Applications of Natural Language Processing to Compare Shakespearean Sonnets to Modern Musical Artists

Mariana Gonzalez Castro*, Carina Kalaydjian*, Dominique McDonald*, Angel Sierra*, and DuoDuo Ying*

*Department of Statistics, UCLA

June 8th, 2022

Abstract

We seek to detect artists that are truly outstanding and stand the trial of time. In this paper, we introduced a metric to predict 45 artists' success based on the similarity between a subset of their discography and Shakespeare's 154 sonnets. We used Natural Language Processing, specially Keyword Extraction and Sentiment Analysis, to measure such similarities. For Keyword Extraction, we measured the number of similar top 10 keywords used between artists and Shakespeare and the Euclidean distance of the frequencies. For Sentiment Analysis, we performed a cluster analysis and measured the Euclidean distance for 8 emotions within the text of artists and Shakespeare. We combined the result for the two methodologies and their rankings to obtain an overall ranking. Out of 45 top artists, Nickelback, Amy Winehouse, and Cake are the most similar artists to Shakespeare using the combined ranking. We conclude that by leveraging Natural Language Processing algorithms, we can predict future success based on the proven success of the past.

1 Introduction

The music industry has been growing in popularity over the last few decades. In 2021, worldwide recorded music revenues totaled $25.9 billion, up 18.5 percent from 2020. [5]There are more and more artists representing different genres, styles, and even subgenres within genres. As a prominent music label, we wonder how to detect artists that are outstanding and stand the trial of time? Who are the artists that are not only relevant now but tens and hundreds of years from now?

In this paper, we propose a method for measuring artists' timelessness using their similarities with Shakespeare. William Shakespeare is a world-renowned poet from 16th-century England. He has recognition for being one of the most revolutionary Literary Artists to this date. Not only has Shakespeare close to one-tenth of the most quoted lines ever written or spoken in English, second-most quoted after the writers of the bible,

but his ability to deeply express emotions also has influenced for hundreds of years.[1] We assume that, given William Shakespeare's success, if modern artists' work is similar to that of William Shakespeare's in terms of word choice and sentiment, then they are likely to be successful as well.

For our data, we use Shakespeare's 154 sonnets from the Gutenberg Project. [13] We obtained a subset of 45 top modern artists' discography through a Kaggle competition. We will use Natural Language Processing, specifically Keyword Extraction and Sentiment Analysis, to measure similarities between the work of 45 artists and 154 sonnets of Shakespeare. The rest of the paper is structured as follows: Section 2 introduces the methodology, both the front-end analysis and back-end data management. Section 3 provides results and plots from our analysis. Section 4 concludes. Appendix can be found in Section 5.

2 Methods

There is a growing demand for the application of Natural Language Processing to drive music knowledge discovery [10]. Within song lyrics there is a wealth of information that can be used to gain insights about the song and its listeners. The workflow used to conduct NLP is outlined in Oramas et. Al [12] as the following steps

1. Corpus creation (collection of separated documents),
2. Text mining (accessing desired info and eliminating the excess),
3. Information extraction (word frequency, collocation, word position, etc.),
4. Knowledge graph generation (a directed labeled graph in which we have associated domain specific meanings with nodes and edges [2]),
5. Sentiment Analysis (Identify feelings and emotions present in a text [8]).

This NLP pipeline can be applied to music to provide recommendations on what songs a listener might like based on what they already listen to. [10] This information is gathered by performing what's called a similarity search. Mahedero et al [10] conducts a similarity search for their project by first calculating a similarity measure. The similarity measure used is known as the Standard Cosine Distance (SCD). The calculations of the SCD are beyond the scope of this paper, but something to note is that the SCD relies on the Inverse Document Frequency as a way to measure the prevalence of words in a document and compare that to other documents.

As previously stated, we relied on Natural Language Processing methods to calculate the similarity between Shakespeare and current artists with the hope of identifying a single artist whose work has similar themes as Shakespeare's. Our pipeline was structured similarly to that proposed in [12] with the main difference being that we do not rely on knowledge graph generation to store any information or display findings. The methods that were most useful for this research were keyword extraction and sentiment analysis. Both methods of analysis provided us with crucial insights, so the details of each seem pertinent to share.

