Stockholms universitet

Full stop: a statistical analysis on the effect of stop words on the appreciation of classical literature

Morris Lundberg Allerholm

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2025:1
Matematisk statistik
Februari 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2025:1**
http://www.math.su.se

Full stop: a statistical analysis on the effect of stop words on the appreciation of classical literature

Morris Lundberg Allerholm*

February 2025

Abstract

This thesis explores the application of linear regression in computational text analysis. The subject of the analysis is the appreciation of literary classics and the primary details of interest are the frequencies of the most commonly used words in English: stop words. To develop the models, both the Lasso and best subset selection are implemented. Their evaluation reveals that there are connections between usage of stop words and general appreciation of classics. Some common words stand out as being especially important for the public's appreciation. The resulting models are also shown to have predictive ability.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: morrislual@gmail.com. Supervisor: Johannes Heiny and Xuechun Hu.

Acknowledgements

Thank you, Johannes Heiny and Xuechun Hu, for your guidance, without which this project would not have been possible.

This thesis has, in part, relied on the large language model ChatGPT for assistance with LaTeX and finding errors in R code.

Contents

1 Introduction and Background

Language is foundational for humans, yet there is limited knowledge regarding its mechanics and their effects on readers. Previously this has been explored by reasoning, but this is difficult; it requires language to be described by language: a paradoxical endeavor. More fruitful results have been gained in the scientific research on language and texts. An even later development in the exploration of texts and their impact on people has been the rise of computational text analysis. Because of the complex, detail-rich nature of language it is a fitting object for statistical studies. Through statistical methods one can obtain deeper insights.

The results of such studies attract diverse groups: interested readers, researchers, people working with literature and so on.

The main question at hand is whether or not we can describe and predict the average rating of books using their text. More specifically: is it fruitful to apply linear regression models to predict and describe the appreciation of classic literature by the use of textual factors, but also external ones. We rely on some linguistic measures, but primarily the frequencies of the most common words: stop words, that carry little meaning by themselves. These words can be seen as the skeleton of our language; they are the structure that carries the meaningful words, and in a sense they therefore define the shape of text.

To explore what these factors tell us about the appreciation of books we implement linear regression as well as Lasso regression. To decide which features the linear regression models should consist of best subset selection is used on different groups of explanatory variables, such that a smaller, final group is promoted on which best subset selection is applied again.

We find that different models show different tendencies but at large tell a similar story: general and external factors are prioritized. Many, but far from all, of the stop words that were originally considered carry great weight in the models. Leave one out cross validation and a smaller test on a separate data set indicate that the models have predictive power.

1.1 Digital humanities

As computational ability has developed, so has the interest in analyzing what would normally be subject to analysis in the humanities with statistical models. This young field of research is called digital humanities, and a part of it focuses on examining text with the tools of statistics.

The wealth of information, essential to this project and many similar projects, is freely available for all thanks to a few websites. Project Gutenberg and Wikisource are projects driven by volunteers that provide and endorse the sharing of ebooks in the public domain. To collect user interactions with books, such as ratings, we have relied on the website Goodreads, where users log and review their reading to share tips and to be inspired to read new books.

An early reference, that also relies on one of the mentioned websites, is the paper by Ganjigunte Ashok et al.[1]. It is a study that uses mainly quantitative data as opposed to qualitative data to analyze texts from many genres. At large they are able to classify successful and unsuccessful literature, based on a threshold in download counts from Project Gutenberg. They employ support vector machines, and achieve an accuracy of up to 84% when classifying books as successful/unsuccessful. The factors which they rely on for this classification are frequencies of words and pairs of words, distribution of word classes, frequencies of different grammatical constructions and sentiment analysis.

Subsequent work by Maharjan et al.[4] continues with a similar project. Unlike their predecessors they use the average rating from Goodreads as a measure of likeability, instead of download counts as a measure for success. They also separate the books into two groups, liked/disliked and attempt to develop models that classify accurately. The methods they use are support vector machines and neural networks, and the features are multiple regarding readability, sentiment analysis, general features of the text such as word- and character count, and words and phrases. The paper also implements a regression model for prediction. Notably they found some stop words to be important in their classification model.

1.2 Mathematical background of linear regression

In this project we use linear regression. The description of the method follows the framework and notation described by Hastie et al.[2], and is partly a summary of their work, especially chapter 3.2.

The quantity of interest is the average rating of books, denoted by Y_i for $i = 1, ..., N$, where N is the number of observations. We make a distinction between Y_i, which represents a random variable, and y_i which represents an observed value. Each y_i can theoretically take on values between 0 and 5, but the range of ratings observed in the data lies between 3.19 (The Castle of Otranto, by Walpole) and 4.48 (Martin Eden, by London).

This project aims to describe and predict the average rating, Y_i, using multiple quantitative factors, concerning the book and its author. Examples of these factors are the length of the book in characters, the gender of the author which is coded on a binary scale, the frequency of the word "it", and a linguistic measure called automated readability index. All of these p factors that adhere to the book with index i, $X_{i1}, ..., X_{ip}$, are together represented as a vector $\bar{X}_i = (X_{i1}, ..., X_{ip})^T$. Again, we differentiate between $\bar{X}_i$ and $\bar{x}_i = (x_{i1}, ..., x_{ip})$, because for $j = 1, ..., p$ X_{ij} is seen as a random variable, while x_{ij} is seen as an observed value.

