DATING OF LITERARY WORKS

180188141

Abstract. This report demonstrates the use of machine learning techniques to predict the publication dates of two of Charles Dickens' literary works, using some of his other work. A variety of different measures, such as word and sentence length, and techniques like random forest classification were tested, to greater and lesser success. Overall, a high level of accuracy was achieved by combining results of multiple methods, most notably a random forest regressor. This demonstrated that machine learning techniques can be useful in certain cases with smaller datasets.

Contents

1. Introduction

This report aims to contribute to the solution of the problem of dating two literary works by Charles Dickens, using extracts from these books and machine learning methods trained on six extracts from other books with known publication dates. Four more extracts from Dickens' books with known publication dates were included for use as a test set. While machine learning techniques are usually not applied to such small datasets, if the methods are successful they could be applied to other problems with small datasets. During the investigation, measures on the extracts such as the average length of words and sentences were created and machine learning algorithms were trained on the six books with known publication dates. The accuracy of the functions was tested and optimised on the four separate test extracts to find the likelihood of each prediction being correct. The functions were then applied to the two books with unknown publication dates to create a prediction of the publication date, and the accuracy of the predictions and performance of the methods were assessed.

2. Data Handling and Exploratory Data Analysis

2.1. **Data.** The data consist of 8 plain text files of extracts from the beginning of eight of Charles Dickens' books. These books are "The Pickwick Papers", "The Old Curiosity Shop", "David Copperfield", "Bleak House", "Little Dorrit", "The Mystery of Edwin Drood", "Hard Times" and "Our Mutual Friend". There are also data on the publication dates of the first six of these books. The goal was to use machine learning techniques to estimate the dates of publication of the latter two books, which were not provided with data on the publication dates.

To do this, the plan was to create a number of measures on the data, apply them to each book, and use these data with machine learning methods to predict the unknown publication dates. This relies on Dickens' writing changing over time, with better results the more consistent the change is. Some changes could be easily measured, such as the number of times he uses a specific word, but others such as his general writing style would not be easily measurable. However, a change such as this may still be indirectly measurable as it may impact other measures, such as the number of times he uses a specific word. To help check the accuracy of predictions, publication dates and extracts from the beginning of four other Dickens' books ("The Chimes", "The Haunted Man and the Ghost's Bargain", "A Tale of Two Cities" and "Great Expectations") were taken from the internet[1] for a test set, so the accuracy of the predictions on new books could be tested. The four books have a variety of publication dates to make tests as effective as possible, but all the books published from 1849-1858 were already included in the given data so no new books from that time could be used for testing. While it would be possible to split the given data into a training set and a test set, there are not enough provided books to warrant reducing the training set at all, so this is the best solution.

2.2. **Data Preprocessing.** To begin, each file was read into separate variables, and anything that could contaminate the data and skew the predictions of publication dates was removed. The texts included escape sequences like "\n", which is in the file to instruct the computer to display the following text on the next line. However, it is not part of the book so it should be removed. A separate variable containing an ordered list of the six known publication dates was also created. Some books were published over a period of multiple years, which could be accounted for by taking an average of the years. However, Dickens originally published his books as serials, meaning new chapters would be published weekly or monthly. This means the beginning of each book was published in the first year of the range, so it would be more accurate to take this value as the publication date.

The extracts are not the same length, which must be taken into account when creating measures. Some measures would differ depending on the length of the text, such as counting the number of times a word appears - it would have more chances to appear in a longer text. One way to compensate for extracts of different lengths is to cut them all down to the same length as the shortest extract. This is not ideal as it reduces the size of the already small dataset, but was used when better methods would not work.

2.3. **Measures.** Due to the small number of books provided, it was necessary to create as many measures as possible to get the most information out of the data. The measures created are outlined below, and their effectivenesses considered.

2.3.1. *Average Word Length.* The first measure created was the average word length of each book. This measure did not need to have any treatment to make sure it was independent of the length of the extracts, since by definition, taking the average accounts for that. The resulting data are shown in Figure 1. With the exception of the 1836 book, ”The Pickwick Papers”, the data do seem to follow a trend. The length of words seems to be low at 1840, rise in the middle of Dickens' life around 1852 and decrease to 1870. The word lengths do not seem to be randomly jumping up and down over time, but increase and decrease with multiple points in a row, suggesting that word length may be useful in predicting the year of publication.

