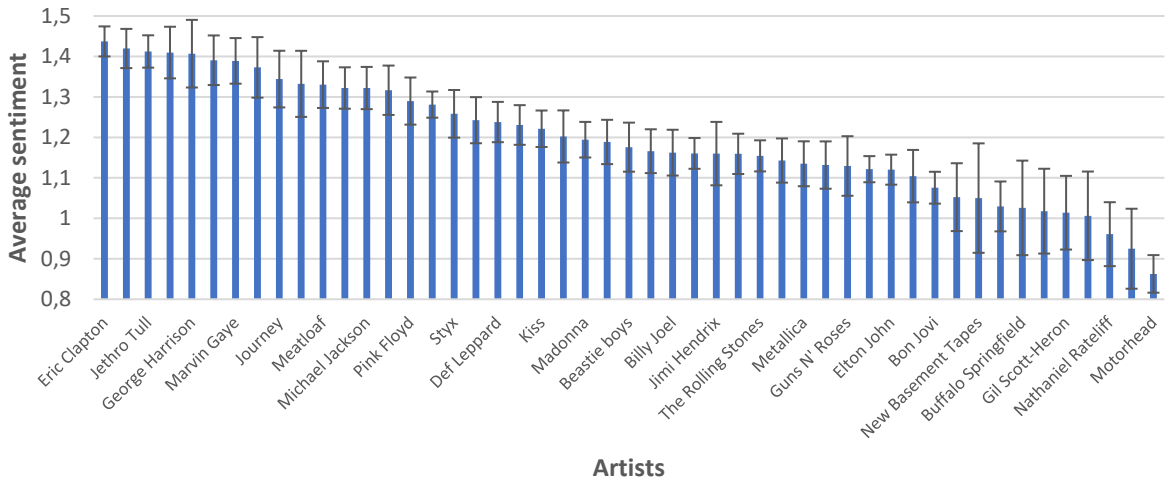
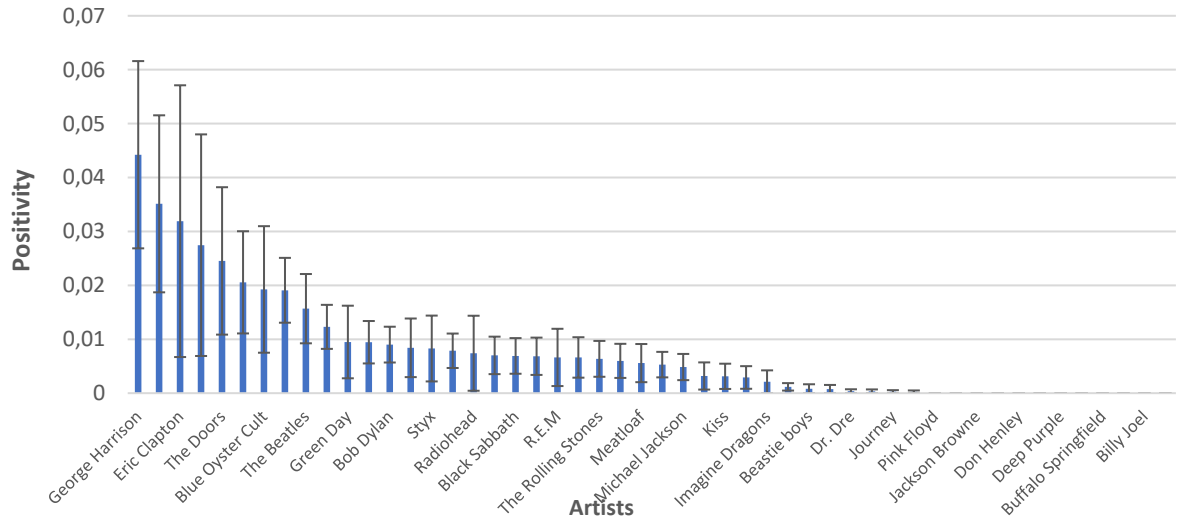


Figure 3 shows the average positivity expressed in the lyrics. That score might not provide full information of the positivity expressed in the lyrics, as very positive sentences might be offset by very negative sentences in other lyrics by the same artist. Figure 4 shows the frequency of sentences annotated as “very positive” by the algorithm. The figure shows that the artists with the highest frequency of very positive sentences is George Harrison, with about 4% of his sentences express very positive sentiments. Other artists that express positivity frequently in their lyrics are John Lennon, Eric Clapton, Commodores, The Doors, and Queen. John Lennon songs are limited to 1970 and later, and his work as a member of the Beatles is excluded. Billy Joel, Boston, and Pink Floyd, on the other hand, rarely express very positive sentiments in their lyrics.

