THE MYSTERY OF THE BARD: A STYLOMETRIC ANALYSIS OF SHAKESPEARE

Claire Gellner

American University: Data641

ABSTRACT

Three of Shakespear's tragedies: *Julius Caesar*, *Hamlet*, and *Macbeth* are analyzed to determine if they were written by the same author. These plays are paired with excerpts from the *King James Bible*, *Paradise Lost*, and poems by William Blake. The six works are used in training classification models, both a Naïve Bayes and a CNN. The model is then used to classify a previously excluded Shakespeare play as Shakespeare or Other. While the results of the Naïve Bayes models are inconclusive, the CNNs accurately identify Shakespeare and classify all three of the plays as Shakespeare, indicating singular authorship. — Shakespeare, Natural Language processing, Stylometry, History, English

1. INTRODUCTION

First officially postulated in the 18th century, the hypothesis that the William Shakespeare did not write the plays and poems attributed to him endures until today [6]. There certainly was a William Shakespeare of Stratford-upon-Avon, whose name did appear as the author of some plays and poems [5]. However, this Shakespeare was an actor and is not thought to have been educated, worldly, or even literate enough to have penned works attributed to him [8]. Thus, even during his own time, there was some doubt that Shakespeare could be the true author of Shakespeare's works [8]. Shakespeare's authorship has always been attributed based almost entirely on the presence of his name on the works themselves.

The fact that Shakespeare's name appears on certain works as the author is not enough to be considered proof of authorship. Pointedly, some works once originally credited to Shakespeare on this basis are now attributed to other authors [5]. Therefore, it is possible that more, if not all the works attributed to William Shakespeare are in fact the work of someone else, or multiple someone-elses. Throughout the centuries, several different candidates have been proposed as the true author of Shakespeare's works. These candidates include notable figures in history such as Henry Neville, Edward de Vere (the 17th earl of Oxford), Sir Francis Bacon, Christopher Marlowe, John Fletcher, and even Queen Elizabeth I herself [8], as well as 83 other possible candidates [6]. These claims to authorship have been heralded by prominent writers and scholars throughout the centuries including Mark Twain, Charlie Chaplin, and Sigmund Freud [8]. However, there is little tangible evidence that any of these are in fact the author of Shakespeare's works [8]. Much of the evidence is circumstantial and is related to persons with whom the author of Shakespeare's works claims to be acquainted and the places this person would have to have been familiar with when compared to the people and places authorship candidates would have known [5]. However, some newer studies explore text-based analyses of word usage in the works compared with the authorship candidates using NLP and Machine Learning [5].

This new avenue of analysis is promising. Using natural language processing and stylometry, it is possible to make a mathematical comparison of Shakespeare's works and those of the other candidates. There have been notable studies conducting stylometric assessments of Shakespeare. Often, these analyses compare Shakespeare's writings with those of the most likely candidates. One of the earliest, Bolton Horton 1987 uses kernel estimation of n-grams to determine the distribution of certain words and groups of words in the writing of Shakespeare and John Fletcher [2], a playwright [11]. Bolton Horton finds that the word pairings in Fletcher's works are, in fact, close to the word pairings used in works attributed to Shakespeare [2], providing some evidence that Fletcher may have written Shakespeare's works. Later, Matthews 1993 uses a neural network to compare Shakespeare's writing to that of John Fletcher, finding that some of the plays in question were more Shakespearian, others more Fletcherian, and some in between [7]. A similar method was repeated later by Merriam 1994 who used a neural network for classification between Shakespeare and Christopher Marlow, finding that there is strong evidence for Marlowe being the author of *The True Tragedy* [9]. Similar methodology has been used since to add more evidence for the Shakespearian authorship question. However, it remains largely contested and inconclusive. Furthermore, what if Shakespeare wasn't the pseudonym of one author, but of multiple? Aljumily 2015 uses a linear hierarchical clustering model (Mean Proximity) and Principal Components Analysis to represent the works of Shakespeare, Sir Francis Bacon, Christopher Marlowe, John Fletcher, and Thomas Kyd in the vector space and assess the similarity of n-grams used by each author, finding support for the hypothesis that Shakespeare's works were written by multiple authors [1].