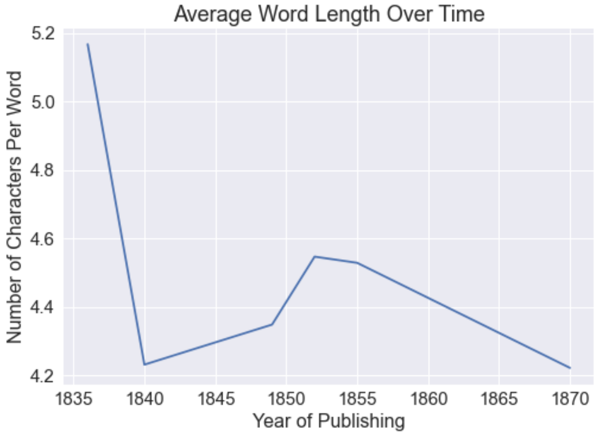

Figure 1

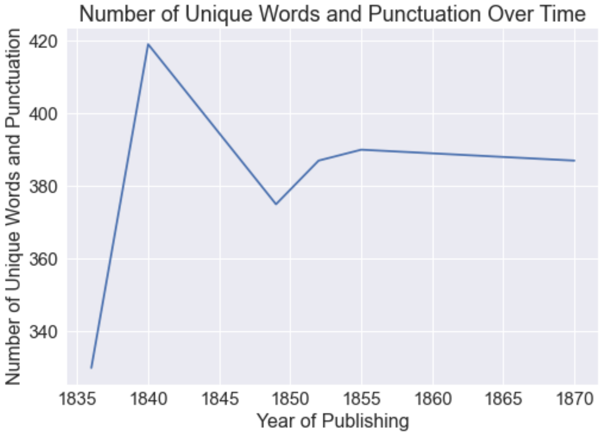

Figure 2

2.3.2. *Number of Unique Words and Punctuation.* The second measure was the number of unique words and punctuation marks from the first 4116 characters of each book. Since the extracts are of unequal lengths, using the whole extracts would result in inaccurate data as longer texts will have more words, so more opportunities to introduce new words and punctuation. Unfortunately, methods to combat this without reducing the size of the dataset, such as dividing the number of unique words by the length (number of characters) of the extract would result in a similar problem. As the number of words in a text increases, the number of unique words per character will decrease, because as each word is used, the next occurrence of that word is no longer unique. Therefore, for this measure, only the first 4116 characters of each extract were used, since this is the length of the shortest extract.

Figure 2 shows the number of unique words and punctuation marks for each book in chronological order. It shows large oscillations in the number of unique words and punctuation for the first three books, fluctuating from very low (∼330) to high (∼420) to low again (∼370) before settling at around 390 for the last three books. This may be useful in predicting the year: if a new book has ∼390 unique words/punctuation marks in the first 4116 characters, it is expected with this measure that the book was written after around 1850. In the same way, a new book with a different result for this measure may be a similar fluctuation to those before 1850. This is not certain though, and it could be that the first two books (1836 and 1840) are quite irregular, or that only the second book (1840) is an anomaly, which is also indicated by a smooth curve instead of Figure 2 if the 1840 book was removed. With more books, the answer to this should become more obvious.

2.3.3. *Punctuation Marks Per Character.* The next measure is punctuation marks per character. This has to be measured per character since there will be more punctuation marks the longer the text is. The plot in Figure 3 shows punctuation marks per character for each book in chronological order. It initially seems quite random, but if it assumed that the first book (1836) is anomalous and ignored, the graph becomes quite linear with one deviation at 1852. This suggests it could be used by an algorithm to predict the publishing date of one of Dickens' books. However, due to the lack of data, there are of course other possible explanations (including randomness), such as the possibility that the 1840 book is anomalous and the graph should start and end high, with a dip in the middle. In any case, it may be useful for machine learning.

