Stylometry: The True Author of Frankenstein

Alastair Thomas, Jack Young and Matthew Dailey

February 2022

1 Executive Summary

Frankenstein is a text of which there has been great debate, since its release in the 1800s, over who the true author of Frankenstein really was. Through the use of stylometry and machine learning techniques, we believe that the true author of Frankenstein was **Mary Shelley**, and we believe this is our best prediction, and it is 82% accurate.

2 Introduction

Stylometry is a field which combines mathematical techniques with the English language; it is the statistical study of literacy texts. We will make use of these techniques to investigate various stylometric methods to, ultimately, identify who the true author of *Frankenstein* was.

The book itself was first published (in English), in January 1818, albeit anonymously [2]. Three years later, a French version of the science-fiction novel was published and it accredited authorship to Mary Shelly. The novel follows the story of a young boy, who uses science to create a monster. The book has well-received publicly and since it was first released, it has been made into multiple films.

Over the years there has, however, been much dispute as to whether Mary Shelly was the true author of the text. Despite this, Mary and her husband Percy both always claimed that Mary was the true author of *Frankenstein*. Here, we will make use of statistical techniques to come to a conclusion on whether Mary was the true, sole author of *Frankenstein*.

To ensure we make accurate predictions, we will use a range of machine learning techniques, namely:

- Discriminant Analysis
- K-Nearest Neighbour (KNN)
- Random Forest
- Support Vector Machine (SVM).

We will discuss these methods in greater depth as we apply them in Section 4.

3 Data Pipeline

3.1 The Corpus

Our data is in the form of a corpus. As mentioned in Section 2, we have an unknown text (Frankenstein) which we denote by $\tilde{x}$, and we have texts from 11 other authors. Let us denote each author by j, and we have n_j texts which we know were written by the j^{th} author. Table 1 highlights the authors we have, and the number of texts written by them (in our corpus).

For each author, we have represented each text in a specific and consistent way, which will allow us to numerically compare the texts using machine learning techniques. We do this by counting how words, such as ‘a’, ‘all’ and ‘shall’, are used. These are known as **function words**. We have converted the word counts into the proportion of totals words in each specific text, before standardising them, which was done by subtracting the mean and dividing by the standard deviation. However, we actually only need to standardise before using KNN and SVM as these are the only techniques we will use which apply a distance metric. Standardisation wouldn’t effect discriminant analysis or the random forest methods. In our investigation, the stylometry functions which we use standardise the data before KNN, SVM and random forest, but not before discriminant analysis.

Author	Number of Books	Author	Number of Books
Bram Stoker	3	Percy Shelley (Poetry)	3
Charles Brockden Brown	5	Thomas Peakcock	4
Mary and Percy Shelley	1	Walter Scott	6
Mary Shelley	6	William Goodwin	5
Mary Wollstonecraft	2	William Polidori	1
Percy Shelley	2	Unknown	1

Table 1: Corpus Details

3.2 Multidimensional Scaling

Before we explore our machine learning prediction techniques, we can use Multidimensional Scaling (MDS) to learn more about the data we are dealing with. This will reduce the data we have from 71 features to just 2, which allows us to plot each text and thus visualise our corpus and see similarities between texts and authors. We will do this in two ways. Firstly, we apply MDS to each text individually and plot them, colouring by author, as seen in Figure 1. Secondly, we find the average of each feature for each author and then apply MDS, which means we get a plot of authors, as seen in Figure 2.