Echoing the findings of Aljumily 2015, this analysis seeks to examine if three of Shakespeare's plays were penned by the same author, or if they were the work of different authors. Assuming that William Shakespeare did not write the works himself, and given that works previously attributed to Shakespeare were reattributed to other authors, then it could be the case that some of Shakespeare's most famous plays are still misattributed, or that multiple authors used the name Shakespeare as a pseudonym.

2. METHODOLOGY

In order to examine the claim of multiple authors, this analysis uses two separate classification analyses. If Shakespeare's works had multiple authors, then it is expected that the Shakespearian works will not consistently be in the same class as they will not be sufficiently similar to each other.

This analysis focused on three of Shakespeare's works: *The Tragedy of Julius Caesar, The Tragedy of Hamlet, Prince of Denmark,* and *The Tragedy of Macbeth.* These three plays are available in the Python Gutenberg package. The first step was to prepare the plays for analysis. All three plays were sentence-tokenized and put in lower case, with punctuation removed. The punctuation was quite stubborn and difficult to remove, with many instances not properly removed. The texts were not lemmatized or stemmed as this would truncate words in such a way that it might obfuscate the author's word choice, therefore style, and thus could affect proper classification. In accordance with Miranda García, A., & Calle Martín 2012, the stop words were not removed for the classification models but were removed for the n-gram frequency analysis [10]. The plays were then set up as data frames with a column added to indicate the author.

This process was repeated with three non-Shakespearian works for comparison. These three works were the *King James Bible*, *Paradise Lost* by John Milton, and an anthology of poems by William Blake. Again, all three works were available as part of the Gutenberg package. These works were chosen as they are closer to Shakespeare's time, thus using a similar lexicon to Shakespeare, and were written in verse and meter, just like Shakespeare's plays. Thus, the contrast between Shakespeare and non-Shakespeare would not be determined by anachronism and modern writing styles or by meter and verse, but by the features of writing style and word choice, yielding more accurate results.

The three non-Shakespeare works were combined into one corpus and taken in subset of 3,500. This combined non-Shakespeare dataset was then appended to a combination of two Shakespeare plays, with the third excluded, serving as the experimental play. For example, to test if *Julius Caesar* was written by Shakespeare, the no-Caesar data set used for model building consisted of the non-Shakespeare subset, *Hamlet*, and *Macbeth*. This way, all three plays could be tested separately without being used to train the models. The 3,500-row subset of the original no-Shakespeare data set allowed a roughly 50:50 split between Shakespeare and not-Shakespeare for each final dataset (no_ceasar: Shakespeare 0.518768, Other 0.481232; no_hamlet: Other 0.540791, Shakespeare 0.459209; no_macbeth: Shakespeare 0.527346, Other 0.472654).

2.1. Frequency Analysis

The experiments were broken into three parts: introductory exploration, a Naïve Bayes classification model, and a neural network (CNN). For the introductory exploration portion, the Shakespearian works were each considered separately as well as combined while the non-Shakespearian works remained combined. All works underwent further preprocessing that removed all non-text characters and English stop words and re-tokenized the texts by word. The word frequency of unigrams and bigrams was calculated for each work.

2.2. Naïve Bayes

The first model for predicting if Shakespeare's works were written by the same author was a Naïve Bayes model. One model was made for testing each play using the sentence-tokenized corpuses excluding one of Shakespeare's works (no_ceasar, no_hamlet, and no_macbeth). These corpuses were first split into training and test sets, and vectorized using CountVectorizer. The vectorizer only kept the top 29% most frequent words as a method of feature reduction. This analysis used the MultinomialNB Naïve Bayes model. The classifiers were intended to classify as "Shakespeare" (1) or "Other" (0). This was repeated using Kfold Cross validation to ensure the results, using 3 folds. The Naïve Bayes models were assessed using accuracy, F1, and confusion matrices. After the model was trained, the excluded play was run through the model as an experiment to see whether the model would classify it as Shakespeare.