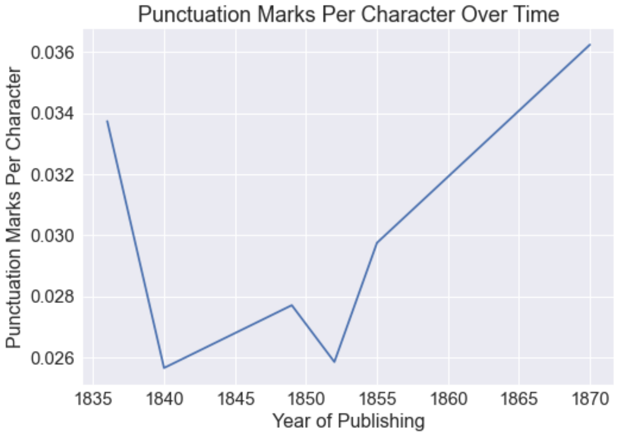

Figure 3

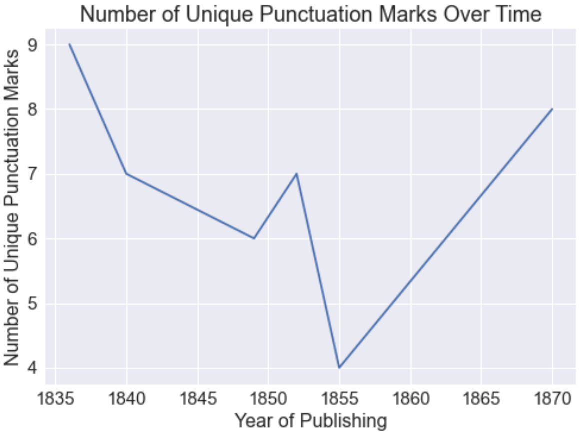

Figure 4

2.3.4. *Number of Unique Punctuation Marks.* The fourth measure created is the number of unique punctuation marks. As with the unique words and punctuation above, the extracts had to be limited to 4116 characters for this measure. Figure 4 shows the results of this measure for each book in chronological order. There is not an obvious trend in the graph, but this does not mean it is of no use for machine learning. There is also still a possibility of a trend excluding an anomaly, but it seems less likely.

2.3.5. *Number of Characters Per Sentence.* Another measure is the average sentence length. This is number of sentences in each extract, divided by the number of characters in the extract to be independent of extract length, resulting in the average number of characters in each sentence. Figure 5 shows how the average sentence length varies with time, which is clearly decreasing with some random noise. Given the limitations of the size of the dataset, this is very good, and it seems highly likely that this measure will be useful in predicting publication dates.

2.3.6. *Natural Language Toolkit Measures.* In order to draw out as much information as possible from the extracts, a Python library Natural Language Toolkit (NLTK) was used. NLTK allows for Part of Speech (POS) tagging, which determines a particular part of speech ”tag” for each word, depending on its meaning and context. NLTK has 36 possible tags[2], such as ”Noun, singular” or ”Adverb, comparative”. A measure was created for each tag divided by the length of the extract, resulting 36 new measures, some of which are shown in Figures 6, 7, 8 and 9. The measures seemed to have varying degrees of effectiveness, but given the need for data and the difficulty in identifying outliers in such a small dataset, all measures were kept for a total of 41 measures.