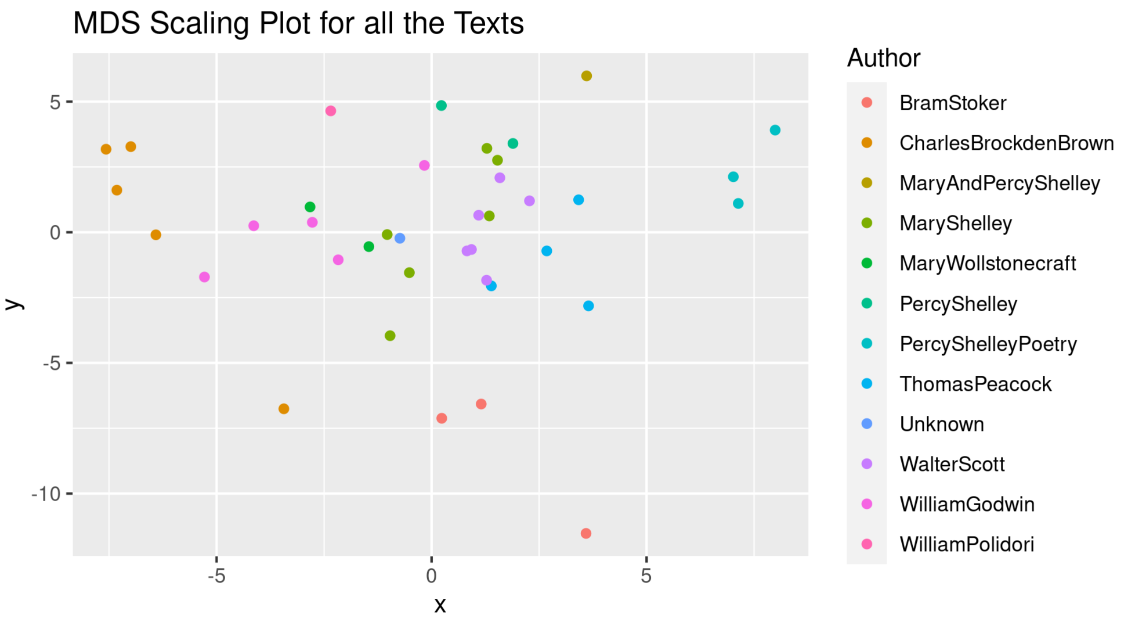

Figure 1: MDS for all texts

In Figure 1, we observe that there is a general cluster around the center of the plot, which contains texts from multiple authors, including all of Mary Shelley's work. Further to this, we see that there is a cluster of Charles Brockden Brown texts on the left-hand side of the plot. It is also clear that Percy Shelley's writing style in his poems is different of that in his books, as shown by his three poems which are isolated on the top right-hand side of the plot.

In Figure 2, we can see the general writing style of each author, where we have taken into account all of the texts by each author, collectively. Firstly, we notice again that it appears that Percy Shelley's writing style is different in his poems to his other texts, and there is a general cluster of authors around the center of the plot. There is also a smaller cluster at the bottom of the plot. We notice that the unknown author is placed very close to both Mary Shelley and Percy Shelley. At this point, the MDS plot would suggest that Mary Shelley is most likely to be the author of Frankenstein, the unknown text, and Walter Scott's text are also similar. We will explore this further using machine learning techniques in Section 4.

