A Study into the Effectiveness of a Variety of Statistical Measures for Distinguishing Between Authorship Styles

University of Sheffield

The following report is an investigation into how different statistical measures can be used for attribution of authorship. By looking at a variety of statistical measures the report will evaluate the difference in authorship between three known texts of Jane Austen, Emily Bronte, and Charlotte Bronte. Furthermore, through building a Random Forrest classifier the report will attempt to classify an unknown text to one of the known authors using the statistical measures evaluated.

Introduction

Authorship attribution is considered to be simply the task of identifying the author of a text. It is deemed an important task as it has great potential to be used in various applications such as determining the author of disputed texts throughout history, in forensics investigations to trace identities of illegal materials spread online, and for plagiarism such as detecting the collaboration in documents.

For this analysis, statistical measures will be used to attempt to resolve a closed set authorship attribution problem where the challenge is to determine the author of a piece of text where the set of authors is known. Four samples of texts have been taken from three different authors, Jane Austen, Emily Bronte, and Charlotte Bronte. Three of the texts, Texts A, B and C, have known authors, whilst D does not. Statistical means will be used to evaluate the differences between authorship between each of the three texts and additionally attribute authorship to the final Text D.

To distinguish the texts effectively between the three authors, it is best first to consider the known differences in styles between them. Jane Austen's novels are often considered "country house novels", which are more comedic and contain fewer narrative and scenic descriptions. Instead, Emily and Charlotte Bronte both wrote gothic novels which contain heavy use of narratives to set 'horror' scenes and descriptions to create tension throughout the plot. Furthermore, Both Bronte Sisters wrote poems as well as novels which use more figurative and rhythmical language in which may have influenced their novel writing as well.

Three main types of statistical measures can be used to distinguish between authorship.

Firstly stylistic measures. These are features such as lengths of words, sentences, and how rich the vocabulary is. These methods are fairly easy to acquire and can additionally help with distinguishing between different genres of texts

Secondly, word-based features. These look at analysing the individual words used in the writings, such as analysing the distribution of function words (non-content words) or word n-gram features, aiming to capture frequencies of different strings of n-consecutive words.

Lastly syntactic cues. This approach attempts to analyse the structure of language through in-depth syntactic analysis of the texts to compare how the sentences have been structured.

To evaluate the texts effectively, the report will aim to use a variety of the measures above. However, statistical measures involving syntactic cues are computationally expensive to analyse and therefore this report will only consider stylistic measures and word-based features.

By applying the statistical methods to each text, this report will also attempt to attribute authorship to the unknown text D through a Random Forrest Classifier. Through this method, the information will also be able to compare the importance of each of the different features used.

Summary of the data

The data consists of three chapters of text from three different books of different authors; Text A was from chapters 16, 34 and 37 from Mansfield Park by Jane Austen, Text B is taken from chapters 2, 22 and 32 of Jane Eyre by Charlotte Bronte, Text C is taken from Chapters 1, 16 and 23 from Wuthering heights by Emily Bronte and lastly Text D is taken from the first chapter of an unknown author. In total each of the texts range from 9,600 – 10,200 words and were sourced online from Project Gutenberg.

The chapters were picked at random from each of the books to minimize the effect that different types of chapters would have on the measures. For example, picking chapters where there was a lot of speech between the characters could give different results to a chapter that was a lot more descriptive.

To classify the final author the texts for each of the known authors is split into multiple documents composed of 10 sentences each, producing a total 150 documents across the three authors, each having between 80 and 400 words. This was deemed the best method as the classifier needed enough documents to be trained however still needed the documents to be of appropriate lengths to provide enough data to capture values that accurately represented the authors styles. As this is the format that the data has been collected to be put into the classifier, the data will be analysed in this format as well.

For the analysis 4 different types of metrics are compared.

Average Sentence Length of the Document: This method has been used in a multitude of studies where it is suggested that different authors may have a different distributions of sentence lengths.Furthermore it was stated by Yule in his work that sentence lengths may help to distinguish between authors as those who had more descriptive writing style wrote in long wandering passages and those who were more perspicuous wrote in briefer and shorter sentences(G. Undy Yule 1938). For this report, sentence lengths were considered over other stylistic measures such as word length or syllable length as it is less effected by the contents of the individual book or chapters. For example, some books may have names of people or places that would have an increased number of syllables or letters, making it seem as if the author used a lot of longer words. For this analysis sentences have been divided wherever there is a sentence ending in the document such as a full stop, exclamation, or question mark.

Frequency of different punctuation: Different types of punctuation can show different authorship styles as not only does it show the variety of punctuation that different authors use however the frequency in which they use them. Furthermore recording different types of punctuation used by authors is considered a good indicator of authorship as there is a lot of opportunity to vary their usage (J.Grieve 2007). For this analysis, the frequencies of four punctuations were recorded, '?', ':','!' and ':'. Full stops and commas were omitted as they were the most common types, given that the documents were split into 10 sentences separated by 'sentence enders' (.,?,!), recording the number of full stops would give inverted results to recording the number of exclamation and question marks and therefore not provide any new information.