2.3. Convolutional Neural Net

Three CNN's were developed. As with Naïve Bayes, one model was developed for each data set excluding one of the Shakespeare plays (no_ceasar, no_hamlet, and no_macbeth). The CNN was built using preprocessed, sentence-tokenized data. The data was tokenized using the tokenizer in the Keras package with the LabelEncoder included in Scikit Learn. The CNNs relied on pre-trained embeddings found in Google Glove6B. Similarly to the Naïve Bayes process, after the model was trained, the excluded play was run through the corresponding model as an experiment to see if the model would classify it as Shakespeare or not. Just as with the Naïve Bayes models, the CNNs classified text as either "Shakespeare" (1) or "Other" (0).

In order to evaluate the CNN's, a confidence matrix was developed in addition to the use of accuracy. Additionally, the matrix of probabilities generated by the CNN was matched to the text lines of the excluded play. This made it possible to see which lines were most likely "Shakespeare" or "Other" and to count how many were in each category.

3. RESULTS

3.1. Frequency Analysis

The results of the frequency analysis were expected. The most frequently used words in Shakespeare's plays are the names of characters, stage direction words such as "enter", and a smattering of other common and expected words such as "lord", "wife", and "king". The most frequent words of the non-Shakespearian works included "lord" and "king" as well, and also other no-longer-used words associated with the Early Modern English period such as "thy", "thee", and "ye" alongside modern English words such as "people", "man", "son", and "come". The results of this analysis indicate that a defining feature of the Shakespearian works is the use of stage direction words as well as character names. These words do not appear frequently in the non-Shakespearian works. The high frequency of words like "thou" and "hath" in the non-Shakespearian works indicates that the more antiquated aspects of Shakespeare's language will not be an issue for classification as these terms are present in both the Shakespearian and non-Shakespearian works.