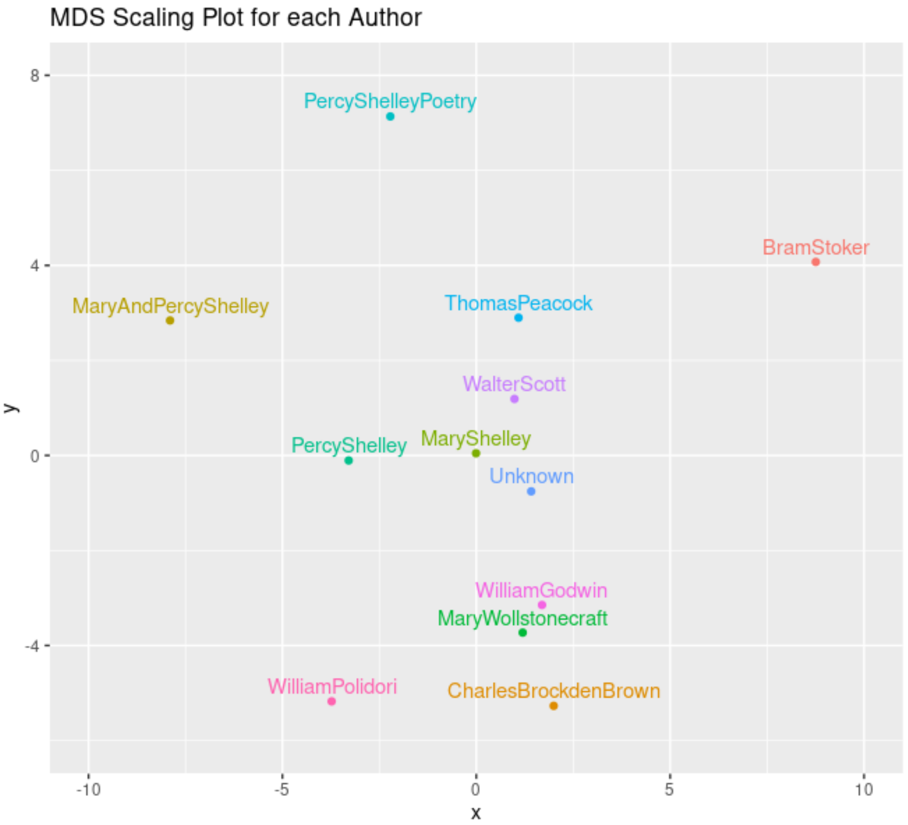

Figure 2: MDS by Author

3.3 Missing Data

The final thing to note about our corpus is that we have three authors with only one text. These are:

- Mary and Percy Shelley (joint authorship texts)
- The unknown author
- William Polidori

Therefore, we will exclude these texts from any testing sets as we cannot predict a text which the model knows nothing about. However they can be included in any training sets as it is possible that the model would predict one of these authors for any of the testing data. We go on to discuss model testing next in section 4.

4 Testing Models

4.1 Cross Validation

Before trying to predict who wrote the unknown text, we wanted to come up with an accurate and reliable model. As mentioned in section 3, we will be splitting our data into training and testing sets. This allows us to test various models on texts with known authors. We can then use various cross-validation methods for each of these models, to identify how well they predict.

We first made use of *leave-one cross validation* to assess the accuracy of our models. This method of cross-validation selects a text to predict, and then trains the model on all of the remaining texts. We then make the prediction for our selected text and check if it is correct. This process is repeated for all the texts, thus allowing us to assess how often our prediction is correct: i.e. the accuracy. Table 2 summarises the accuracy of our models. We can see that K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) are jointly most accurate, at 91.667%. This means that the model correctly classifies the unknown text almost 92% of the time.

However, this method overestimated the accuracy of our model due to over-fitting. To avoid this we then used *6-fold cross validation* to give a more realistic view of the accuracy of the prediction on the unknown texts. This method works by splitting the corpus into 6 folds, then we will take each fold in turn as a test set, train the model on the remaining 5 folds and then predict our selected test fold, comparing the predictions to the known true author. Again, the results are highlighted in Table 2, where we see that KNN and SVM are equally as accurate. If we did require more precision in our estimates of model accuracy, we could run our 6-fold cross validation 100 times and average the results.

Model Type	**Discriminant Analysis**	**K-Nearest Neighbours**	**Random Forests**	**SVM**
Leave-one Accuracy (%)	86.111	91.667	86.111	91.667
6-fold Accuracy (%)	78.125	84.375	75	84.375

Table 2: Model Accuracy

4.2 Confusion Matrix

To further develop our understanding of our models and their predictions, we can analyse the confusion matrix by creating Alluvial plots. Figure 3 summarises all four models and shows us where the predicted authors differ from the actual author. Each coloured set of lines represents the incorrect predictions made by each model. Firstly, notice that 7 authors are being incorrectly classified into 6 prediction classes. Interestingly, texts from Charles Brockden Brown, Mary Shelley and William Goodwin are being predicted as the unknown author, which agrees with what we know already from the MDS scaling; the styles of these authors are very similar to that of the unknown text. We can also see that these incorrect classifications to the unknown author happen only via Discriminant Analysis and the KNN model. The Random Forest and SVM models do not predict any texts to be written by the unknown author. However, the later two methods do still make six and four classification errors respectively.