Frequencies of different lexical words: Recording the percentage of verbs, adjectives, nouns and adverbs can show significant stylistic traits in authorship and have been used in a variety of studies. It has been suggested that a more refined intellectual habit of thinking can increase the number of nouns used, whilst a more dynamic empathy and active attitude can be expressed by an increase in number of verbs (H.Sommers 1966). Frequencies of lexical words can be used to distinguish between different genres as well as some genres comprise of more narrative and descriptive scenes that others. The frequencies of verbs, adjectives, nouns and adverbs were recorded and also normalized by dividing by the total number of words in the text in order to gain a percentage frequency of the word in the document.

Distribution of Function Words: Function words are words which don't contain content such as 'a', 'the', 'it' etc... They are considered an effective measure as authors do not consciously control the usage of these words in their writing and so are able to show their natural authorship style quite well. Furthermore they can capture the authors individual style, not just the genre or period of their work as all authors across different writing styles and periods use the very same function words. For the analysis, first a list of 277 function words were collected from an online Blog named Semantic Similarity. Then each of the texts were analysed and the frequencies of each of the function words are recorded and compared. For the model a list of all the top function words across the texts was comprised and the frequencies for each of these words per document was recorded. The values were also normalized by dividing by the total number of words in the text to gain a percentage frequency of the word in the document. Furthermore to remove biases that may have been caused by the narrative or gender of the main character, words relating to gender or narrative were removed.

Overall, the data sources seem to be sourced reliably. Each of the chapters were taken online from project Gutenberg, a volunteer effort created to encourage the distribution of E books and make them accessible to the public. Given that the aim of the project is to allow more people to have access to a variety of books it is unlikely that they have been edited in anyway. Furthermore, the list of function words was taken from an online blog created by a researcher who has released his own publications on a variety of topics in text processing. Although it would have been more reliable to have created a list tailored to the texts being used, this would have been very time consuming so using a pre-curated list was more efficient and is still reliable as it has been used in a variety of published works already.

Analysis

Figure one shows a boxplot of the distributions of average sentence lengths between the four texts. It can bee seen that median average sentence lengths does not show a lot of variation between the texts however does show a divide between Texts A and Texts B which have similar median average sentence lengths around 19, and Texts C and D have medians of just under 14. The mean of the average sentence length however does show more variation, likewise do the quantiles

of the average sentence lengths between the texts suggesting that the texts have different ranges of average sentence lengths.

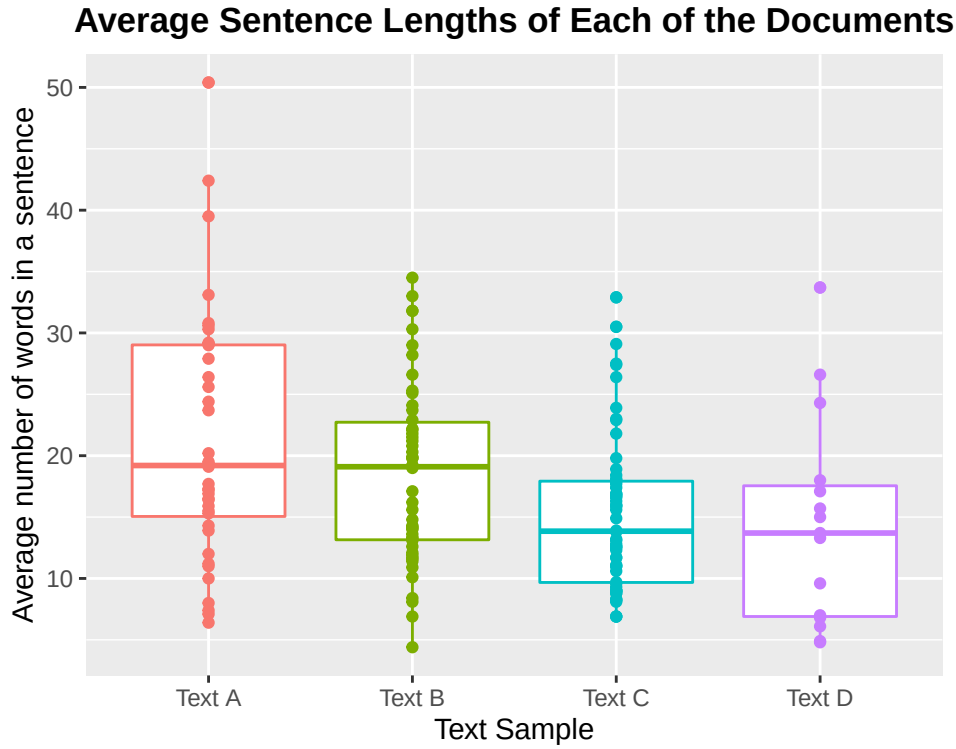

Figure 1: Box plots showing the average number of words in a sentence for each of the texts. It can be seen that there is not a large difference in distributions between the texts however text A does have the largest range of values and overall the longest sentences where as Text D has the shortest.

Table 1: Summary statistics of the average sentence length of each.

	class_label	mean	sd	Q1	Q2	Q3
1	Text A	21.220	10.020	15.100	19.200	29
2	Text B	18.540	7.300	13.100	19.100	22.700
3	Text C	15.160	6.350	9.700	13.900	17.900
4	Text D	14.440	8.580	6.900	13.700	17.600

Table 1 shows that Text A has the largest mean and range of average sentence length suggesting that the author Jane Austen varies her sentence lengths the most out of the three. The boxplots suggest that the distribution of average sentence lengths suggest of Text D is most similar to that of Text C - Emily Bronte, as they have the most similar medians and furthermore the most similar upper and lower quartiles.