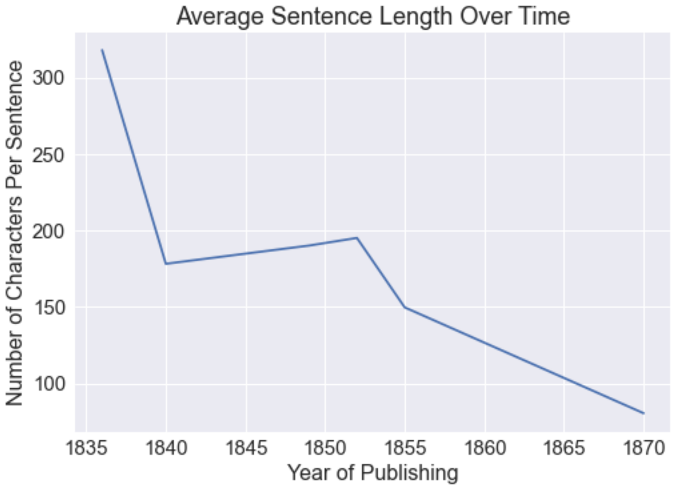

Figure 5

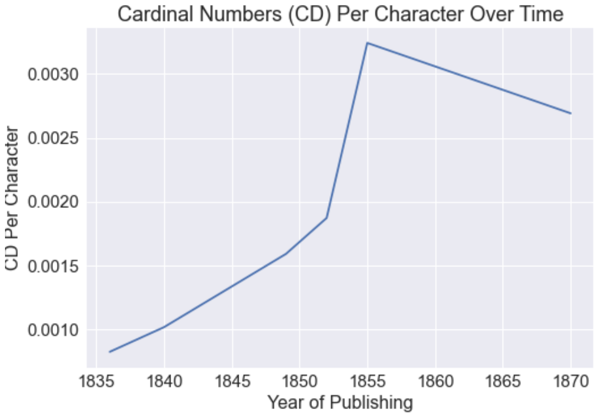

Figure 6

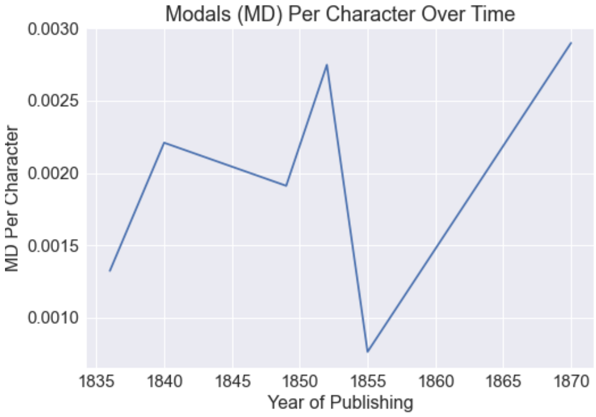

Figure 7

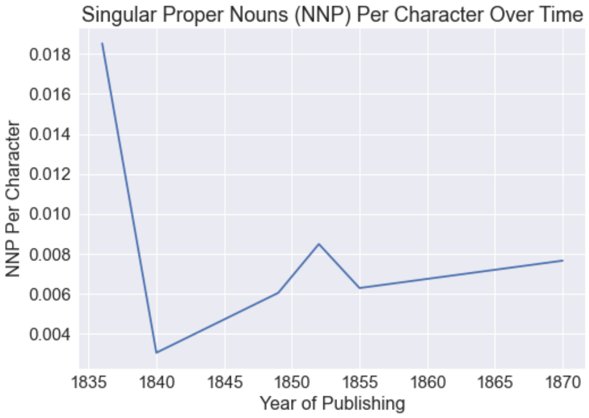

Figure 8

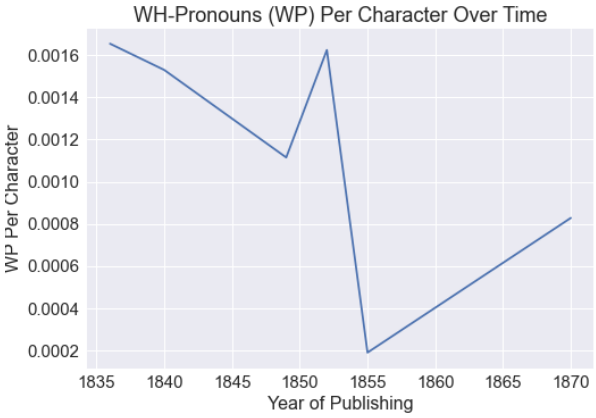

Figure 9

2.4. **Final Preprocessing.** After creating the measures, they were combined into a single dataframe of the measures, or features of each book. Each feature is also scaled to have zero mean and unit

Set	Title	Date	5-year	10-year	20-year
Training	The Pickwick Papers	1836	1	1	1
	The Old Curiosity Shop	1840	1	1	1
	David Copperfield	1849	3	2	1
	Bleak House	1852	4	2	2
	Little Dorrit	1855	4	3	2
	The Mystery of Edwin Drood	1870	7	4	2
Test	The Chimes	1844	2	1	1
	The Haunted Man and the Ghost's Bargain	1848	3	2	1
	A Tale of Two Cities	1859	5	3	2
	Great Expectations	1860	5	3	2

Table 1. The 8 books with known publication dates separated into categories by date.

variance, so they are all normally distributed and have the same variance, which is a requirement for a variety of machine learning algorithms.

Machine learning algorithms work best with lots of data, and some methods may not be at all effective with only 6 books to train on. Overall, it seems there are some trends over time - especially in sentence length. In other measures, there are also possible trends with an anomaly, so it appears that some machine learning methods could work.

3. Analysis

In this section, the effectiveness of a variety of machine learning algorithms will be tested in order to find the best methods to predict the two unknown publication dates of the books provided.

A limited amount of data greatly reduces the effectiveness when training machine learning models. To attempt to counteract this, the books were grouped into categories based on publication date, and classification algorithms were used to predict the category. The categories were 5 year groups, 10 year groups and 20 year groups (or before/after midnight 31/12/1850), as shown in Table 1. This meant the algorithms could be trained on multiple books in the same category, but with less precision. The three different categories were used since they mean it is possible to test the accuracy of each classifier at three levels of precision. The least precise category (20 years) should result in a higher accuracy, but if the classifier is good enough, it may also be accurate classifying into more precise categories.

3.1. **Linear Discriminant Analysis.** The first classification algorithm used was linear discriminant analysis (LDA). LDA is a classifier which fits a linear decision boundary in the form of hyperplanes (n-dimensional plane e.g. a straight line in 2D or a flat plane in 3D) to distinguish between at least two classes. It creates a linear combination of features to try to represent the output variable (similar to analysis of variance, ANOVA).

The LDA function was trained on the features from the six books with known publication dates as well as the categories that each book falls in. As an initial test, it was used to predict the publication dates of the data it was trained on. Due to the low number of these books, it was expected that the function should be able to do this fairly accurately, however it achieved 67% accuracy classifying the books into 5 year classes, 83% into 10 year classes, and 67% into 20 year classes. 83% accuracy into 10 year classes is not bad as only one of the six books was predicted the wrong class, however it is surprising that it was worse predicting 20 year classes, a supposedly easier task. This suggests the model is underfitted to the data, which is possibly due to the limitations of a linear decision boundary. The accuracy may be low for the 5 year classes for the same reason.