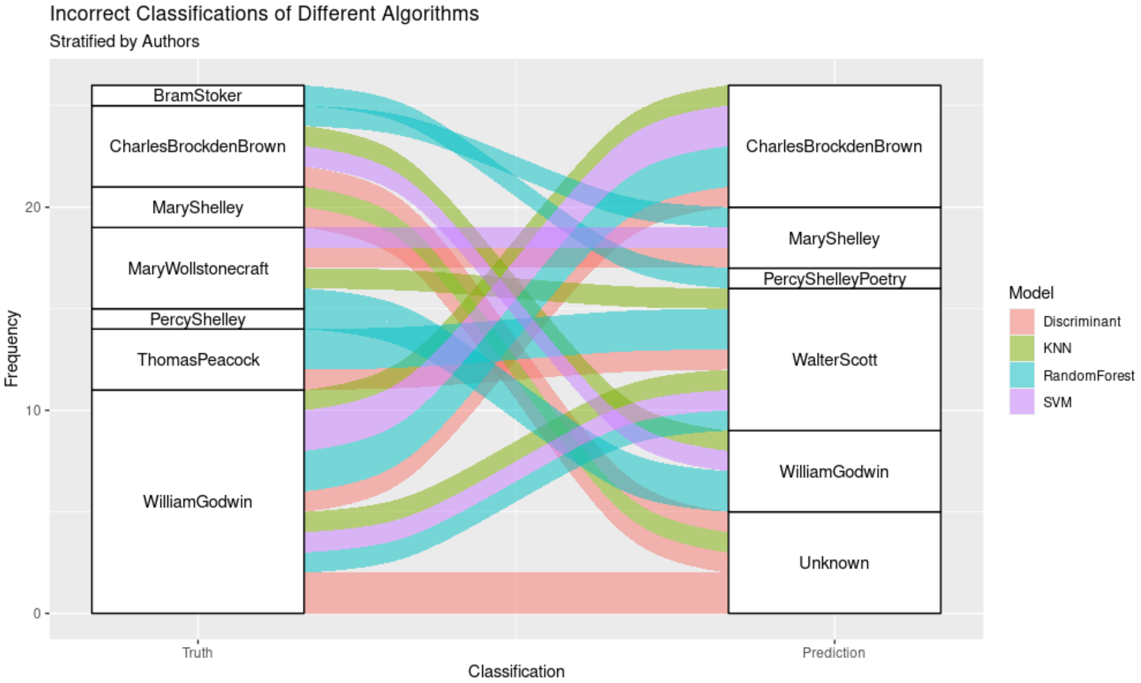

Figure 3: Model Prediction Analysis

4.3 Kappa Statistic

We can make use of the *kappa statistic* which is given when creating a confusion matrix using the caret package on R. This helps us understand our models and their predictions by putting a number on the difference between the model's observed and predicted accuracy. Each model's kappa statistic [1] was calculated as

$$\frac{\text{observed accuracy - expected accuracy}}{1 - \text{expected accuracy}}.$$

These values are shown in Table 3. We know that the higher the Kappa statistic, the closer the predictions are to the truth. We observe that SVM model has the highest kappa coefficient.

	Discriminant Analysis	**K-Nearest Neighbour**	**Random Forest**	**SVM**
Kappa Coefficient (%)	75.000	81.839	74.253	81.756

Table 3: Model Kappa Statistics

4.4 KNN Optimisation

The KNN method, with $k = 1$ provides, jointly, the most accurate predictions. We can further optimise this, by exploring how the accuracy varies for further values of k, once again applying 6-fold cross validation for each k value. The results are shown in Figure 4. We can see that as the value of k increases away from one, the accuracy falls. We can therefore conclude, that $k = 1$ should be used in the final model as this provides the most accurate prediction. However, we need to be careful as a small k value is indicative to an over-fitted model, but in this case as we don't have many texts, we are expecting a smaller k to give higher accuracy's.