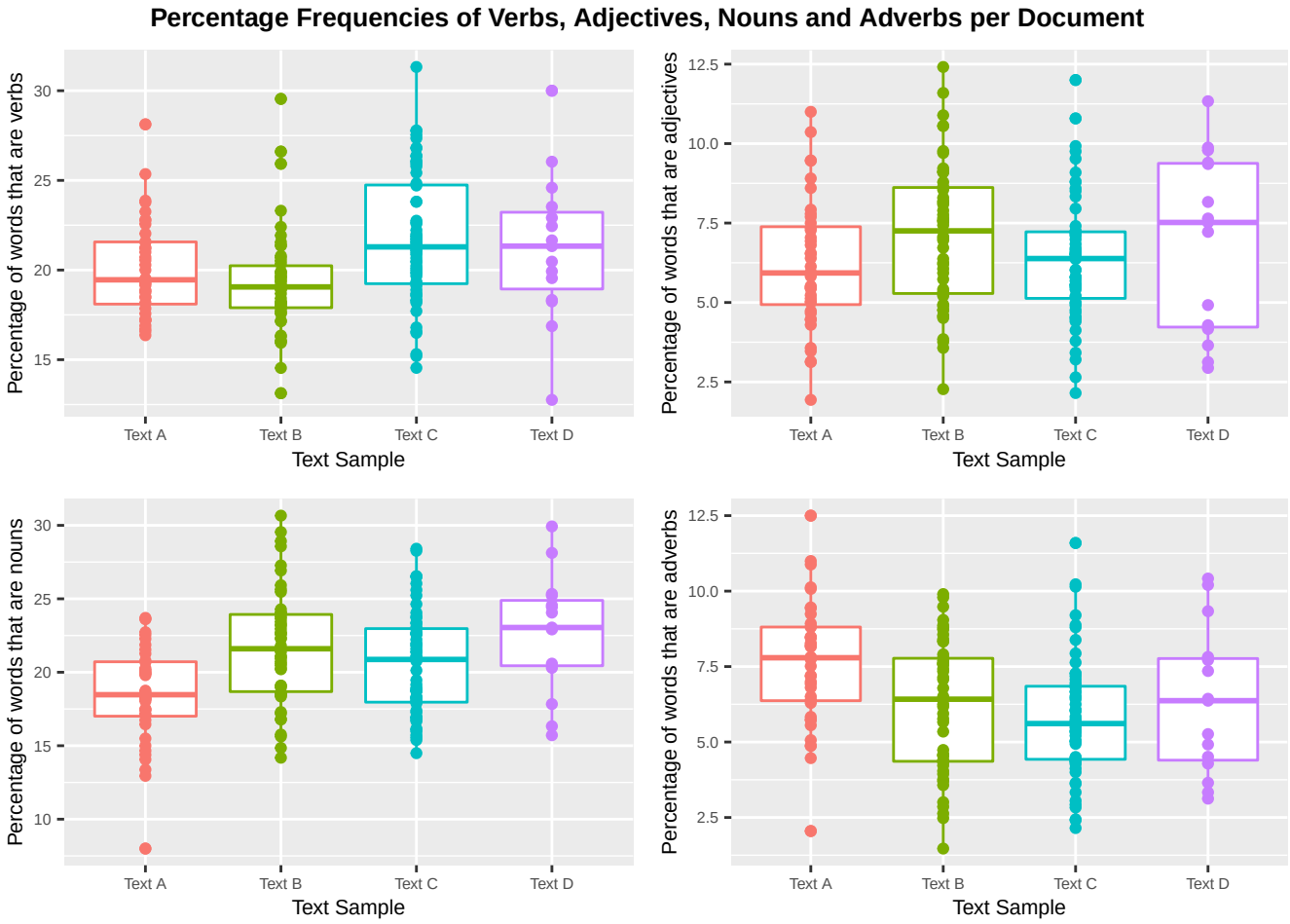

Figure 2: Box plots showing the frequencies of different punctuation type

The box plots above show the distributions of the average number of verbs, adjectives, nouns and adverbs per document.Although there are not many major differences between the distributions, some subtle patterns can still be seen.

For frequency of verbs, the plots show that Text A and B have lower median numbers of verbs than C and D however the range of Text D is most similar to that of Text B

For frequency of adverbs there is not a big difference between the medians of Texts B, C and D however Text A not only has the highest median but also has higher values in general. The plots show a general trend that Text A tends to have lower percentages for each of the lexical words except for adverbs, where it has a significantly higher median.

Text D tends to have the highest medians for each of the percentage frequencies except for adverbs suggesting that Texts A and D are the most dissimilar and were potentially written with different literary styles. This can especially be seen in Table 2 for the percentage frequency of nouns which show the largest difference between the medians of texts overall are between Text A and Text D at 4.5%.