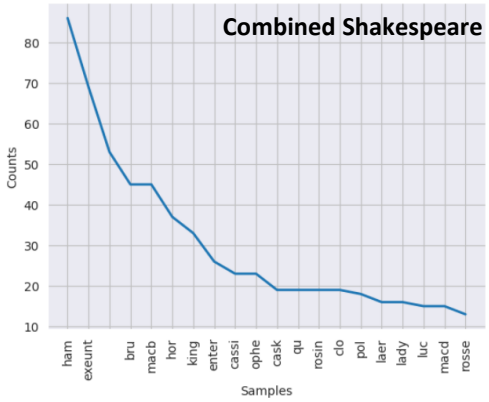

Figure 1: Frequency Analysis for All Shakespeare

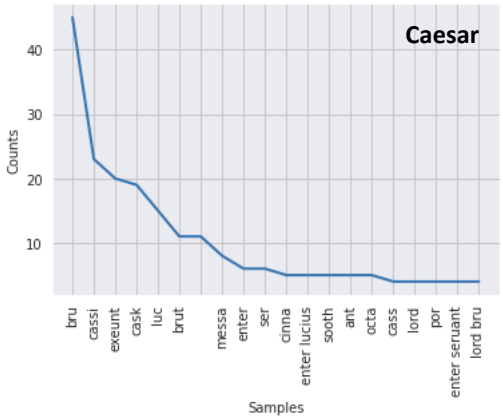

Figure 2: Julius Caesar Frequency Analysis

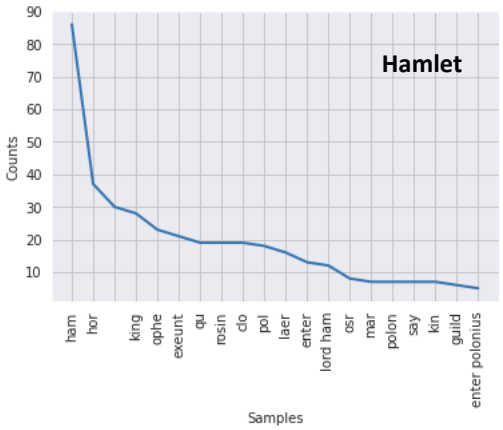

Figure 3: Hamlet Frequency Analysis

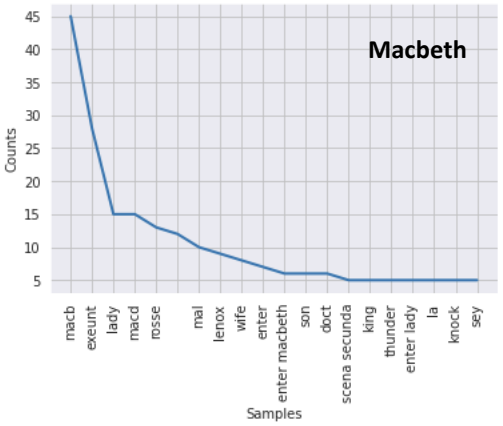

Figure 4: Macbeth Frequency Analysis

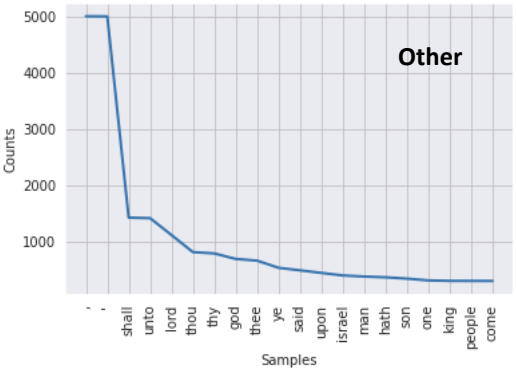

Figure 5: Non-Shakespearian Frequency Analysis

3.1. Naïve Bayes

The second stage was the Naïve Bayes classification modeling. The three models were trained on the composite data sets of 2 Shakespeare plays and a random sampling of the non-Shakespearian works. The Caesar classifier (trained on *Hamlet* and *Macbeth*) yielded an accuracy of 0.54 and an f1 score of 0.66. The Hamlet classifier model (trained on *Julius Caesar* and *Macbeth*) yielded an accuracy of 0.48 and an f1 score of 0.57. The Macbeth classifier model (trained on *Julius Caesar* and *Hamlet*) yielded an accuracy of 0.58 and an f1 score of 0.65.

Once the three classification models were trained, the remaining Shakespeare play was fed to the model with which it was not trained. This served to determine if the excluded play would be classified as Shakespeare or not. It was found that in two cases, the excluded play was predominantly classified as not-Shakespeare. Of the 1552 sentences in *Julius Caesar*, only 11 were classified as Shakespeare. For *Hamlet*, 75 lines were classified as Shakespeare, and 2278 were not. For *Macbeth*, 1346 lines were classified as Shakespeare, and 73 were not.

Even when the Naïve Bayes analysis was rerun using cross validation, the results were not much more accurate. In this case, the Julius Caesar classifier remained with accuracy, f1, and precision between 0.5 and 0.7 for all three-folds. The Hamlet classifier fared worse with accuracy around 0.47, f1 at 0.04, and precision at 0.1. The Macbeth classifier fared best with accuracy nearing 0.6, f1 around 0.7, and precision around 0.57. Interestingly, this time all three models predicted that the excluded play was Shakespeare. The *Julius Caesar* classifier predicted 1543 lines as Shakespeare and 9 lines as Other. The *Hamlet* classifier predicted 2245 lines as Shakespeare, and 108 as Other. The *Macbeth* classifier predicted 1351 lines as Shakespeare and 69 lines as Other.