Figure 4

The Frequency of Very Positive Sentences in The Lyrics of Different Artists



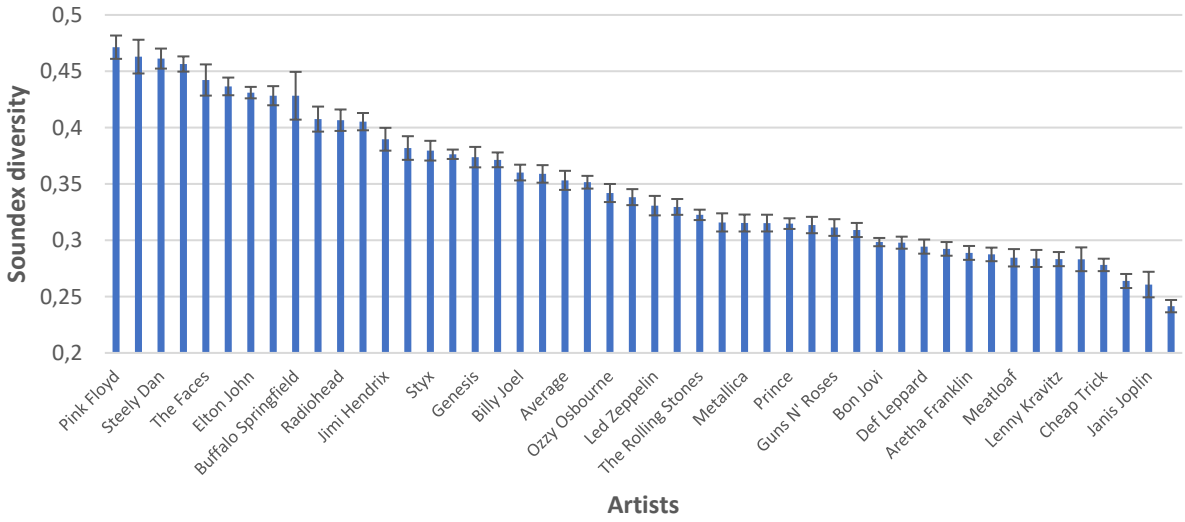
The frequency of very positive sentences is not necessarily a reflection of the overall positivity of the lyrics, as very positive sentences and negative sentences can be used by the same artists and even in the same song. The data shows that while different artists express very positive sentences differently, that cannot be associated with a certain genre or lyrics style. For instance, Pink Floyd and Billy Joel tended to express a darker lyrics style and often focused on political and social topics in their songs (Rozinski, 2015; Salkin & Crisci, 2015). The analysis shows that very positive sentences are infrequent in their work. On the other hand, the post-Beatles John Lennon also tended to focus in his lyrics on political and social issues, while doing that by expressing a much higher number of very positive sentences. That could be linked to Lennon's practice of "sugarcoating" the political messages in his lyrics (Hewett, 2016). For instance, Lennon was quoted for "Now I understand what you have to do. Put your political message across with a little honey" (Fricke, 2001).

Differences in the use of sounds

Figure 5 shows the distribution of sounds in the lyrics using the Soundex algorithm. The graph shows that artists such as Imagine Dragons, Janice Joplin, Bon Jovi, and Ringo Starr tend to use in their lyrics the same sounds repetitively, while other artists such as Pink Floyd chose a more diverse set of sounds of words in their lyrics.

Figure 5

The Distribution of Sounds in The Lyrics of Different Artists by Using the Soundex Algorithm



Note: The error bars show the standard error of the mean of the Soundex sound diversity.

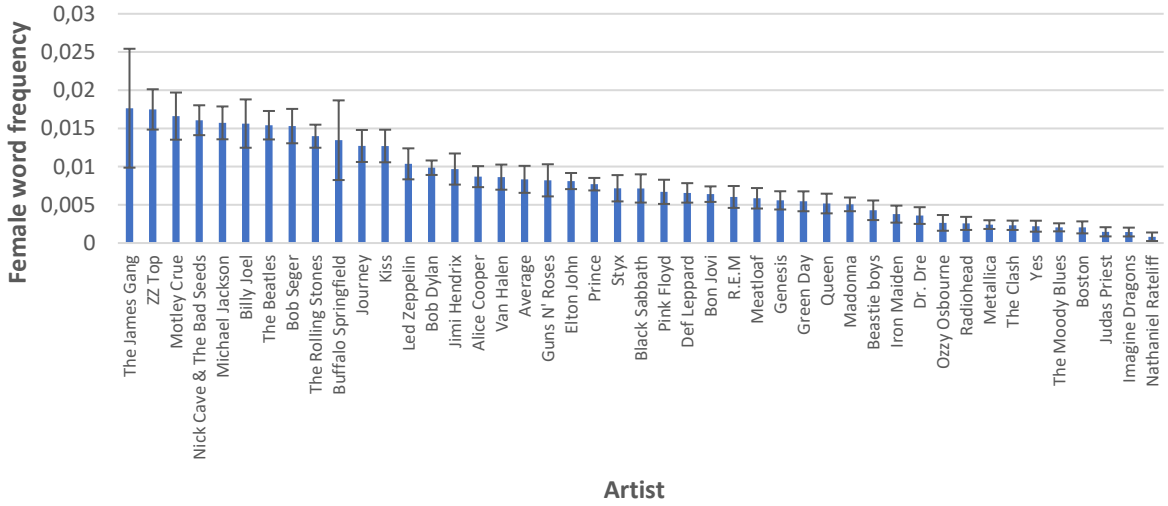
The diversity of sounds in Steely Dan lyrics can be associated with the different thematic tones used as a dominant aspect in the lyrics (Alper, 2022; Borshuk, 2021). Traffic is a rock band that made use of long jazz-like tracks converted into rock-style music (Ray, 2013; Strong, 2000), and the unique style could be also expressed in the selection of sounds of the lyrics to fit that style. That jazz-rock style of long improvisation-like tracks is somewhat common to the style of Steely Dan, Pink Floyd, and Jethro Tull. These rock bands, active mostly during the 1970s and 1980s, all share that music style, and that style is different than most other popular music artists.