Table 2: Table of the Avergae Frequencies of Different Lexical Words.

	class_label	Verbs	Adjectives	Nouns	Adverbs
1	Text A	19.460	5.930	18.470	7.790
2	Text B	19.060	7.250	21.600	6.420
3	Text C	21.290	6.380	20.870	5.610
4	Text D	21.330	7.520	23.050	6.370

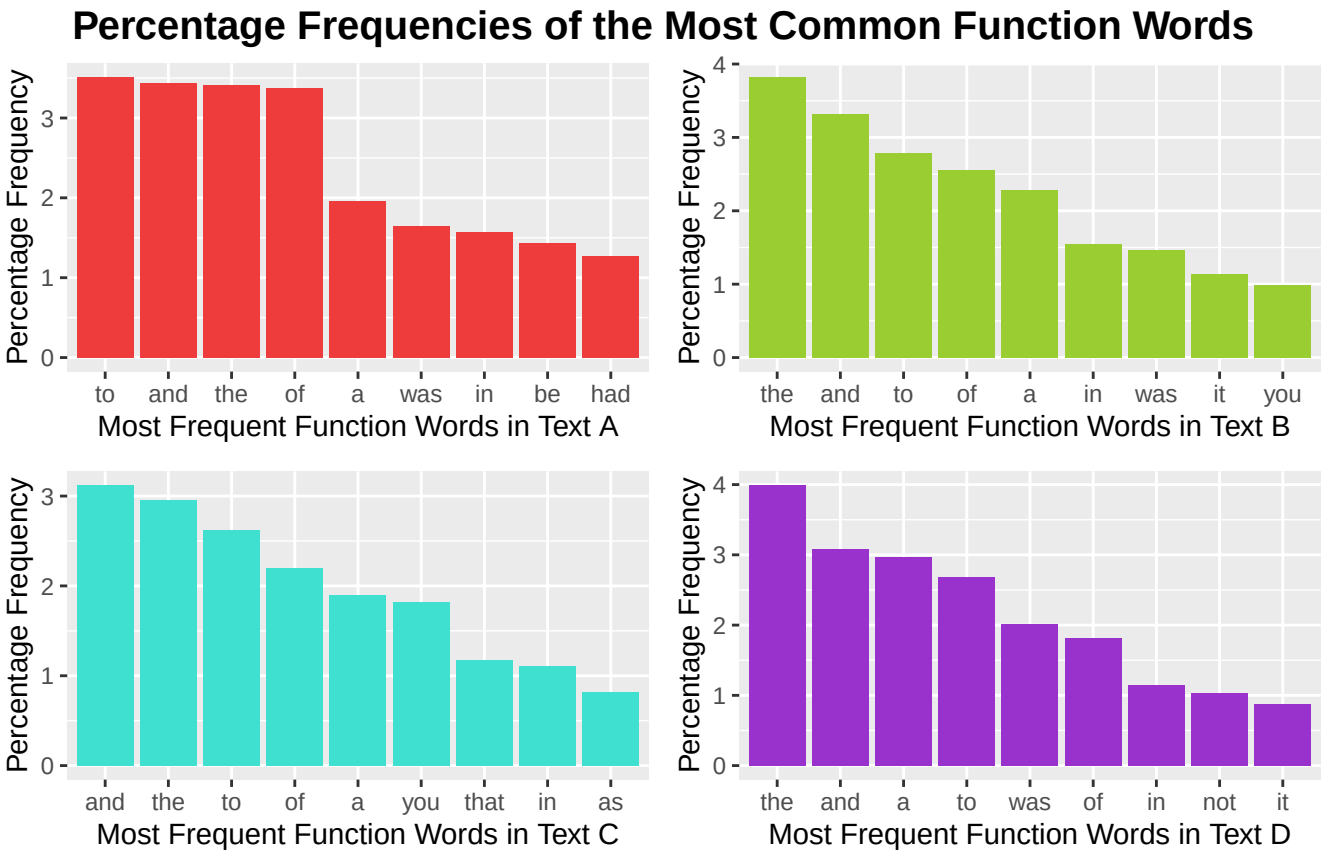

Figure 3: Bar graphs showing the 9 most common function words of each of the texts ranked in order of frequency . Overall the most frequent across all of the graphs appeared to be similar includding words such as 'to', 'the', 'and', 'of' and 'a'.

Figure 3 shows Bar charts for the 9 most common function words of each of the texts.

It can be seen that all of the Texts have very similar words in their top 9 list of Function words however this may be more down to the English language itself as these are all in the 5 most common words in the English language together.

One of the largest difference between the texts are the shapes of the distributions of the bar graphs, especially Text A.The bar chart for Text A shows that it has similarly high frequencies for each of its top 4 function words all at around 3.5/3.4% and then drops significantly to frequencies of 2% and under for the rest of the function words. The other texts tend to show a much more gradual distribution with not as large difference between the frequencies suggesting that the texts B,C and D had a more varied use of function words where as Text A, had a smaller set of function words used more frequently.This can further be seen as the 9th most frequent word for Texts B,C and D all have frequencies bellow 1% where as for text A, the word 'had' still has a frequency of 1.27%

One of the main differences between Text D and the other texts is that it has a significantly higher frequency of the word 'a' at around 3% compared to Texts B which 'a' has a frequency of 2.3% and Texts A and C with even lower frequencies both under 2%.

One of the main differences between Text C and the other texts is that the word 'was' does not appear in its top frequencies. Each of the other texts show percentage frequencies of 'was' to be at least 1.4% where as the lowest frequency on C is 0.82%, suggesting was has an even lower frequency than this.

Texts D and Texts B appear to show some similarities as they both have a significantly higher percentage frequency of the word 'the' with frequencies close to 4% as opposed to Texts A and C have frequencies of 3.4%.

Although there are some words which appear in some charts and not in others, it is not appropriate to say that these show significant differences in the texts. The graphs only show the frequencies of the top 9 function words, so although some function words rank higher for different texts, they may still have similar frequencies overall.