2.1 Keyword Extraction

Keyword extraction is largely aimed at identifying the most relevant words in different texts and utilizing those words to understand common theme or popular topics.[8] We used the tm library in R to perform keyword extraction[7] on the sonnets. This process involves first prepping the text by removing unwanted punctuation or numbers, eliminating stop words, changing everything to lower case. These transformations are necessary because when working with strings, you not only have to be precise, but you also must be exact. The cleaning process sometimes includes stemming the words, but we opted not to do that. Once the text is clean the idea is to create a table containing each word used in the text and the frequency with which it is used. To identify the top ten key words we sorted the words in the matrix by their frequency.

Figure 1: Shakespeare Top Keywords Wordcloud

Once Shakespeare's keywords were identified, the next step was to identify the keywords for each of the 45 music artists. This process mirrored that of extracting keywords from the sonnets except we utilized an algorithm to automate the process for each artist. Once keywords were identified we converted the word frequencies for Shakespeare and all other artists into proportions, in order to standardize for comparison. To find the artists who were most like Shakespeare based on keywords, we checked for artist who had the highest number of matching keywords. Finally, then we ranked those by who used keywords in similar proportions to Shakespeare by calculating the Euclidian distance between the proportions of each of the matching keywords. This is similar to what Mahedero et. Al [10] does for their similarity search except we rely on word proportions and Euclidean distance, while their research utilizes Inverse Document Frequency and the Standard Cosine Distances.

2.2 Sentiment Analysis

A similar sort of ranking was achieved from Sentiment Analysis, but the process has notable differences. The library syuzhet was used for the analysis. There is a wide variety of paths one could explore when performing sentiment analysis. Two that seemed viable for the purpose of this research were calculating an overall sentiment score and identifying the different emotions present in each text. [6] The latter option proved to be more fruitful, as the analysis is more detailed. This process of classification based on emotion is known as NRC sentiment analysis. NRC Sentiment Analysis uses the National Research Council (NRC) Word-Emotion Association Lexicon to classify words in a text into eight categories of emotions. [4] It is important to note that a word may be associated with more than one emotion. The eight emotions are anger, anticipation, disgust, fear, joy, sadness, surprise, trust (include simple equation). [4] The objective of NRC Sentiment Analysis is to calculate the frequency with which each emotion is conveyed. This is calculated by identifying the emotions associated with the unique words in a text and summing up all the instances of each emotion[6].

Once the NRC sentiments were calculated for both Shakespeare and the music artists, the frequencies were converted to proportions for accurate comparison. Using the proportions of each emotion, Euclidean distance was calculated for between each music artist and Shakespeare. The artists with the shortest distances from Shakespeare were considered the most similar to him, and therefore ranked higher in regard to comparison of the emotions conveyed in their works.

After comparing results from each analysis method, a final ranking was calculated. This overall ranking was calculated by summing up the rankings from both analysis methods. Because similarity to Shakespeare was assessed using Euclidean distance, lower rankings signify higher similarity to Shakespeare. This meant that the artists with the lowest overall ranking mirrored Shakespeare's work emotionally more than other artists.

2.3 K-Means Clustering

The NRC sentiments were utilized even further as predictors in K-Means Cluster Analysis. The objective of cluster analysis within the scope of this research is to utilize an unsupervised learning model [3] to assess commonalities between the work of each music artist and Shakespeare. K-Means clustering is an iterative method that categorizes each data point into one of k predefined groups, or clusters. The process is driven by two objectives. The first being maximizing the distance between clusters, so that they are distinct. The second is minimizing the data points within a cluster, so the clusters themselves are homogenous. [3] By employing K-means Clustering we were able to identify a group of artists that whose work most closely matches Shakespeare's.

2.4 Data Management

Before discussing the results of our various methods, it is important to address what made such an undertaking possible: data management. The size of this project necessitated multiple team members using multiple platforms. The data was stored in a relational database and then hosted on Amazon Web Services(AWS) service called Relational Database Services(RDS). Using RDS preserves the database from alterations, each

memeber each member provided the read only user credentials to their database to enable the team to access data without making changes to the database itself. Once we had the data it the next steps were to process and analyze it.