The chance level accuracy of the Naïve Bayes classifiers may be the result of a lack of clear distinction between Shakespeare and not-Shakespeare. As seen in the frequency analysis, there was some overlap in most frequent words; perhaps the two groups are not significantly dissimilar. However, because the classifiers are so inaccurate, the results of the predictive experiment cannot be taken as valid.

3.2. Convolutional Neural Net

The CNNs yielded high accuracy in distinguishing Shakespeare from non-Shakespeare works. The test accuracy of the Caesar-classifying CNN, based on the validation set, was 0.85. The accuracy for the Hamlet-classifying CNN was 0.89, and for Macbeth-classifying it was 0.85.

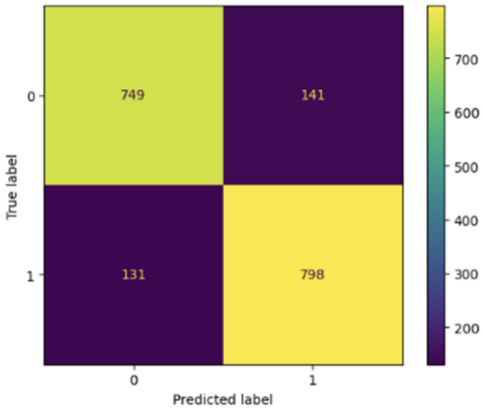

Figure 6: Caesar Classifier CNN

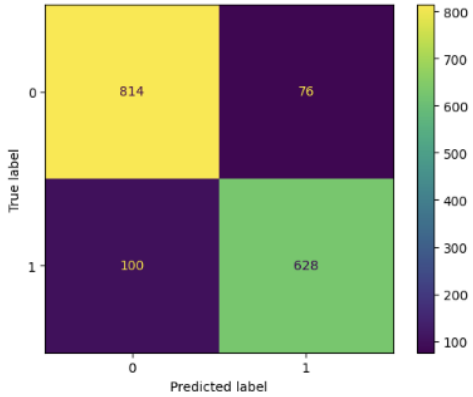

Figure 7: Hamlet Classifier CNN

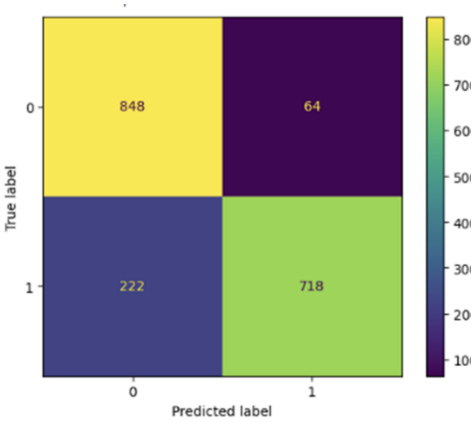

Figure 8: Macbeth Classifier CNN

The conclusion of the remaining play experiment showed a high probability of the Shakespeare's plays being classified as Shakespeare. *Julius Caesar*'s CNN prediction matrix ranked 1380 lines of 1552 as being "Shakespeare". *Hamlet* had 2027 lines of 2353 in the play classified as "Shakespeare". *Macbeth* had the fewest lines classified as "Shakespeare", with only 1002 of the 1420 lines classified as "Shakespeare".

A final experiment was run using the neural net. In this experiment, 10 random, well-known lines from Shakespeare were run through the Hamlet classifier (the highest accuracy classifier) to be classified as Shakespeare or Other. Of the 10 lines, only 4 were not classified as Shakespeare. These included well-known lines from *Romeo and Juliet* and other plays. One of the lines classified as "not Shakespeare" was "What's in a name? A rose by any other name would smell as sweet" (*Romeo and Juliet*, Act 2, Scene 2), as well as "All the world's a stage, and all the men and women merely players." (*As You Like It*, Act 2, Scene 7) [12][13].