Figure 4 shows the frequency distributions for different types of punctuation.

The two most distinguishing features in the plots are number of colons and number of exclamation marks. Text B has significantly higher number of colons than the other texts, as not only does it have the highest median however has

Table 3: Frequencies of the Most Common Function Words for Texts A and B.

	Most.Frequent.Words.A	Percentage.Frequency.A	Most.Frequent.Words.B	Percentage.Frequency.B
1	to	3.510	the	3.820
2	and	3.440	and	3.320
3	the	3.410	to	2.790
4	of	3.370	of	2.560
5	a	1.960	a	2.280
6	was	1.650	in	1.540
7	in	1.570	was	1.460
8	be	1.440	it	1.140
9	had	1.270	you	0.990

Table 4: Frequencies of the Most Common Function Words for Texts C and D.

	Most.Frequent.Words.C	Percentage.Frequency.C	Most.Frequent.Words.D	Percentage.Frequency.D
1	and	3.120	the	3.990
2	the	2.960	and	3.080
3	to	2.620	a	2.970
4	of	2.200	to	2.690
5	a	1.900	was	2.020
6	you	1.820	of	1.820
7	that	1.170	in	1.150
8	in	1.110	not	1.030
9	as	0.820	it	0.870

the largest upper quartile and individual value as well. This suggests that it would be a key feature for distinguishing texts of Charlotte Bronte.

Text C has the highest median number of exclamation marks and also the highest interquartile range, however it still shows some similarities with text B which also has occasional documents with high number of exclamation marks.

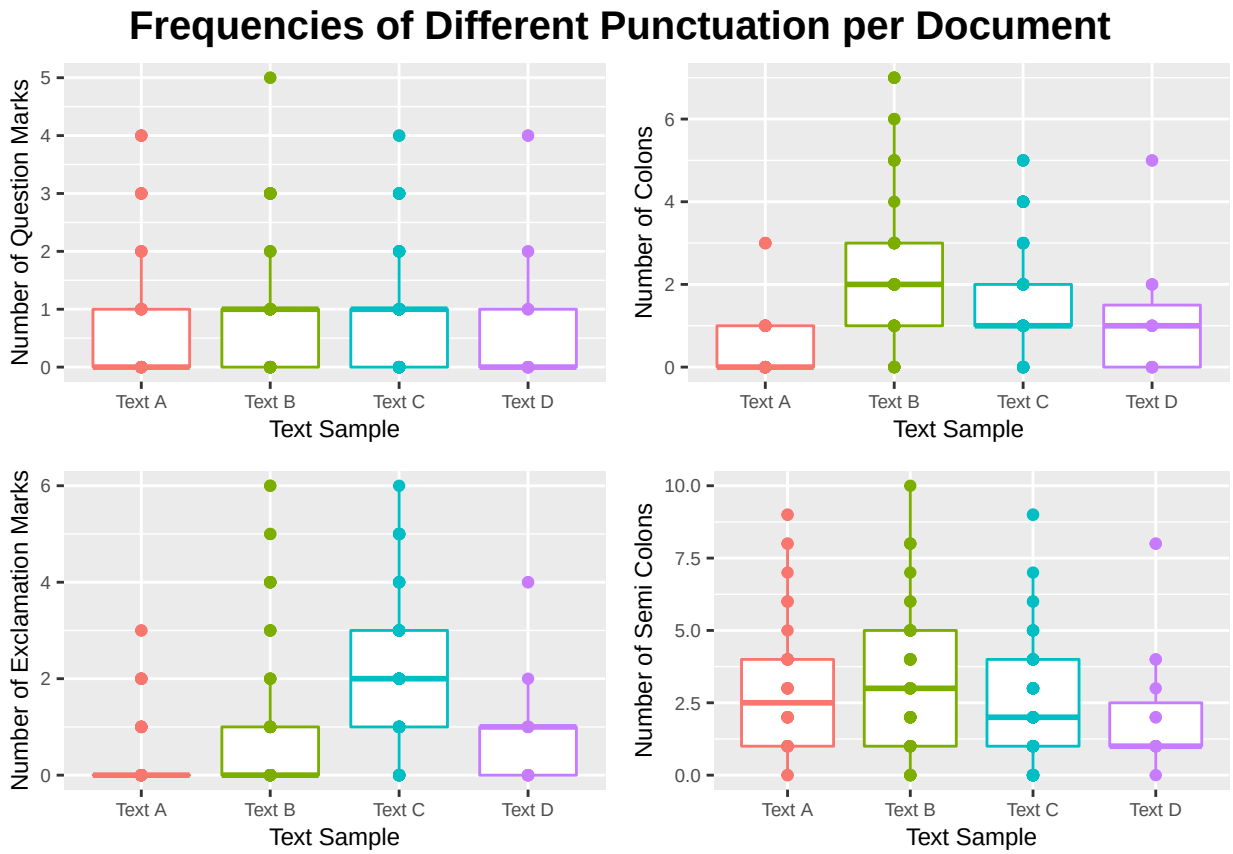

Figure 4: Box plots showing the frequency distributions of different punctuations in each of the Texts.

Frequencies of semicolons show the biggest distribution and range between the texts themselves however overall do not show many patterns that could be used to distinguish between the texts. This is especially as all the texts appear to show a frequent use of semi colons apart from Text D which suggest that this is not a very good metric to distinguish the authorship of Text D. Likewise can be said for frequencies of question marks as all of the Texts appear to have quite low values and there is not much difference between the texts.