Expression of gender identity

Gender expressed in popular music has been a topic of interest in popular music research (Cohen, 2001; Werner, 2012; Whiteley, 2013). For instance, Flynn et al. (2016) showed differences in objectification of men and women that are also sensitive to the era and genre. It is therefore expected that gender identity could be an element that is expressed differently in lyrics of different musicians, and consequently expressed differently in a computational analysis of gender-related terms. Figure 6 shows the frequency in which terms related to women identity are mentioned in lyrics. Artists that mention terms related to women more often are ZZ Top, Billy Joel, Michael Jackson, and The Beatles. On the other end of the graph, bands such as Radiohead, Yes, Boston, or Imagine Dragons use such terms far less frequently.

Figure 6

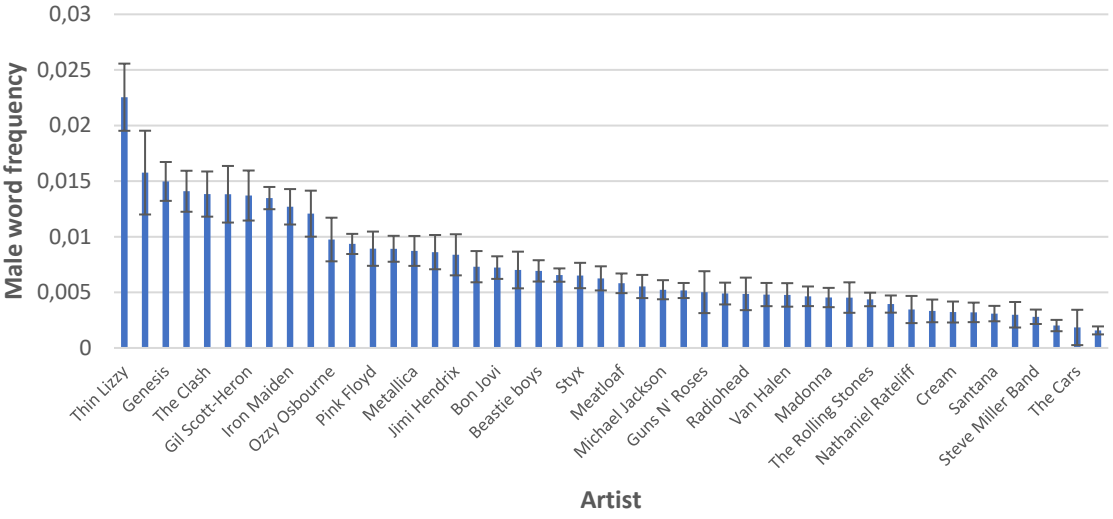
The Frequency of Woman-Related Terms in Lyrics of Different Popular Music Artists



Similarly, Figure 7 shows the frequency of terms related to man identity in the lyrics. The frequency of words related to man identity in the entire song dataset is 0.007, somewhat less frequent than words related to woman identity, with overall frequency of 0.008. Artists with the most frequent use of man identity words include Dire Straits, Genesis, Bob Dylan, and Nick Cave. The Cars, Imagine Dragons, and Peter Frampton use man identity words less frequently compared to most other artists in the dataset. The correlation between the frequency of man-related words and woman-related words is 0.045.