The LDA function was then used to predict the classes of the publication dates for the four books in the test set (for each category size), and those predictions were compared with the known correct date to assess its accuracy. The function was 0%, 25% and 75% accurate in classifying the test set into 5, 10 and 20 year classes respectively. Unfortunately, due to the size of the training set and the distribution of publication dates of the books in this set, for 5 year classification, there were no books in the training set in classes 2 or 5. This meant the function could not predict classes 2 or 5 for the test set, of which 3 of the 4 books were classes 2 or 5 (Table 1). Still, the function could have achieved 25% accuracy by correctly predicting the other book, but failed to do so. The 5 year classes will be removed from future tests. The 25% accuracy of the 10 year predictions is also poor, as randomly guessing classes would result in the same average accuracy. With twice as many classes as the 20 year category, it might be difficult to fit linear hyperplanes effectively to all the boundaries. The 75% accuracy for the 20 year classes is not bad, however it is not very precise as effectively it only predicts whether the book was published before or after 1850. It seems that this accuracy figure is good by chance since it performed worse than this on the training set, suggesting that the accuracy of the prediction would not be high consistently.

3.2. **Quadratic Discriminant Analysis.** Quadratic discriminant analysis, or QDA, is very similar to LDA but allows for a curved (quadratic) decision boundary. QDA could be more suited to this data than LDA, since it is less prone to underfitting with a more flexible decision boundary. However, for a small dataset such as this, overfitting is generally more likely than underfitting, and QDA is more likely to suffer from this than LDA.

QDA predicted the 20 year classes of the books it was trained on as well as the test books with 100% accuracy, which suggests it is an accurate and well-fitted model that could be used on new books. While it is a fairly imprecise prediction, its accuracy suggests it could at least be used to narrow down the range of dates a new book was published in. However, due to the way decision boundaries are generated, the training set must have more than one book per class, which rules out the use of the more precise 10 year category.

3.3. **Decision Tree Classifiers.** A decision tree classifier learns simple if-else decision rules from the features of the training data[3]. The maximum depth of the statements within the tree can be specified, the higher this is, the more closely fitted the model will be.

To create the best decision tree classifier possible, the optimum maximum depth was found by plotting each maximum depth 1-10 against the average of 500 tests of accuracy with that depth (Figure 10 shows this for the 10 year category). A decision tree classifier with maximum depth 2 was found to be most effective for predicting 10 year classes with an average accuracy of ∼32%. This is slightly above the average accuracy of a random guess (25%), but is not high enough to be useful in this context. For 20 year classes, a similar plot did not show any conclusive evidence for a specific maximum depth, and was inconsistent even when averaging across 2000 tests (Figure 11). However, the accuracy across different depths was quite consistent (deviation of $< 1\%$), so a maximum depth of 8 was chosen as it seemed to deviate slightly less on each test. With this, the decision tree classified the test books into 20 year classes with an accuracy of ∼63%, again, a little above the average accuracy of a random guess, 50% (higher due to the fewer classes).

3.4. **Decision Tree Regressors.** A decision tree regressor is very similar to the classifier above, but outputs a number - in this case representing year of publication - instead of a class. This was included with the aim of outputting a continuous variable predicting the specific year. However, this was not successful as the function attempted to output the year of the book from the training set that the test book was most similar to, rather than a new floating point prediction of the year

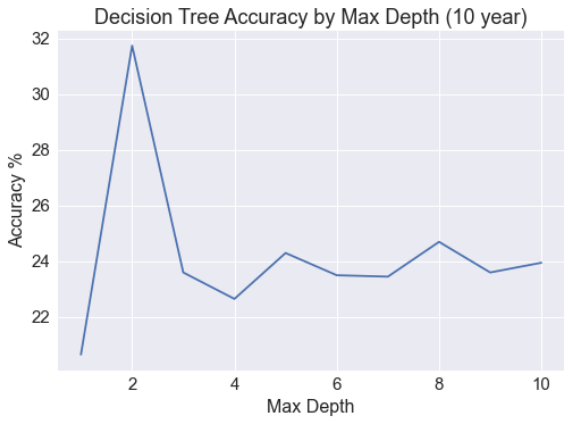

Figure 10

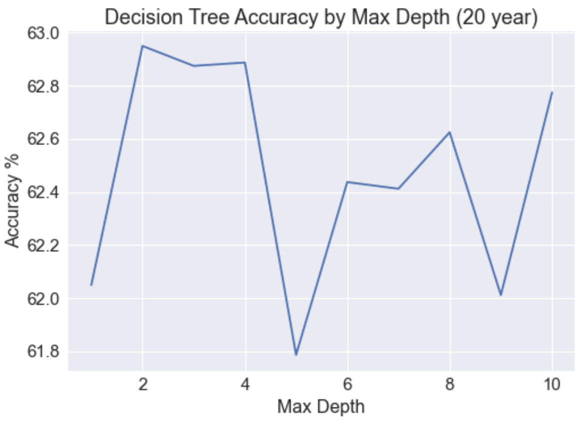

Figure 11

based on the training set. The accuracy of this function was measured by the average of the mean squared error (MSE) in years, given by:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