Again, with so many contributors working to advance the project it was necessary to have an avenue for efficient and organized sharing of code. For this aspect of the project, GitHub was utilized and it allowed team members to work on the same files from different locations and share them as frequently as necessary. Along with our code we are also able to store and share important information that aided us in our research. The different information sharing structures employed allowed for efficient progression and ultimately valuable results.

3 Results

Natural Learning Process requires the creation of a corpus to perform keyword extraction and sentiment analysis on the sonnets of Shakespeare. Separate corpora of 45 sampled musical artists were made to perform more keyword extraction and sentiment analysis for comparison to Shakespeare. Shakespeare's sonnets were collapsed such that a single corpus contained all 154 sonnets, each line representing a single line from a single sonnet. Each of the 45 artists had associated with a single corpus that contained multiple songs of their total discography, with each line representing a single line from a single song for the respective artist. With such corpora made, analysis of the text may begin.

Keyword extraction and sentiment analysis of Shakespeare's 154 sonnets against the discography of 45 sampled musical artists allow for an ordinal ranking of the artists in terms of similarity with Shakespeare. Amy Winehouse ranks most like Shakespeare considering a combined analysis. This can be interpreted as Amy Winehouse being most similar to Shakespeare in terms of keywords used by both persons and in terms of the eight measured emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) exhibited by their respective bodies of work. Shakespeare frequently writes about love, as seen in the keyword extraction of his sonnets. "Love" is also a keyword for Amy Winehouse. As for sentiment, Shakespeare and Amy Winehouse's bodies of work display a similar amount of the eight mentioned emotions being measured, relative to the other 44 musical artists. Essentially, of the 45 musical artists being examined, Amy Winehouse writes about the same topic in the most similar manner to Shakespeare.

Artist	Word Rank	Sentiment Rank	Overall Rank
amy-winehouse	6	1	7
nickelback	2	8	10
cake	9	3	12
adele	1	21	22
joni-mitchell	11	11	22
leonard-cohen	8	15	23
paul-simon	10	22	32
blink-182	18	14	32
bob-marley	27	6	33
rihanna	25	10	35

Table 1: Ranked Top 10 Most Similar Music Artist to Shakespeare

Keyword extraction alone allows for insight into which of the 45 sampled musical artists are most similar to Shakespeare's sonnets in terms of theme. The artist Adele ranks most similar to Shakespeare when considering only keyword extraction. This can be interpreted such that both Shakespeare and Adele write about similar themes in their works. For example, Shakespeare and Adele have the keywords "love" and "heart," and they both indeed write about facets of love. While Adele ranks highest among the 45 sampled musical artists, she ranks 21st in terms of sentiment. So while Shakespeare and Adele both may write about similar topics, the emotions expressed in their works differ enough to place Adele as 4th overall ranked in similarity to Shakespeare. This makes sense, as Adele tends to write about the frustrations and anxieties of love, while Shakespeare tends to write about the joys and excitement associated with love.

Artist	Word Count	Frequency	Keyword	Word Rank
adele	3	566.00	love time heart	1
nickelback	3	11777.00	love time yet	2
bieber	2	580.00	love time	3
dolly-parton	2	617.00	love time	4
dj-khaled	2	928.00	mine time	5
amy-winehouse	2	3944.00	love time	6
bjork	2	5625.00	love yet	7
leonard-cohen	2	6280.00	love time	8
cake	2	8042.00	love time	9
paul-simon	2	8836.00	love time	10

Table 2: Ranked Top 10 Most Similar Music Artist to Shakespeare Based on Keywords

Sentiment analysis alone allows for insight into which of the 45 sampled musical artists are most similar to Shakespeare's sonnets in terms of how the emotions anger, anticipation, disgust, fear, joy, sadness, surprise, and trust are expressed. Amy Winehouse ranks as most similar to Shakespeare when considering only sentiment. In analyzing the sentiment Shakespeare expresses in his sonnets, all 154 poems were collapsed into a single corpus. This resulted in an analysis that concludes that Shakespeare's most observed emotions are sadness, fear, and trust. While the poetry of Shakespeare is often not associated with such emotions, the manner of the analysis reveals that the work of Amy Winehouse is associated with the emotions of fear, trust, and sadness, respectively. Amy Winehouse tends to write songs of trusting old and current lovers along with songs about being in sad about loves ending.