Figure 7

Average Frequency of Terms Related to Men in Different Artists



The gender-based use of gender-related terms changes between different artists. For instance, Imagine Dragons use gender-related terms infrequently regardless of whether the terms are related to men or women. On the other hand, The Clash use men

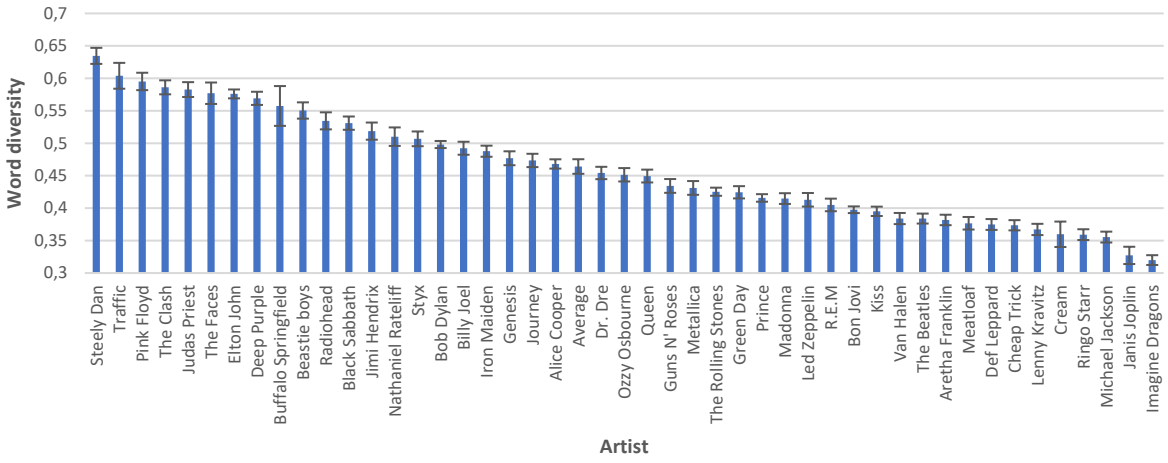
related terms in high frequency, while women-related terms are far less frequent in their lyrics. For the Rolling Stones and the Beatles, however, women terms are more frequent than men-related terms. Both bands have been focusing in their lyrics on topics related to love and romance. For instance, it was only until their sixth studio album that the Beatles included their first notable song that did not focus on love and romance - "Nowhere man". The graphs do not show a clear difference between different genres, and musicians within the same genres show substantial differences in the way they use gender-related terms.

Diversity in the use of words

Figure 8 shows how different artists use the same words repetitively in the lyrics. The graph shows that popular music artists differ from each other in the way they tend to repeat the same word multiple times in the same song. The lyrics of Elton John and Jimi Hendrix, or bands such as Deep Purple and Pink Floyd do not tend to repeat the same words in their lyrics. Less than 50% of the words that these artists use in lyrics are repeated. Imagine Dragons and Ringo Starr, on the other hand, tend to repeat the same words in their lyrics much more often.

Figure 8

The Average Word Diversity in The Lyrics of Different Popular Music Artists



The graph shows substantial differences between the degree of word repetition. The artist with the most diverse lyrics as reflected by the use of different words is Steely Dan. The lyrics of Steely Dan has attracted a relatively significant interest in the music research community (Alper, 2022; Borshuk, 2021; Clements, 2009; Everett, 2004). Specifically, it has been suggested that the lyrics of Steely Dan represent different thematic tones while expressing human nature stories as cautionary tales (Alper, 2022). Another band with diverse use of words is The Clash, which as a punk band in the late 1970s was also recognized for expressing a variety of topics and social matters (Bindas, 1993; Setiawan, 2013). The rock band Traffic is also shown to use a diverse set of words. As Figure 1 show, the band's lyrics has higher readability index compared to other artists. These are two measurements are not mathematically related, but their combination can be

viewed as a reflection of a more complex lyrics style. The band's work has not attracted substantial interest from the academic community, and future qualitative work will be required to fully understand the lyrics of that band, mostly active in the *rock* genre.

On the other end of that graph, Ringo Starr is characterized by low diversity of words and low readability index. That combination can be an indication of a simple approach in Starr's lyrics style. With Starr's positive sentiments shown in Figure 3, Starr's lyrics can be viewed as simple and joyful, rather than an attempt to communicate more complex ideas in the lyrics or provoke the thoughts of the listener. Starr's songs used here exclude his work while he was a member of the Beatles. While Starr's is credited for merely two songs during his Beatles era ("Octopus's Garden" and "Don't pass me by"), he contributed partial lyrics to other Beatles songs. His unique style as a lyricist was coined by his band members as "Ringoism" (Hobson, 2021). Preeminent songwriters such as Bob Dylan also have high diversity in the words they use, although their words diversity is not exceptionally high, indicating the diversity of words is not necessarily a single indication of high impact of the lyrics on society.

CONCLUSION

The availability of digital platforms allows quantitative analysis that was impractical to perform in the pre-information era. For instance, even a simple task such as counting the number of words or the frequency of terms that appear in popular music lyrics is a daunting task that requires substantial labor. Computer analysis allows to quantify these text elements in a large number of songs, enabling a new approach to the studying of popular music. Here we extract several intuitive text measurements from popular music lyrics. The fact that a machine learning classifier can identify the artists by the lyric's elements in accuracy much higher than mere chance shows that popular music artists have a certain style in their lyrics reflected by the collection of quantitative measurements used here.