Classification

The majority of the data was kept the same when put into the classifier except for a few differences. Firstly given that the texts had different words in the top 9 frequencies of function words, the function words were taken to be list of frequencies of 13 words for each of the texts in order to capture all of the words present.

Secondly the Documents in text D were combined to get average values for the whole chapter in order to achieve one classification for the whole document instead of several classifications that could produce conflicting results. Once again, frequency of lexical words, punctuation and function words were normalized to account for a larger document size.

Before training the model, a standard scalar was used so that each feature could be scaled to a unit variance. This reduces the bias of features in the model as variables measured at different scales do not contribute equally to the model fitting.

The model was then trained using 80% of the documents and fit to a Random Forrest Classifier which was chosen to be the most suitable for a number of reasons. Firstly Random Forest can handle a large number of features as it has embedded feature selection to regulate feature importance itself maning dimensionality reduction does not have to be performed. This also means that Random Forrest makes it possible to evaluate feature importance and compare how each of the features affected the results of the classifier.Furthermore, Random Forrest is quite robust and can handle data with outliers well itself so they do not have to be removed.Therefore apart from scaling the values, no other pre-processing steps were needed for the data and all of the features used in the analysis were used in the model.

Lastly some basic hyper parameter tuning was done by comparing different accuracy values for number of trees and depth of the trees for the model to average.It was found that the Random Forrest Classifier with a maximum number of

trees of 60 and with maximum depth of 10 nodes produced the highest accuracy.

Results

The model predicted the author of Text D to be the same as Text B, Charlotte Bronte.

Overall the model had quite an average performance.It was able to produce an accuracy of 0.7 suggesting that out of all of the data points predicted, 70% were predicted correctly. The model was able to produce a high recall for documents of Texts A and C however not for documents for Texts B in which it had a very low recall of 0.36. Furthermore The model also produced fairly decent precision scores all of them around 0.7.

From Table 4 it can be seen that the model has high recall but low precision for Texts A and C, this means that the model is able to predict a lot of documents that are of Texts A and C correctly however also over predicts the amount of documents (especially of text B) to be of those classes, as can be seen from the heat map bellow as well.

In contrast, for documents of Text B the classifier has a higher recall (0.67) but very low precision(0.36). This suggests that it does not miss-classify a lot of documents that are actually of Texts A or C to be of Text B however it also means that it does not predict a lot of documents to be of Text B in general. This also suggests that a lot of the documents it classifies to be of Text B are in fact of Text B.

Furthermore the heatmap shows that the majority of miss-classified documents were between Texts B and C - authors Charlotte Bronte and Jane Eyre. Two documents were predicted to be written by Charlotte Bronte when they were in fact written by Emily Bronte, and furthermore 10 documents were miss-classified the other way round.

Table 5: Evaluation Metrics of the Classifier

	Class	Precision	Recall	F1_score
1	A	0.700	1	0.820
2	B	0.670	0.360	0.470
3	C	0.710	0.830	0.770

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```

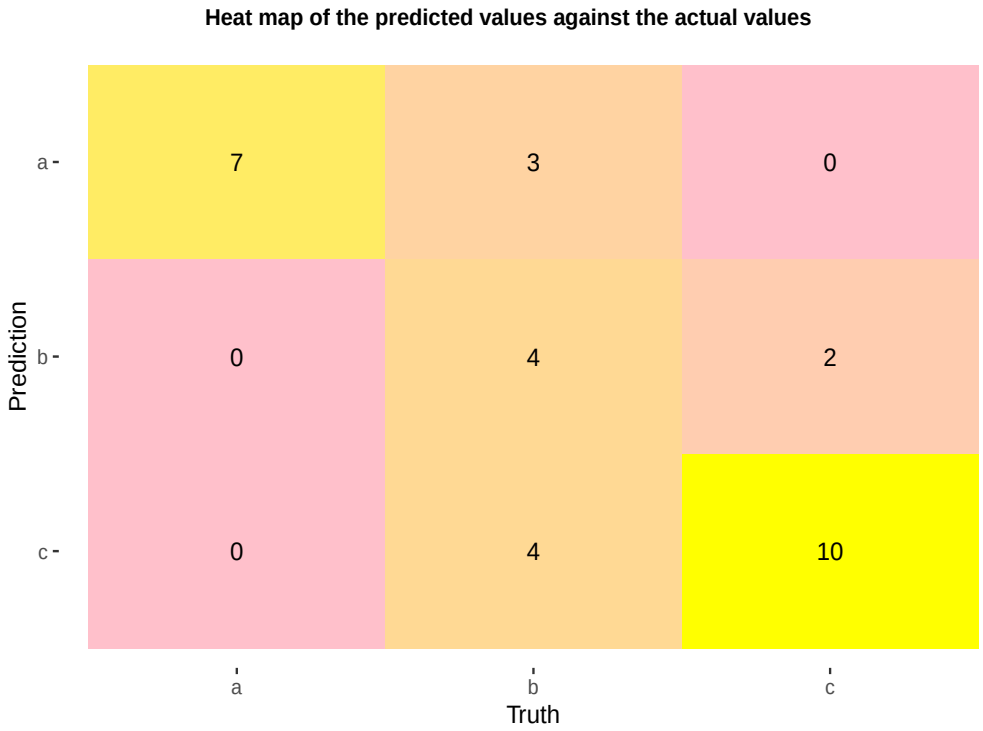

Figure 5: Heat Map of the confusion matrix of the results of the classifier. The Y axis shows How many of each of the texts were predicted to be of which class and the X axis shows what the actual class of the text was. For example the top middle box has a value of three suggesting that three documents were predicted to be of the author A - Jane Austen however were actually of author B - Charlotte Bronte

The figure bellow shows a break down of the importance of each of the features in the model for classifying each of the documents.