Artist	Sent. Euclidean Distance	Sentiment Rank
amy-winehouse	0.001227	1
eminem	0.001237	2
cake	0.001304	3
nirvana	0.001365	4
bob-dylan	0.001471	5
bob-marley	0.001492	6
johnny-cash	0.001589	7
nickelback	0.001827	8
britney-spears	0.002097	9
rihanna	0.002249	10

Table 3: Ranked Top 10 Most Similar Music Artist to Shakespeare Based on Sentiments

Sentiment analysis of the 45 sampled artists were also implemented in k-means cluster analysis . Ten separate clusters were determined with Shakespeare being seen in the bottom right corner cluster, in orange.

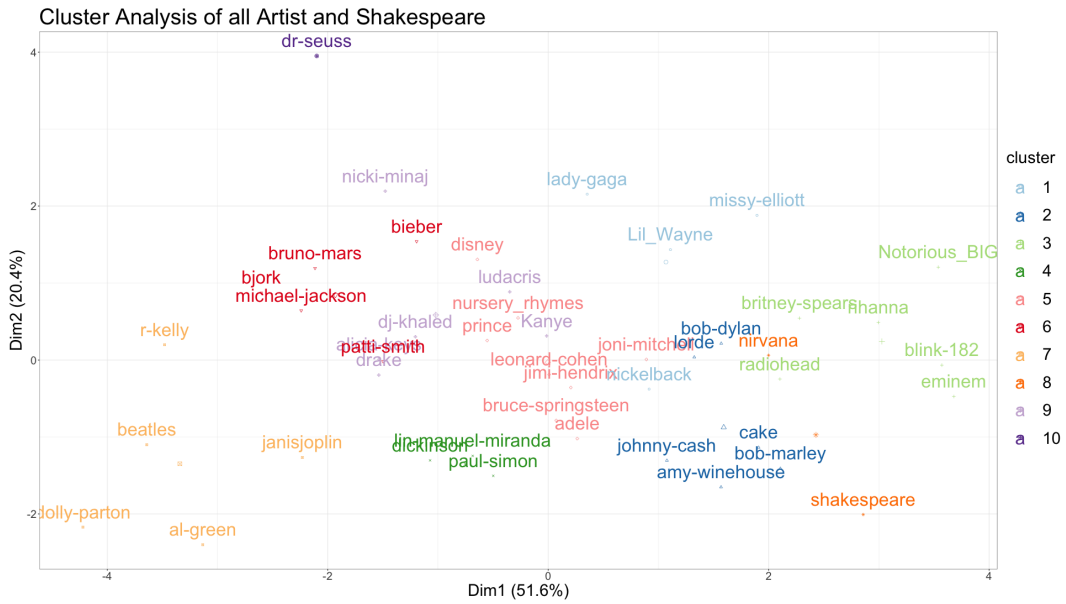

Figure 2: Cluster Analysis using 8 Emotions from 45 Artists and Shakespeare

4 Conclusions

In order to compare the Elizabethan works of Shakespeare to the modern musical artists of today, Keyword extraction and Sentiment analysis were used to compare Shakespeare's original 154 sonnets to the bodies of work of 45 sampled artists. Keyword extraction, alone, revealed that Adele is most similar to Shakespeare with crossover words including "heart" and "love," while sentiment analysis on its own revealed that Amy Winehouse is most similar to Shakespeare. Combined keyword extraction and sentiment analysis show that Amy Winehouse is most similar to Shakespeare, since she ranked 4th most similar to Shakespeare in terms of keyword extraction and 1st most similar to Shakespeare in terms of sentiment.