Machine learning classifiers that can associate lyrics with their creating artist are not new. However, identifying the artist automatically does not necessarily lead to new knowledge about popular music. Machine learning-based document classifiers tend to work by complex data-driven rules that act as a "black box", and therefore often do not provide the user with substantial new knowledge that the user does not already know. The study here is limited to intuitive explainable measurements of the lyrics, and therefore can be used to profile the differences and identify specific quantitative features that can reflect differences between songwriters. Because the text elements are explainable and intuitive, the analysis is interpretable, and can therefore lead to new knowledge about popular music. Such differences can provide ques for further studying of the uniqueness of lyricists or identifying similarities and influential links between popular music songwriters. The analysis shown here identified several differences between songwriters and demonstrates that songwriters are different from each other in a manner that can be quantified and measured. Lyrics elements expressed by songwriters, whether intentional or subconscious, can be identified and quantified by

applying computer analysis. That approach can expand the set of tools that can be used for analyzing popular music.

The results of the computer analysis led to several observations regarding the songwriters. For instance, several mathematically unrelated measurements of Ringo Starr's lyrics show a positive and simple lyrical style. Early Hip-hop artists such as Gil Scott-Heron or the Beastie Boys, as well as rock bands that featured improvisation-like musical pieces tended to express more complex lyrics. The analysis also showed that artists with lower readability index also tend to express more positivity in their songs. When applying the analysis to all songs in the dataset, the correlation between the positivity expressed in the lyrics and the readability index showed that lyrics that express positive sentiments are also easier to read, while less positive songs also have more complex lyrics. Although the link between sad songs and the expression of complex has been proposed (Mori & Iwanaga, 2014), that analysis can be difficult to apply to a large number of songs without using computer analysis.

The quantitative approach used in this study can be used for more detailed analysis, involving different artists and different research questions. Therefore, the Udat software developed for this study is available for free download (Shamir, 2017), and can be used by the research community for future studies that involve quantitative analysis of popular music lyrics. While the topic shown in this paper is studying differences between artists, the software and analysis can be applied to a variety of other research questions related to the analysis of popular music lyrics.

Funding

This research was supported in part by NSF grant AST-1903823.

REFERENCES

Alper, G. (2022). Not just derision and darkness: The interplay of lyrics and music in Steely Dan's compositions. *Rock Music Studies*. https://doi.org/10.1080/19401159.2022.2008161

An, Y., Sun, S., & Wang, S. (2017). Naive Bayes classifiers for music emotion classification based on lyrics. *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, 635–638. https://doi.org/10.1109/ICIS.2017.7960070

Ballard, M. E., Dodson, A. R., & Bazzini, D. G. (1999). Genre of music and lyrical content: Expectation effects. *The Journal of Genetic Psychology*, *160*(4), 476–487. https://doi.org/10.1080/00221329909595560

Bindas, K. J. (1993). The future is unwritten: The Clash, punk and America, 1977-1982. *American Studies*, *34*(1), 69–89. https://journals.ku.edu/amsj/article/view/2851

Bindas, K. J., & Houston, C. (1989). "Takin' care of business": Rock music, Vietnam and the protest myth. *The Historian*, *52*(1), 1–23. https://doi.org/10.1111/j.1540-6563.1989.tb00771.x

Borshuk, M. (2021). "Steely Dan at 50." *Rock Music Studies*. https://doi.org/10.1080/19401159.2022.2008165

Clements, P. (2009). Cultural legitimacy or 'outsider hip'? Representational ambiguity and the significance of Steely Dan. *Leisure Studies*, *28*(2), 189–206. https://doi.org/10.1080/02614360902769886

Cohen, S. (2001). Popular music, gender and sexuality. In J. Street, S. Frith, & W. Straw (Eds.), *The Cambridge Companion to Pop and Rock* (pp. 226–242). Cambridge University Press. https://doi.org/10.1017/CCOL9780521553698.014

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284. https://doi.org/10.1037/h0076540

Condit-Schultz, N., & Huron, D. (2015). Catching the lyrics: Intelligibility in twelve song genres. *Music Perception: An Interdisciplinary Journal*, *32*(5), 470–483. https://doi.org/10.1525/mp.2015.32.5.470

Davies, P. (1990). "There's no success like failure": From rags to riches in the lyrics of Bob Dylan. *The Yearbook of English Studies*, *20*, 162–181. https://doi.org/10.2307/3507528

de Boise, S. (2020). Music and misogyny: A content analysis of misogynistic, antifeminist forums. *Popular Music*, *39*(3–4), 459–481. https://doi.org/10.1017/S0261143020000410

Dunlap, J. (2006). Through the eyes of Tom Joad: Patterns of American Idealism, Bob Dylan, and the Folk Protest Movement. *Popular Music and Society*, *29*(5), 549–573. https://doi.org/10.1080/03007760500238510

Echard, W. (2005). *Neil Young and the poetics of energy*. Indiana University Press.