This was averaged across all four books in the test set from 1000 tests to give an average MSE of 112. This means, on average, each prediction was over 10 years out from the actual date, so this is not a very useful model.

3.5. **Random Forest Classifiers.** A random forest classifier uses a specifiable number of decision tree classifiers (estimators) constructed with some randomness and used on smaller parts of the dataset. It averages results from these classifiers to predict a class.

To find the best number of estimators for the random forest classifier, a similar method to that used in the decision tree section was used. The average accuracy of 50 tests of the classifier constructed with each multiple of 5 estimators up to 100 was plotted (Figures 12 and 13). It was found that classifying into 10 year classes was most accurate with a low number of estimators, so the value 20 was used. This achieved an average accuracy of only 26%, approximately the same as the average accuracy of a random guess. For 20 year classes, the accuracy of the classifier seemed to improve with the number of estimators until it plateaued. A random forest constructed with 100 estimators achieved an average accuracy of 73%, which means this method could be of some use, however more accurate and more precise methods have been discussed above.

3.6. **Random Forest Regressors.** Random forest regressors are very similar to random forest classifies, but they can output continuous variables. The optimum number of estimators was found to be 80 using the previous method but plotting MSE (in years) instead of percent accuracy. This regressor can output a floating point value that has not been previously included in the training set, so can be used to predict a specific year of publication after being trained. This resulted in an average mean square error of 38, meaning the model's predictions differed from the actual publication year by around 6 years on average. This is not amazing in a general context, but is perhaps surprisingly good for the limitations of the dataset, and is useful as it provides a specific predicted year.

3.7. **Further Improvements.** Feature importance is a measure of the relative importance of features based on their usefulness in making a prediction[4]. When using the decision tree classifier, the feature importance showed only one measure was being used. This indicated a problem, meaning it was possible the algorithm could be more effective. In the measure creation, some of

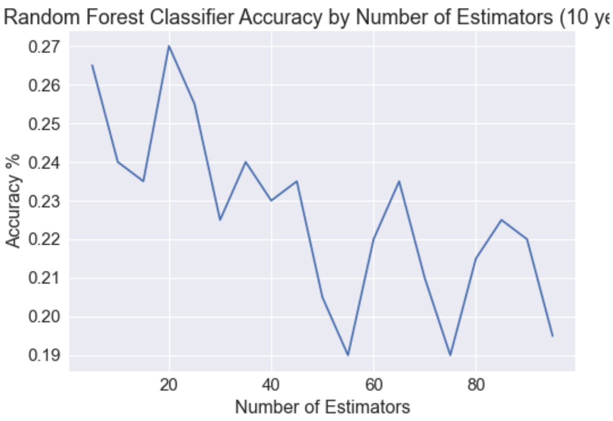

Figure 12

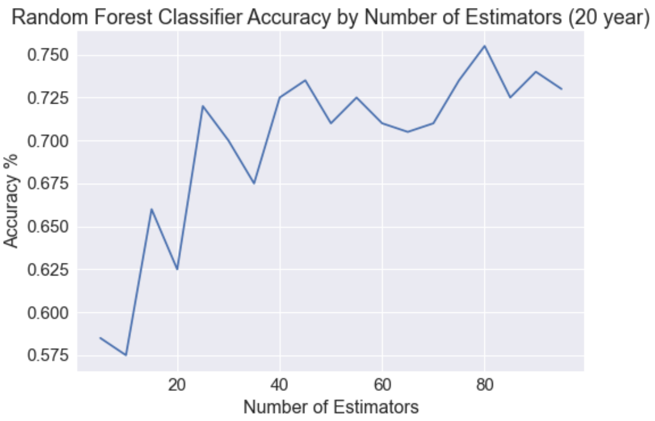

Figure 13

the NLTK features counted very few of certain types of words, possibly meaning it was picking up a lot of random noise. This could highly increase chances of algorithms overfitting, for example by fitting to measures that were just random noise. It is unlikely to be able to distinguish between noise and trends due to the small dataset. Rather than inspecting all 36 NLTK measures to see which ones were useful, the algorithms were attempted again using just the five other features. There were varying amounts of success between algorithms, but the most useful prediction to become more accurate would be from the random forest regression as it is the most precise.