The 154 sonnets were all collapsed into a single corpus for analysis. If the sentiment analysis is done on the corpora of sonnets and not a single corpus, Shakespeare's sentiments tend to be more positive, with joy and trust being the most common emotions. If the sentiment analysis is done on a single corpus of all the sonnets, Shakespeare's sentiments tend to be more negative, with fear and sadness being the most common emotions. This inconsistency can be seen as a limitation of the study. The sentiment analysis done also did not include the positive and negative measures of analysis. Including these measures in future analysis could provide insights not otherwise seen. For future research, one solution would be to not collapse the sonnets and instead perform sentiment analysis on each individual sonnet and add up the 154 sentiment scores. A concern of this would be that if the sonnets were not collapsed into a single corpus, why collapse the songs of the 45 sampled artists. In experimenting with the methodologies, it was observed that collapsing the songs of the artists and not collapsing the sonnets resulted in sentiment analysis that is more easily interpretable. Future analysis would need to determine the exact effect collapsing all the work of a single artist has on the sentiment analysis of the corpus.

Since the bodies of works of each artist were provided as a single corpus, there was no feasible way to separate the corpus into individual songs. This limitation prevented the possibility of performing sentiment analysis on individual songs. A continuation of this project would need to find an automated method to separate the songs from each other for a given artist. One possible method includes a strict k-means clustering - like algorithm that would separate clusters based on criterion that would include repetition of lyrics (repeated chorus in a song would be an indicator of a single cluster), and sentiment associated with a single song. A BERT (Bidirectional Encoder Representations from Transformers) model could also be applicable here. If a BERT model determines the next line of a song to be probabilistic, it can be reasoned that it would be in the same song. Implementation of such an algorithm would allow for an automated way to separate the songs of the corpus.

Different dimensions could have been considered for cluster analysis as well. For example, the reading level associated with Shakespeare's sonnets could have been considered and compared to the reading levels of the bodies of work of the 45 artists and plots of k means cluster analysis could have been made. Other dimensions include line compositions such as the amount of type of word (adverb, adjective, noun), popularity of an artist compared to the popularity of sonnet, and use of punctuation.

Notably, some of the songs in the bodies of work of the musical artists contained lyrics or songs in different languages. Theoretically, this would have little effect on the sentiment analysis, but the methodology applied does not support cross language analysis. Because of this, it would be necessary to translate the lyrics to English, for more proper sentiment analysis. This could have had an effect on the keyword extraction as well. There could have been more instances of keywords that were not counted since they were not in English. Another worry could be the fact that Shakespeare's sonnets are written in early modern english. This makes it difficult to read and understand, and could have some influence on sentiment analysis. Solutions to this could include using new sonnets that translate the early modern Shakespearan sonnets to modern english. This could clarify the emotions of the sonnets, while not losing any meaning.

In the process of gathering results, there were some noted inconsistencies in the generation of the final results. When downloading the data from the RDS, special characters were added to the download data. These characters were cleaned before analysis; however when pulling a new version of the code from GitHub it was noted that the special characters were changing from machine to machine which intern was leading to different results. in order to reproduce the results please make sure to remove these special characters in the data cleaning sections of `_complete_analysis.R`.

Other methodologies could have been implemented as well to examine similarities among the corpora. Further emphasis could have been placed on n-grams to measure the collocation of keywords from the corpora. Identification of such collocated words would have allowed for a more accurate measure for keyword extraction, since collocated words are counted as a single word. The sonnets and bodies of works could have undergone tagging as well, where certain tags such as "nature", or "love," or "politics" could have been assigned to corpora that detail the theme of the work. A naive bayes classifier could then have been implemented to determine which sonnets belong to which tags and which sonnets correspond to artists that also possess similar tags. Finally, a BERT model could also be used in this scenario to provide a more complex methodology for sentiment analysis. BERT models often are used for next word predictions in text analysis, but a classification layer could be added on top of the transformer output to create a token for classification needs.