Edwards, W. (2002). From poetry to rap: The lyrics of Tupac Shakur. *Western Journal of Black Studies*, *262*, 61–70. https://www.vonsteuben.org/ourpages/auto/2016/2/24-/51380098/PoetrytoRapTupac.pdf

Everett, W. (2004). A royal scam: The abstruse and ironic bop-rock harmony of Steely Dan. *Music Theory Spectrum*, *26*(2), 201–236. https://doi.org/10.1525/mts.2004.26.2.201

Fell, M., & Sporleder, C. (2014). Lyrics-based analysis and classification of music. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 620–631. https://aclanthology.org/C14-1059

Flynn, M. A., Craig, C. M., & Anderson, C. N. (2016). Objectification in popular music lyrics: An examination of gender and genre differences. *Sex Roles*, *75*, 164–176. https://doi.org/10.1007/s11199-016-0592-3

Fox, W. S., & Williams, J. D. (1974). Political Orientation and Music Preferences Among College Students. *Public Opinion Quarterly*, *38*(3), 352–371. https://doi.org/10.1086/268171

Freudiger, P., & Almquist, E. M. (1978). Male and female roles in the lyrics of three genres of contemporary music. *Sex Roles*, *4*, 51–65. https://doi.org/10.1007/BF00288376

Fricke, D. (2001, December 27). "Imagine": The anthem of 2001. *Rolling Stone*. https://www.rollingstone.com/music/music-news/imagine-the-anthem-of-2001-83559/

Gosa, T. L. (2017). Hip hop, authenticity, and styleshifting in the 2016 presidential election. *Journal of Popular Music Studies*, *29*(3), e12236. https://doi.org/10.1111/jpms.12236

Hess, M. (2005). Hip-hop realness and the white performer. *Critical Studies in Media Communication*, *22*(5), 372–389. https://doi.org/10.1080/07393180500342878

Hewett, M. R. (2016). Two linguistic case studies of the craft of songwriting: "Imagine" and "Like a Rolling Stone." *Lingua Frankly*, *3*. https://doi.org/10.6017/lf.v3i0.9345

Hobson, J. (2021). A hard day's night. *Occupational Medicine*, *71*(9), 398–400. https://doi.org/10.1093/occmed/kqaa170

Kresovich, A., Reffner Collins, M. K., Riffe, D., & Carpentier, F. R. D. (2021). A content analysis of mental health discourse in popular rap music. *JAMA Pediatrics*, *175*(3), 286–292. https://doi.org/10.1001/jamapediatrics.2020.5155

Kutschke, B. (2016). Political music and protest song. In K. Fahlenbrach, M. Klimke, & J. Scharloth (Eds.), *Protest Cultures* (1st ed., pp. 264–272). Berghahn Books; https://doi.org/10.2307/j.ctvgs0b1r.33

Lammer, J. (2016). *The impact of Bob Dylan on the Beatles* [Universität Graz]. http://unipub.uni-graz.at/obvugrhs/1356263

Logan, B., Kositsky, A., & Moreno, P. (2004). Semantic analysis of song lyrics. *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, 827–830. https://doi.org/10.1109/ICME.2004.1394328

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, 55–60. https://doi.org/10.3115/v1/P14-5010

Martin, P. J. (2006). Musicians' worlds: Music-making as a collaborative activity. *Symbolic Interaction*, *29*(1), 95–107. https://doi.org/10.1525/si.2006.29.1.95

Mise, U. (2020). Rap music as resistance and its limits, two diverging cases: Sulukule and Bağcılar rap. *Anthropology of East Europe Review*, *37*(1), 27–51. https://scholarworks.iu.edu/journals/index.php/aeer/article/view/28763

Mori, K., & Iwanaga, M. (2014). Pleasure generated by sadness: Effect of sad lyrics on the emotions induced by happy music. *Psychology of Music*, *42*(5), 643–652. https://doi.org/10.1177/0305735613483667

Napier, K., & Shamir, L. (2018). Quantitative sentiment analysis of lyrics in popular music. *Journal of Popular Music Studies*, *30*(4), 161–176. https://doi.org/10.1525/jpms.2018.300411

Nielson, E. (2009). "My president is black, my lambo's blue": The Obamafication of rap? *Journal of Popular Music Studies*, *21*(4), 344–363. https://doi.org/10.1111/j.1533-1598.2009.01207.x

North, A. C., Krause, A. E., & Ritchie, D. (2021). The relationship between pop music and lyrics: A computerized content analysis of the United Kingdom's weekly top five singles, 1999–2013. *Psychology of Music*, *49*(4), 735–758. https://doi.org/10.1177/0305735619896409