The graph shows that punctuation frequency had the most variable results. Colon and exclamation mark frequency had two of the highest feature importance where as semicolon and question mark frequency had the two lowest, with colon frequency being almost 8 times more significant.

Frequencies of different lexical types were able to be of somewhat importance and also performed more consistent as a whole. Adverb frequency was the 2nd most important feature and furthermore verb and noun frequency were in the upper half.

Average sentence length and distribution of function words do not show to have any significant feature importance in the model, all producing feature importance of 0.05 and lower.

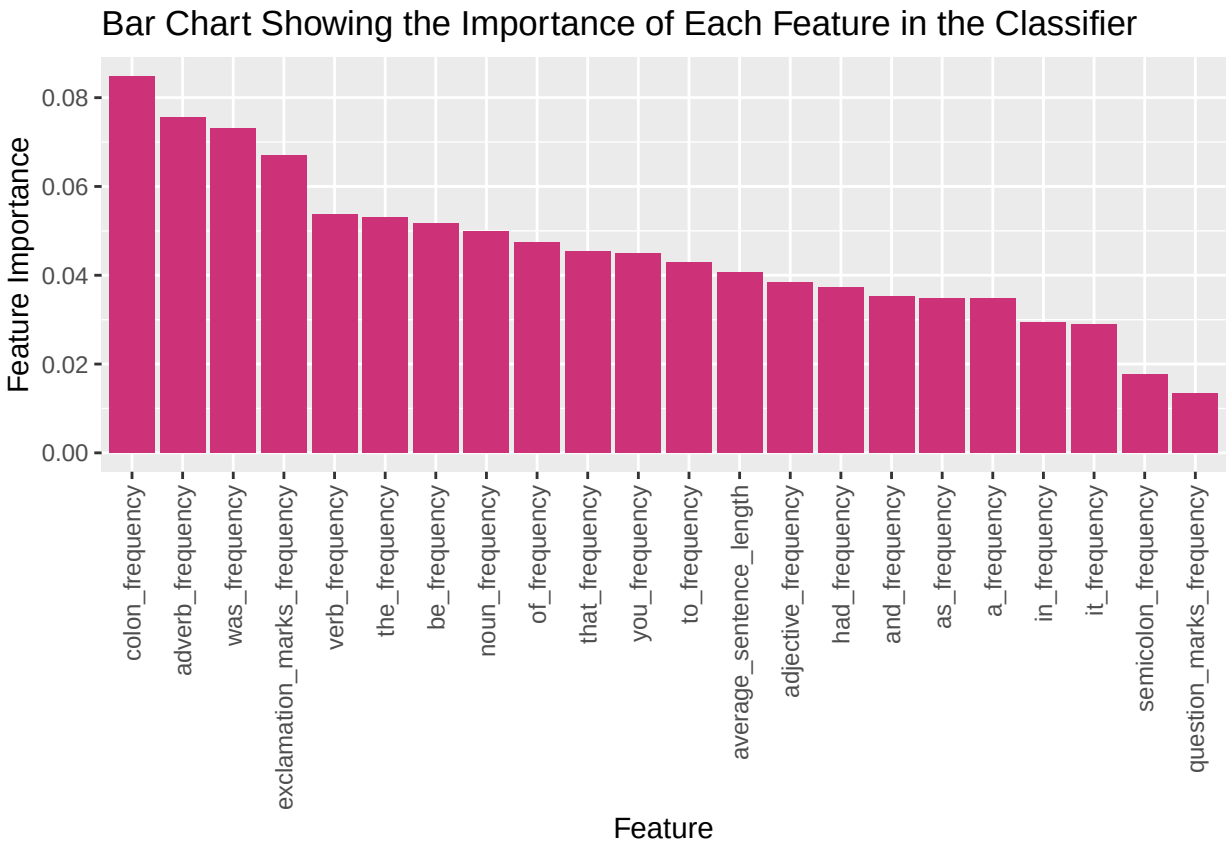

Figure 6: Bar graph showing the importance of each of the featues in the model in order of importance. Frequency of colons, adverbs and the word 'was' were seen to be the features that the classifier used the most to distinguish between the authors

Table 6: Importance of each feature in the classifier

	feature_importances	feature_names
1	0.041	average_sentence_length
2	0.013	question_marks_frequency
3	0.085	colon_frequency
4	0.067	exclamation_marks_frequency
5	0.018	semicolon_frequency
6	0.054	verb_frequency
7	0.038	adjective_frequency
8	0.050	noun_frequency
9	0.075	adverb_frequency
10	0.073	was_frequency
11	0.029	it_frequency
12	0.047	of_frequency
13	0.037	had_frequency
14	0.043	to_frequency
15	0.035	a_frequency
16	0.052	be_frequency
17	0.035	as_frequency
18	0.029	in_frequency
19	0.053	the_frequency
20	0.035	and_frequency
21	0.045	you_frequency
22	0.045	that_frequency

Conclusion

To conclude, the report suggests that the most likely author for Text D is Charlotte Bronte. Although the model did not produce particularly high accuracy overall, given that the model is overly cautious but also very accurate at predicting texts that are of in fact to be written by Charlotte Bronte, it can be said that it is likely to be quite an accurate prediction given by the model.