Removing the NLTK measures reduced the average MSE of the random forest regression by 9, meaning the average error of 6.2 years was reduced to 5.4 years, so the NLTK measures were not included again. Some of the other algorithms were also made more effective, but in some the accuracy was reduced (Table 2). To reduce the MSE further, the average feature importances were calculated, as well as an average of feature importances that lead to an MSE of below ∼25. These were compared, and it was found that two of the five measures, the number of punctuation marks per character and number of unique punctuation marks were less important in the more accurate predictions (with lower MSE). These two measures were removed and the algorithms tested again (Table 2), and the random forest MSE was reduced to 22. This meant the average error in the prediction was reduced further from 5.4 to 4.7 years. To check whether this could be reduced any further, the same process was applied, which resulted in the ”number of unique words” measure being removed. This caused the MSE to reduce to 19, or in other words the average error reduced again from 4.7 to 4.4 years. The accuracy of the other algorithms (excluding the decision tree regressor) at each of these steps can be found in Table 2.

It is likely that some of the removed measures would be useful when used on a large amount of data. Due to the limitations in the size of the dataset, the algorithms are less able to distinguish the useful information from the noise in these variables.

3.8. **Summary.** Some of the most useful functions found, based on accuracy and precision are: QDA with 20 year classes and all features (100% accuracy), the decision tree and random forest classifiers with 10 year classes and only word length and sentence length (75% accuracy), and the random forest regressor with only word length and sentence length (mean squared error of 19). Other functions with slightly lower accuracies could be used to attempt to validate results.

	All	No NLTK	Num. Unique, Word + Sentence Len	Num. Unique, Word Len
LDA, Y10	25%	67%	25%	25%
LDA, Y20	75%	25%	75%	50%
QDA, Y20	100%	50%	50%	50%
TreeCLF, Y10	32%	34%	36%	75%
TreeCLF, Y20	63%	69%	70%	75%
RF CLF, Y10	26%	39%	29%	75%
RF CLF, Y20	73%	65%	75%	75%
RF REG, MSE	38	29	22	19

Table 2

4. Results

In this section, the trained algorithms were applied to the features of "Hard Times" and "Our Mutual Friend", the two books with unknown publication dates, in an attempt to predict the year of publication for each.

4.1. **Prediction.** Firstly, QDA with all features (100% accuracy on test set) predicted that both books were published before 1850. To check this, other algorithms with a slightly lower accuracy can be used. There were five classifiers that predicted the 20 year classes of the test set with 75% accuracy: LDA with all features, LDA with unique words, word length and sentence length, the decision tree classifier with unique words and word length, the random forest classifier with unique words, word length and sentence length, and the random forest classifier with unique words and word length. Only the decision tree also classified the two books as before 1850, and overall the results look randomly distributed. No predictions can be made from this.

There were two classifiers that were 75% accurate in predicting the 10 year classes of the test data. These were the decision tree and random forest classifiers with the features unique words and word length. Both of these predicted that the first book was published between 1846 and 1855, and the second book between 1866 and 1875 (effectively 1866-1870 since this is the year Dickens died). Combined with the previous results, it seems that the first book "Hard Times" was probably published around 1846-55, since that would explain the inaccuracy when attempting to place it before or after 1850. Given the two 10 year predictions for the second book, "Our Mutual Friend", are over 15 years later than 1850, it is likely that this book was published after 1850.

The random forest regressor achieved a minimum mean squared error of 19 using the features unique words and word length. It also achieved a mean squared error of 22 when also supplied with the sentence length measure. This means its predictions on the test set were an average of 4.4 and 4.7 years out, respectively. When applied to the two unknown books, the average prediction by the 19 MSE regressor of "Hard Times" and "Our Mutual Friend" was 1851 and 1862 respectively. For the 22 MSE regressor, this was 1858 and 1863 respectively. Given this and the previous results, it appears "Hard Times" was published in the 1850s, probably ∼1853-57. It appears "Our Mutual Friend" was published around the early 1860s, probably ∼1862-63.