Odell, M. K. (1956). The profit in records management. *System Magazine (New York)*, *20*, 20.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D. M., & Goldberg, I. G. (2008). WND-CHARM: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, *29*(11), 1684–1693. https://doi.org/10.1016/j.patrec.2008.04.013

Ortega, J. L. (2021). Cover versions as an impact indicator in popular music: A quantitative network analysis. *PLOS ONE*, *16*(4), e0250212. https://doi.org/10.1371/journal.pone.0250212

Petrie, K. J., Pennebaker, J. W., & Sivertsen, B. (2008). Things we said today: A linguistic analysis of the Beatles. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(4), 197–202. https://doi.org/10.1037/a0013117

Ray, M. (2013). *Disco, punk, new wave, heavy metal, and more: Music in the 1970s and 1980s*. Britannica Educational Pub.: in association with Rosen Educational Services. http://site.ebrary.com/id/10627012

Richardson, J. E. (2017). Recontextualization and fascist music. In L. C. S. Way & S. McKerrell (Eds.), *Music as multimodal discourse: Semiotics, power and protest*. Bloomsbury Publishing.

Rozinski, T. (2015). Using music and lyrics to teach political theory. *PS: Political Science & Politics*, *48*(3), 483–487. https://doi.org/10.1017/S1049096515000293

Ruth, N. (2019). "Where is the love?" Topics and prosocial behavior in German popular music lyrics from 1954 to 2014. *Musicae Scientiae*, *23*(4), 508–524. https://doi.org/10.1177/1029864918763480

Salkin, P., & Crisci, I. (2015). Billy Joel: The chronicler of the suburbanization in New York. *Touro Law Review*, *32*(1), 111–138. https://digitalcommons.tourolaw.edu/lawreview/vol32/iss1/8

Setiawan, A. (2013). Analysis on anti capitalism in the "Clampdown" lyric by The Clash. *LANTERN (Journal on English Language, Culture and Literature)*, *2*(2), 35–45. https://ejournal3.undip.ac.id/index.php/engliterature/article/view/2396

Shamir, L. (2017). UDAT: A multi-purpose data analysis tool. *Astrophysics Source Code Library*, ascl:1704.002. https://ui.adsabs.harvard.edu/abs/2017ascl.soft04002S

Shamir, L. (2021). UDAT: Compound quantitative analysis of text using machine learning. *Digital Scholarship in the Humanities*, *36*(1), 187–208. https://doi.org/10.1093/llc/fqaa007

Shamir, L., Macura, T., Orlov, N., Eckley, D. M., & Goldberg, I. G. (2010). Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, *7*(2), 1–17. https://doi.org/10.1145/1670671.1670672

Shamir, L., Orlov, N., Eckley, D. M., Macura, T., Johnston, J., & Goldberg, I. G. (2008). Wndchrm – an open source utility for biological image analysis. *Source Code for Biology and Medicine*, *3*(1), 13. https://doi.org/10.1186/1751-0473-3-13

Smith, E. A., & Senter, R. J. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (U.S.)*, 1–14.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Strong, M. C. (2000). *The great rock discography*. Mojo Books.

Thrasher's Wheat. (2004, April 14). *Neil Young lyric analysis*. http://thrasherswheat.org/fot.htm

Tomiyama, H. (2017). Neil Young: Some complexities in his songs. In T. Connolly & T. Iino (Eds.), *Canadian Music and American Culture: Get Away From Me* (pp. 61–76). Springer International Publishing. https://doi.org/10.1007/978-3-319-50023-2_4

Tsaptsinos, A. (2017). *Lyrics-based music genre classification using a hierarchical attention network* (arXiv:1707.04678). arXiv. https://doi.org/10.48550/arXiv.1707.04678

Vandagriff, R. S. (2015). Talking about a Revolution: Protest Music and Popular Culture, from Selma, Alabama, to Ferguson, Missouri. *Lied Und Populäre Kultur / Song and Popular Culture*, *60/61*, 333–350. https://www.jstor.org/stable/26538872

Werner, V. (2012). Love is all around: A corpus-based study of pop lyrics. *Corpora*, *7*(1), 19–50. https://doi.org/10.3366/cor.2012.0016

West, A., & Martindale, C. (1996). Creative trends in the content of beatles lyrics. *Popular Music and Society*, *20*(4), 103–125. https://doi.org/10.1080/03007769608591647

Whiteley, S. M. (2013). Popular music, gender and sexualities. *IASPM Journal*, *3*(2), 78–85. https://doi.org/10.5429/604

Williams, M. L. (2020). "Meditate, don't medicate!" An analysis of addict rap, black men's social issues, and J. Cole's K.O.D. album. *Howard Journal of Communications*, *31*(5), 415–428. https://doi.org/10.1080/10646175.2020.1717679