Not Shakespeare	Shakespeare	Line
0.472847	0.527153	To be, or not to be: that is the question.
0.102138	0.897862	The lady doth protest too much, methinks
0.058416	0.941584	Good night, good night! Parting is such sweet sorrow.
0.030796	0.969204	If you prick us do we not bleed? If you tickle us do we not laugh? If you poison us do we not die? And if you wrong us, shall we not revenge?
0.901941	0.098059	All the world's a stage, and all the men and women merely players. They have their exits and their entrances; And one man in his time plays many parts.
0.121150	0.878850	Romeo, Romeo! Wherefore art thou Romeo?
0.169055	0.830945	If music be the food of love play on.
0.812713	0.187286	What's in a name? A rose by any other name would smell as sweet.
0.724526	0.275474	Shall I compare thee to a summer's day? Thou art more lovely and more temperate.
0.244594	0.755406	How sharper than a serpent's tooth it is to have a thankless child!

Figure 9: Test sentence classification

4. LIMITATIONS & FURTHER STUDY

The main limitation of this study is the availability of data. This study used only three of Shakespeare's plays and three non-Shakespeare works. The next step would be to repeat the experiment using Shakespeare's remaining plays. It may also be of interest to consider Shakespeare's sonnets and any other writings separately.

This belies the other issue: the Shakespearian works used here are plays, while the works of the "other" category were not. It was hypothesized that using works from a similar time-period that use meter and verse may be similar enough, especially when breaking the texts down to the sentence level. However, the results of this experiment may be affected by using other plays to serve as the "other" category instead. It is possible that the distinction between plays and literature, even at the sentence level is too big, leading to it being too simple to classify Shakespeare simply because it is a play.

5. DISCUSSION

This study set out to provide evidence for or against the hypothesis that Shakespeare's plays may have been written by different people. It was assumed that William Shakespeare did not write them himself, but his name used as a pseudonym, and thus the name may have been used by more than one person. From the classification models used in this analysis, there is evidence that this is in fact not the case. Most likely, the three plays used here were all written by the same author.

The results of the Naïve Bayes classifiers were inconclusive given that the classifiers had such poor accuracy. This also led to difficulties classifying the excluded play. Of the 6 Naïve Bayes models, 3 classified the excluded play as Shakespeare, and 3 classified it as other. However, the models were consistent between themselves in their classifications; all three models trained on a train-test split classified the excluded play as Other, and all 3 trained using Kfold CV classified it as Shakespeare. All the classifiers also classified the bulk of the excluded play as belonging to the same class, and results were not randomly split between classes as might have been expected for classifiers with accuracy barely above chance. Perhaps it was the amount of training content that allowed the Kfold Naïve Bayes classifiers to correctly identify the excluded plays as Shakspeare. Or perhaps, using Kfold allowed the classifier to be exposed to specific quintessential Shakespeare markers that may be randomly excluded from the training sets when using train-test split. Unfortunately, there is not much to be said about the Naïve Bayes classifiers except that the issue at hand may be too nuanced and complex for Naïve Bayes.

The CNN's were able to classify Shakespeare and non-Shakespeare with much higher accuracy. Similarly, the CNNs consistently classified the excluded Shakespeare play as Shakespeare's work. This may be a result of too much dissimilarity between the Shakespeare plays, and the works making up the Other category. However, if it was the result of dissimilarity between the works, it would be unlikely that the Naïve Bayes classifiers would have struggled to the extent they did; Shakespeare would not be mistaken for something else. Most likely, the success of the CNNs is indicative of single authorship of the three Shakespeare plays; the findings of the CNN experiment confirm that Shakespeare is easily classified as the work of a single author.

This study provides evidence that the works of Shakespeare were written by a single author. However, whether that author was 16th century actor William Shakespeare of Stratford-Upon-Avon or of some other and yet anonymous, more prestigious writer remains to be seen. Looking at the test sentences and the results of the excluded play experiment, a trend appears: the Shakespeare sentences tend to be shorter. This is in addition to the high frequency of character names and stage direction words. These stylistic markers can be considered some of the key elements indicative of Shakespeare. However, these could also be elements indicative of a play when compared to other literature styles, casting some doubt on if all three plays have the same author, or are all merely in the same class because they are plays.