To further develop our investigation, we could have optimised the parameters we used in our other models, both by trial and error or by using in-built optimisation methods. Examples of parameters to be optimised are *ntree* in the random forest model, or the cost and γ in the SVM model.

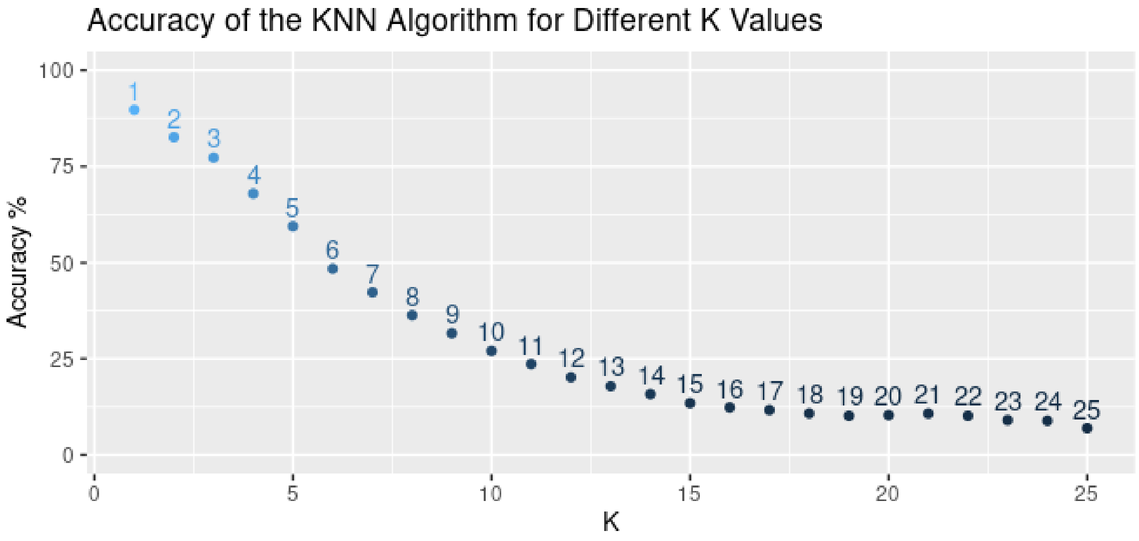

Figure 4: KNN Accuracy as k Varies

5 Frankenstein's True Author

Finally, we will use our model(s) to predict who the author of the unknown text is. This is done by simply training each model on all the texts, minus the unknown one and then predicting the unknown one. Table 4 summarises our findings. We can see that three our of our four models predicts Mary Shelley to be the true author of Frankenstein, with Discriminant Analysis predicting Walter Scott. This links back to our MDS Scaling plot by authors in Figure 2 where Mary Shelley was most similar to the unknown text and Walter Scott only slightly less similar than Mary Shelley. We can also say that the second most probable author of Frankenstein in the Discriminant Analysis model was

Charles Brockden Brown, which links back to Figure 3 where we saw that the Discriminant methods was incorrectly predicting Charles Brockden Brown's texts to be the unknown author.

	Discriminant Analysis	**K-Nearest Neighbour**	**Random Forest**	**SVM**
Frankenstein Prediction (%)	75.000	81.839	70.473	81.756
Predicted Author	Walter Scott	Mary Shelley	Mary Shelley	Mary Shelley

Table 4: Model Predictions

We have that 75% of our models predict Mary Shelley to be the true author and our two most accurate models (KNN and SVM) both predict Mary Shelley too. Hence, we can conclude that the true author of Frankenstein was Mary Shelley, and we can predict this with 82% accuracy using KNN, which was our most accurate method.

References

[1] Statistics How To. *Kappa.* URL: https://www.statisticshowto.com/cohens-kappa-statistic/. [accessed: 23/02/22].

[2] Wikipedia. *Frankenstein.* URL: https://en.wikipedia.org/wiki/Frankenstein. [accessed: 23/02/22].