Results

Natural Learning Process requires the creation of a corpus to perform keyword extraction and sentiment analysis on the sonnets of Shakespeare (***Data Table?***). Separate corpora of 45 sampled musical artists were made to perform more keyword extraction and sentiment analysis for comparison to Shakespeare (***Data Table?***). 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 (Table <#>***Insert Table of Combined Rankings***). 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 (Table <#>***Insert Table of Shakespeare Keywords***). “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.

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 (Table <#> ***Insert Table of Keyword Rankings***). 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.

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. (Table <#> ***Insert Table of Sentiment Rankings***). Bob Dylan 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 (Table <#> ***Insert Table of Shakespeare Sentiment scores***). While the poetry of Shakespeare is often not associated with such emotions, the manner of the analysis reveals that the work of Bob Dylan is associated with the emotions of fear, trust, and sadness, respectively (Table <#> ***can include Bob Dylan sentiment table***). Bob Dylan tends to write songs of political engagement and the want for change, so his body of work reflecting such emotions seems to correlate appropriately.

Sentiment analysis of the 45 sampled artists were also implemented in k-means cluster analysis (Table <#>). Seven separate clusters were determined with Shakespeare being seen in the bottom right corner cluster, in blue. Notably, Bob Dylan and Amy Winehouse are in the same cluster as Shakespeare. This, naturally makes sense since both are highly ranked as being similar to Shakespeare in terms of sentiment analysis.

Conclusion

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 Bob Dylan 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 5th 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.

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. These methodologies have been used in other fields of study such as pharmacology. For instance, NLP methodologies were used to measure the quantifiable differences in clinically significant information (CSI) between the descriptions of original and generic antimicrobials (Shimazawa *et al.,* 2018). These methodologies were used to identify what differences existed between original and generic medications in an automated way allowing pharmacists to quickly be able to determine the safety and potential dangers of generic medications. In the field of biology, NLP methodologies are used to compare gene co-expression networks (GSNs) between organisms to find areas of the gene that are topologically similar, thus finding organisms that have similar gene expressions (Ovens *et al.*, 2021). The automated process of NLP methodologies allowed these biologists to find potential connections in organisms that could have an impact on determining the clades of a species evolutionary biology, since organisms with similar gene expressions are likely to have evolved from a common ancestor. These applications showcase how significant the use of these NLP methodologies can be, outside of a social science setting.