While this analysis lends support to a single author for Shakespeare's works, it could also still be the case that others of what are currently considered Shakespeare's writings are misattributed, just not the ones analyzed here. Through the continued efforts of comparing Shakespeare's works against works of other potential authors, it may be possible to conclude if any or all of Shakespeare's works were written by another author, and who the author was. Continuing analysis of Shakespeare's works, or a similar analysis of other famous writing may help spotlight previously overlooked authors who were not able to claim their works during their lifetimes for socio-political reasons. Soon, we will be better able to appreciate these authors and their works. Or perhaps we will praise the enduring legacy and genius of the bard.

11. REFERENCES

[1] Aljumily, Refat. "Hierarchical and Non-Hierarchical Linear and Non-Linear Clustering Methods to "Shakespeare Authorship Question"." *Social Sciences* 4, no. 3 (2015): 758-799. doi:https://doi.org/10.3390/socsci4030758. https://www.proquest.com/scholarly-journals/hierarchical-non-linear-clustering-methods/docview/1721905109/se-2.

[2] Bolton Horton, T. (1987). *The Effectiveness of the Stylometry of Function Words in Discriminating between Shakespeare and Fletcher* [Ph D]. https://era.ed.ac.uk/bitstream/handle/1842/6638/Horton1987.pdf?sequence=1&isAllowed=y

[3] Craig, H., & Kinney, A. F. (2017). Shakespeare, Computers, and the Mystery of Authorship. In G. Taylor & G. Egan (Eds.), *The New Oxford Shakespeare : Authorship Companion*. Oxford University Press. https://web.p.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=7ef6a82e-cf6e-48aa-a85f-fa1fbf66dc2c%40redisMcCrum,

[5] Leigh, R. J., Casson, J., & Ewald, D. (2019). A Scientific Approach to the Shakespeare Authorship Question. *SAGE Open*, *9*(1), 215824401882346. https://doi.org/10.1177/2158244018823465

[6] *List of Shakespeare authorship candidates*. (2023, February 24). Wikipedia. https://en.wikipedia.org/wiki/List_of_Shakespeare_authorship_candidates#:~:text=Claims%20that%20someone%20other%20than%20William%20Shakespeare%20of

[7] MATTHEWS, R. A. J. (1993). Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, *8*(4), 203–210. https://doi.org/10.1093/llc/8.4.203

[8] McCrum, R. (2010, March 14). *Who really wrote Shakespeare?* The Guardian; The Guardian. https://www.theguardian.com/culture/2010/mar/14/who-wrote-shakespeare-james-shapiro

[9] MERRIAM, T. V. N. (1994). Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, *9*(1), 1–6. https://doi.org/10.1093/llc/9.1.1

[10] Miranda García, A., & Calle Martín, J. (2012). Testing Delta on the "Disputed Federalist Papers." *International Journal of English Studies*, *12*(2), 133. https://doi.org/10.6018/ijes/2012/2/161791

[11] Mowat, B., & Werstine, P. (n.d.). *About Shakespeare and Fletcher's The Two Noble Kinsmen | Folger Shakespeare Library*. Www.folger.edu. Retrieved March 6, 2023, from https://www.folger.edu/explore/shakespeares-works/the-two-noble-kinsmen/about-shakespeares-the-two-noble-kinsmen/.

[12] Simran, K. (2020, October 29). *10 of the Most Famous Shakespeare Quotes*. ThoughtCo. https://www.thoughtco.com/top-shakespeare-quotes-2833137

[13] *Ten unforgettable Shakespeare lines*. (n.d.). Www.bbc.com. https://www.bbc.com/culture/article/20150423-ten-memorable-shakespeare-lines