The analysis and model evaluation suggest that there is not one statistical measure that is particularly better than the other at attributing authorship as not all features from one measure perform better than all features from another. Therefore it is not possible to determine if one of the four metrics used on a whole is better at distinguishing between authorship and suggests that in reality, to evaluate between different authors effectively, a combination of metrics need to be used, not just minor variants of each other. One of the features of most importance was colon frequency, which was used mainly to distinguish Charlotte Bronte's work. This is probably as during the Victorian era punctuation such as colons and semicolons were used a lot more frequently than in the modern era. Although the distribution of colons did not seem to show much feature importance, the increased use of colons by Charlotte Bronte suggest that this was more due her choice of writing style rather than something that could not have been controlled such as function words.

Another feature of high importance was adverb frequency. Jane Austen appeared to use significantly more adverbs than Charlotte and Emily Bronte as seen in the analysis and was probably the main reason that the classifier was able to not miss classify any of the documents from either of the Bronte's to be of Austen's. This has also been noted in other analysis where it has been seen that Austen used words such as 'really', 'very' and 'quite' in particularly high frequencies (D. Morris, 2017).

The model did not appear have much trouble classifying Jane Austen's work however was not able to distinguish between either of the Bronte's work that well suggesting that the measures used were better at distinguishing authors of different genres.As in the report it was known that the author of Text D would belong to one of Text A, B or C's it was suitable to use the known genres of each of the authors to distinguish writing styles. However if a classifier was to be made to distinguish texts where the author was not known to be in the training set, it would be more appropriate to use metrics that focused more on the individual writing style than the style for the genre as well.

Discussion

The main way in which the model could be improved is by increasing the number of documents used. For complexity reasons this was kept small as the more data used, the more computational power and time it takes to train the model.However this also meant that the data was less likely to represent the actual authorship style of each of the authors, just that of those few chapters used. Realistically to properly capture the style of the author and evaluate the metrics, more documents would be needed to train the model and furthermore documents would need to would be significantly larger in size and range across multiple novels from each of the author's work. This can be seen especially for capturing the frequencies of punctuation used in each of the documents, as the analysis showed many of the documents recorded low number of punctuation frequencies which is most likely due to using small sample sizes. Using larger sample sizes would have helped to see a clearer distribution of the data as more would have been collected.

Some of the measures that were used in the report could have potentially produced better results if they were not averaged. This was mainly true of sentence lengths which was averaged so that it could fit into the model better, however by averaging out the data information is lost on the true distribution and variability. Another method could have been to treat it as a categorical variable such as having multiple feature for different lengths e.g. 1-10, 11-20, 21-30 etc... However, there were over 150 different sentence lengths recorded and so this method would have dramatically increased the number of features in the model if it was to be done in a way that was still able to capture significant amounts of data. Although random forest can handle many features, this is still in relation to the number of documents provided and so averaging the sentence lengths was chosen instead to better suit the model and for simplicity reasons as overall.

Another way in which the model could be improved is through using different metrics. A lot of recent research has evaluated authorship styles using syntactic measures which look at the structure of the document and the order of words in relation to each other. These have been seen to produce some of the best results, for example models which included metrics recording the numbers of different types of clauses used that were found by breaking down the structure of the documents were able to distinguish texts between 26 different types of authors with accuracy scores of over 0.77 (J.Soler-Company, L.Wanner,2018). However these metric were were omitted from this report as they are quite complex and require in depth look at the syntactic structure. Furthermore, analyzing the syntactic structure of a document is computationally expensive and there is a lack of text–processing resources to carry out these methods.

Lastly, different types of classifiers and machine learning methods could be used to produce a model with higher accuracy. The Random Forrest Classifier was chosen mainly as it allowed a breakdown of feature importance to help

evaluate how useful each of the features were in classifying the document, however using more complex models such as support vector machines and neural networks have the potential to be more expressive than decision trees. Furthermore better methods could have been used to tune the hyper parameters such as using k-fold cross validation where instead of just training the model to one set of training and testing data, the data is split up into batches to train the model across multiple sets.

References

George Udny Yule (2014). The statistical study of literary vocabulary. Cambridge: Cambridge Univ. Pr.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. Literary and Linguistic Computing, 22(3), pp.251–270.

Inc, S. (2017). Jane Austen's English: Thoughts for JAFF Writers | Diane H. Morris. [online] www.moorgatebooks.com. Available at: https://www.moorgatebooks.com/07/jane-austens-english-thoughts-for-jaff-writers/ [Accessed 8 Apr. 2022].

Leed, J. (1966). The computer & literary style: introductory essays and studies. Kent, Ohio, Kent State University Press.

Soler-Company, J. and Wanner, L. (2018). On the role of syntactic dependencies and discourse relations for author and gender identification. Pattern Recognition Letters, 105, pp.87–95.

Sources of data

Function words : https://semanticsimilarity.wordpress.com/function-word-lists/

chapters : www.gutenberg.org.