4.2. **Accuracy.** The actual publication dates of "Hard Times" and "Our Mutual Friend" are 1854 and 1864, respectively. The accuracy of every algorithm was tested on the two books, and can be seen in Table 3. Generally, the algorithms achieved a lower accuracy than on the test set, with some exceptions. The inconsistency suggests many models had a bad fit to the data. QDA was notably poor, seeming to be overfitting and never correctly predicting that either book was published after 1850, which is especially surprising since it correctly predicted the 20 year classes of all four books in the test set when given all features. Fortunately, the most important and precise predictor, the random forest regressor, increased in accuracy substantially from its two best performances on the

	All	No NLTK	Num. Unique, Word + Sentence Len	Num. Unique, Word Len
LDA, Y10	50%	50%	50%	50%
LDA, Y20	50%	100%	50%	0%
QDA, Y20	0%	0%	0%	0%
TreeCLF, Y10	33%	21%	0%	50%
TreeCLF, Y20	51%	50%	53%	0%
RF CLF, Y10	28%	36%	33%	50%
RF CLF, Y20	60%	92%	99%	55%
RF REG, MSE	56	22	9.2	6.3

Table 3

test set, reducing the MSE from 22 and 19 to 9.2 and 6.3 respectively. This means there was an average difference of approximately 3 and 2.5 years respectively between the prediction and the correct year.

5. Conclusions

The preliminary tests of the algorithms found a variety of accuracies, and some algorithms showed promise, such as QDA. The algorithms were applied to the two unknown books and an overall prediction was made, between 1853 and 1857 for one book, and around 1862 and 1863 for the other book, which was very accurate. The real publication dates were 1854 and 1864, so this was close especially considering the limitation of the small dataset. The accuracy of many of the predictors, especially QDA, differed greatly between the test books and the books with unknown date, indicating a bad fit. The random forest regressor was the most useful predictor in this project, due not only to its accuracy but also its precision. In conclusion, using a variety of machine learning techniques proved successful in predicting the publication dates of two unknown books within the wider context of Dickens' literary works, which has shown that machine learning can be effective even in projects with smaller datasets.

6. Discussion and Further Study

Overall, the prediction was a success, and surprisingly accurate given the amount of data provided. The accuracy seems especially fortunate because the MSE of the final regressor prediction was lower than that of the regressor applied to the test set. It was expected that it would have the same or higher MSE, since the regressor was optimised using the test set data, not the books with unknown date. It is possible that the regressor was more accurate on the later books, explaining why the books with a variety of dates (1840-60) in the test set were predicted with lower accuracy than the other two books, with dates after 1850. Alternatively, it could be random.

There was a lot of noise in the classification predictions, which made the regressor even more useful. Had the regressor not been used, the prediction would have to be almost solely based on the two 10 year classifications. While this already results in a weak prediction, the two classifiers were a decision tree and random forest using the same features, so were closely linked. The random forest uses decision trees for its prediction, meaning the two identical predictions were not independent, further reducing their strength.

There were many limitations to overcome for this report, mainly due to the small number of books in the dataset. While it would not be possible to include more books than the number Dickens wrote, including more books for training as well as the entire books rather than extracts could increase the accuracy of predictions. There were also limitations in the accuracy of testing, since there were only four extracts from test books which had to be of the same form (short). There was also a ”blind spot” for testing, since there were no books available for testing that were published from 1849 to 1858. However, more test books that were published in other years could have been included to increase the reliability of tests.

The small size of the extracts also means there is more noise. Books that begin with speech tend to have very different punctuation, and the words used would depend on the characters in the book that are speaking. Having the whole book would average away some of this noise. The small number of books also means anomalies are not detectable, since slow changes over time seem more rapid, less predictable, and can look like random noise.

The algorithms should work best with consistent changes in measures over time, but that is very unlikely in this scenario, due to the data being written language. Any trends are especially likely to be inconsistent because of the small size of the dataset.

When classifying books into 10 year ranges, it would have been beneficial to denote the ranges starting with 1831-1840, since that would result in the final category ending in 1870, the year of Dickens' death. This would help to ensure each class has a similar date range. In the context of this report, this should have a very small impact due to the small size of the dataset. There are not many books in the training set, and they are not evenly distributed over time, so they cannot be evenly distributed between the classes.

Other measures could have been experimented with, such as counting of specific words. For example, it may be expected that a writer will use fewer ”weak words”[5] as they improve. Measures of specific punctuation marks could also be considered. The multitude of measures that were removed during the refining of the random forest regressor could be inspected more closely to determine whether any were useful. It is possible that a measure that is very descriptive of the publication year was disregarded.

Confidence intervals for the regressors and for the accuracy of classifiers would be very useful in determining the expected accuracy of a predictor when applied to new data. It may become clear that some of the classifiers have such a large confidence interval that they should be discarded, which would increase the ease of interpreting results.

More complex algorithms could be attempted, such as a neural network, however these are unlikely to be effective as it is especially important to have a larger dataset.

7. References

(1) https://www.gutenberg.org/

(2) https://pythonspot.com/nltk-speech-tagging/

(3) https://scikit-learn.org/stable/modules/tree.html#tree

(4) https://machinelearningmastery.com/calculate-feature-importance-with-python/

(5) https://mybookcave.com/authorpost/17-weak-words-to-avoid-in-your-writing/