The original idea of this project included the classification and analysis of synthetically made Shakespearean pseudo sonnets. A Markov Model was created and trained with the original 154 Shakespearean sonnets and then 5000 sonnets were created and stored on an off site location. Classification and analysis of these

sonnets would include methodologies similar to those seen in this project, with sentiment analysis allowing for a thematic tagging of the pseudo sonnets. Classification trees could then be used to separate the pseudo sonnets. With appropriate tags associated with each pseudo sonnet, analysis would follow as frequency counts that could lead to regression analysis. Analysis could have gone as far as to see whether the tag frequencies of the generated sonnets were proportional to the tags of the original sonnets. Incorporation of these pseudo sonnets with the comparison of the 45 sampled artists would have also been possible. Similar analysis would be implemented on the individual songs of the artists that would then allow for a one to one association between pseudo sonnet and song. Based on analysis done with tagging and sentiment analysis, a model could theoretically provide similar pseudo sonnets and songs, given either an original sonnet or song.

In all, this project highlights the applications of Natural Learning Process methodologies with the intent of comparing separate corpora. With the use of NLP methodologies, a quantifiable comparison between the corpora of many persons was done with a final quantifiable comparison being made between which musical artist is most like that of an Elizabethan bard. This should highlight the significance and application that is capable of the process and methodology. This methodology could be used in more professional settings. From the perspective of a record label looking to sign potential artists, there is currently a large demand for a guideline for implementing a model in the music industry to increase the success rate of artist selection[11]. There have also been many instances of gender and racial discrimination of record labels in their selection of who to sign onto record labels[9]. A solution to this could include the implementation of NLP methodologies much like the ones used in this project that compare the corpora of a single artist's work to the work of an artist that is already considered popular or successful. This would allow for a non biased approach in comparing the possible success of a potential signing artist. The methodology of this project compared two artists in terms of keywords and sentiment analysis, but expansion of this analysis to compare a potential artist to an already successful artist could result in a measure for potential success of the signing artist. If such a methodology is then implemented in the music industry, a new framework can be created surrounding a quantifiable measure of success between potential artist and current artists.

References

[1] 56 Fun Facts about William Shakespear. https://nosweatshakespeare.com/resources/shakespeare-facts/. Accessed May 31, 2022.

[2] An Introduction to Knowledge Graphs. http://ai.stanford.edu/blog/introduction-to-knowledge-graphs/. Accessed June 3, 2022.

[3] K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawback Accessed May 31, 2022.

[4] NRC Word-Emotion Association Lexicon. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. Accessed June 3, 2022.

[5] Streaming Drives Global Music Industry Resurgence. https://www.statista.com/chart/4713/global-recorded-music-industry-revenues/#:~:text=Apr%208%2C%202022-,Music%20in%20the%20U.S.,year%27s%20total%20of%20%2421.9%20billion. Accessed June 3, 2022.

[6] Syuzhet Package Documentation in R. https://www.rdocumentation.org/packages/syuzhet/versions/1.0.6. Accessed May 20, 2022.

[7] Tm: Text Mining Package Documentation in R. https://rdrr.io/rforge/tm/man/. Accessed May 20, 2022.

[8] What is Text Analysis? A Beginner's Guide. https://monkeylearn.com/text-analysis/. Accessed May 20, 2022.

[9] Luis Aguiar, Joel Waldfogel, Sarah B. Waldfogel. PLAYLISTING FAVORITES: MEASURING PLATFORM BIAS IN THE MUSIC INDUSTRY. *NBER Working Paper*, (No. 29017), 2021.

[10] Jose P. G. Mahedero, Álvaro MartÍnez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon. Natural language processing of lyrics. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, page 475–478, New York, NY, USA, 2005. Association for Computing Machinery.

[11] Matthias Seifert and Allegre L. Hadida. Facilitating talent selection decisions in the music industry. 47(4):790–808, 2006.

[12] Sergio Oramas, Luis Espinosa-Anke, Francisco Gómez, and Xavier Serra. Natural language processing for music knowledge discovery. *Journal of New Music Research*, 47(4):365–382, 2018.

[13] William Shakespear. Shakespear's Sonnets. Project Gutenberg, 1997 [Online]. EBook-No. 1041, Accessed May 20, 2022.