Yang, Y. (2020). "Musicalization of the culture": Is music becoming louder, more repetitive, monotonous and simpler? *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 750–761. https://ojs.aaai.org/index.php/ICWSM/article/view/7340

Yeh, C.-H., Tseng, W.-Y., Chen, C.-Y., Lin, Y.-D., Tsai, Y.-R., Bi, H.-I., Lin, Y.-C., & Lin, H.-Y. (2014). Popular music representation: Chorus detection & emotion recognition. *Multimedia Tools and Applications*, *73*(3), 2103–2128. https://doi.org/10.1007/s11042-013-1687-2

Yoo, Y., Ju, Y., & Sohn, S. Y. (2017). Quantitative analysis of a half-century of K-Pop songs: Association rule analysis of lyrics and social network analysis of singers and composers. *Journal of Popular Music Studies*, *29*(3), e12225. https://doi.org/10.1111/jpms.12225

Appendix

Table 1

Artists in The Dataset, Number of Songs, And Range of Years During Which Songs of Each Artist Were Released

Artist	# Songs	Start Year	End Year
Alanis Morissette	126	1991	2020
Alice Cooper	330	1969	2019
Aretha Franklin	310	1961	2011
Bad Company	127	1974	1996
Beastie boys	121	1986	2011
Billy Idol	111	1981	2014
Billy Joel	129	1971	1993
Black Sabbath	177	1970	2016
Blue Oyster Cult	123	1972	2001
Bob Dylan	484	1962	2017
Bob Seger	210	1969	2017
Bon Jovi	286	1984	2020
Bruce Springsteen	354	1973	2019
Cheap Trick	229	1977	2017
Chicago	275	1969	2019
David Bowie	349	1967	2017
Deep Purple	189	1968	2020
Def Leppard	179	1980	2015
Electric Light Orchestra	189	1971	2019

Artist	# Songs	Start Year	End Year
Elton John	419	1969	2016
Eric Clapton	357	1970	2018
Fleetwood Mac	255	1968	2013
Genesis	180	1969	2000
George Harrison	130	1970	2002
Green Day	159	1990	2020
Guns N' Roses	101	1987	2008
Heart	170	1976	2016
Imagine Dragons	106	2008	2018
Iron Maiden	183	1980	2015
Jackson Browne	189	1967	2014
Jeff Beck	120	1968	2016
Jethro Tull	253	1968	2003
Jimi Hendrix	103	1967	2013
John Fogerty	121	1973	2013
John Lennon	121	1970	1980
John Mellencamp	258	1976	2018
Journey	170	1975	2011
Judas Priest	223	1974	2018
Kiss	221	1974	2012
Lenny Kravitz	144	1989	2018
Lou Reed	243	1972	2011
Lynyrd Skynyrd	174	1974	2012

Artist	# Songs	Start Year	End Year
Madonna	265	1983	2019
Marvin Gaye	243	1961	2019
Meatloaf	156	1977	2016
Metallica	146	1983	2016
Michael Jackson	217	1972	2014
Motley Crue	156	1981	2008
Motorhead	266	1977	2015
Neil Young	576	1969	2019
Nick Cave & The Bad Seeds	191	1984	2019
Oasis	117	1994	2008
Ozzy Osbourne	153	1980	2020
Paul McCartney	343	1970	2018
Peter Frampton	126	1972	2019
Phil Collins	119	1981	2019
Pink Floyd	139	1967	2015
Prince	507	1978	2019
Queen	186	1973	2008
R.E.M	192	1983	2008
Radiohead	144	1993	2016
Ramones	191	1976	1994
Ringo Starr	248	1970	2019
Robert Plant	149	1982	2017

Artist	# Songs	Start Year	End Year
Rod Stewart	398	1969	2018
Rush	170	1974	2012
Sammy Hagar	181	1976	2019
Santana	211	1969	2019
Scorpions	222	1972	2015
Steve Miller Band	172	1968	2011
Stevie Nicks	128	1981	2014
Stevie Wonder	275	1962	2005
Styx	173	1972	2017
Ted Nugent	111	1974	2018
The Beatles	295	1963	1996
The Clash	130	1977	1993
The Doobie Brothers	132	1971	2010
The Doors	109	1967	1978
The Grateful Dead	167	1967	1989
The Kinks	402	1964	1993
The Moody Blues	209	1965	2003
The Rolling Stones	360	1964	2016
The Who	195	1965	2019
Thin Lizzy	127	1971	1983
Tom Petty and the Heart Breakers	201	1976	2014
U2 Band	224	1980	2017

Artist	# Songs	Start Year	End Year
Van Halen	119	1978	2012
Yes	172	1969	2014
ZZ Top	166	1971	2012

Source: AZLyrics.com, authors' estimation.

This page intentional left blank.