As previously stated we wish to estimate the values of Y_i given $\bar{X}_i$. The estimate is therefore $E[Y_i|\bar{X}_i]$ and the key assumption for linear regression is that

$$E[Y_i|\bar{X}_i] = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j = f(\bar{X}_i). \tag{1}$$

This means that the expected value of Y_i given the value of $\bar{X}_i$ can be described as a linear function of $\bar{X}_i$, $f(\bar{X}_i)$. To account for the variance that is not described by this linear function an error term is added, which gives the full model:

$$Y_i = f(\bar{X}_i) + \epsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i. \tag{2}$$

Such that, given $\bar{X}_i$, all the randomness is contained in the error term ϵ_i which we assume is normally distributed with mean 0 and constant variance σ^2 for all $i = 1, ..., N$. This means that homoscedasticity – constant variance for all random variables defined by (2) – is assumed. To estimate the vector $\beta = (\beta_0, ..., \beta_p)^T$, given the data, least-squares approximation is used. The least-squares method is aptly named; it minimizes the residual sum of squares defined by

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2. \tag{3}$$

Where y_i denotes the i-th observed average rating and x_{ij}, for $j = 1, ..., p$, the observed values of the explanatory variables.

The design matrix $\mathbf{X}$ is defined by its i-th row being the vector $(1, \bar{x}_i^T) = (1, x_{i1}, ..., x_{ip})$. So $\mathbf{X}$ has N rows and $p + 1$ columns: one row for each observation and one column for each coefficient that is to be estimated. As a consequence of this definition, in addition to representing all observed average scores, response values, as $\mathbf{y} = (y_1, ..., y_N)^T$, it is possible to describe the residual sum of squares, as seen in (3) using matrix multiplication:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta. \tag{4}$$

Because of the fact that $\mathbf{y}$ and $\mathbf{X}$ consist of observed values the first term is a real value, the second and third terms are equal, so that they can be written together as $-2\beta^T\mathbf{X}^T\mathbf{y}$ and the last term is the product of two vectors and a symmetric matrix. This means that differentiating the terms in (4) with respect to β gives:

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta). \tag{5}$$

To ensure that the solution of the equation $\frac{\partial RSS(\beta)}{\partial \beta} = 0$ is a global minimum, by showing that the second derivative of $RSS(\beta)$ is positive definite, the derivative (5) is differentiated again:

$$\frac{\partial^2 RSS(\beta)}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}.$$

If $\mathbf{X}$ does not have linearly dependent columns, then it has full rank. To show that $\mathbf{X}^T\mathbf{X}$ is positive definite, consider a non-zero column vector $\mathbf{a}$ with $p + 1$ rows. The criterion for $\mathbf{X}^T\mathbf{X}$ to be positive definite is $\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} > 0$. We get

$$\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} = (\mathbf{X}\mathbf{a})^T(\mathbf{X}\mathbf{a}) \geq 0$$

which is simply the square of the Euclidean norm of the vector $\mathbf{Xa}$. What remains then is to show that it cannot be equal to zero, which can only be the case when $\mathbf{Xa} = \mathbf{0}$, but since $\mathbf{X}$ only has linearly independent columns its kernel is $\{\mathbf{0}\}$ and $\mathbf{a} \neq \mathbf{0}$ by definition, so that $\mathbf{Xa}$ cannot be equal to $\mathbf{0}$.

This shows that the matrix $\mathbf{X}^T\mathbf{X}$ is positive definite, guaranteeing its invertibility and that the solution to the optimization equation is a global minimum. The equation

$$-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

is solved for β, and the vector that solves it is denoted by $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{6}$$

The resulting vector in (6), $\hat{\beta}$, is the vector of estimated coefficients that minimizes the residual sum of squares. It is important that the design matrix, $\mathbf{X}$, has full rank, otherwise this solution is not unique. Because of this it is important to avoid having two ore more collinear explanatory variables.

By the Gauss-Markov theorem the least-squares approximation has the smallest variance among linear, unbiased estimators (Hastie et al.[2], page 51).

There are multiple measures of how good the model fits the data. The value of interest is the residual sum of squares used in the calculation above. Even if it is minimized by least-squares it might still be relatively large if the assumption that $E[Y_i|\bar{X}_i]$ is a linear function is faulty.

To examine this, by testing the significance of the model as a whole, and moreover the estimated coefficients by themselves, the vector $\hat{\beta}$, as defined by (6), is treated like a random variable. Because we are interested in $E[Y_i|\bar{X}_i]$, all $\bar{X}_i$:s are seen as given, such that the randomness of $\hat{\beta}$ is contained in the random vector $\mathbf{Y}$, defined by $\mathbf{Y} = (Y_1, ..., Y_N)$. We calculate that

$$\begin{aligned} Var(\hat{\beta}) &= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\mathbf{Y})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})((\mathbf{X}^T\mathbf{X})^{-1})^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

Where $\sigma^2 = Var(\epsilon_i)$ is the variance of Y_i when $\bar{X}_i$ is given, for all i:s, and the last step is due to $\mathbf{X}^T\mathbf{X}$ being symmetric, which means that its inverse is symmetric. To estimate this variance an estimate for σ^2 is needed: $\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

Because $\hat{\beta}$ is unbiased, and because $\hat{\beta}$ is considered to be a matrix multiplied with a multivariate normal vector, $\hat{\beta}$ is also a multivariate normal vector with the following distribution

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}). \tag{7}$$

To test if a single feature has a significant impact, a Z-score is calculated. Since the variance of $\hat{\beta}$ is given by the covariance matrix in (7) the variance for each estimated coefficient is $Var(\hat{\beta}_j) = \sigma^2 v_j$ where v_j is the element on the j-th row and the j-th column in $(\mathbf{X}^T\mathbf{X})^{-1}$. This means that $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$ such that under the null hypothesis for the individual coefficient, that $\beta_j = 0$, $\hat{\beta}_j \sim N(0, \sigma^2 v_j)$. Dividing the coefficient by its standard deviation gives a fraction that has a standard normal distribution. And replacing σ with $\hat{\sigma}$ creates the Z-score:

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}. \tag{8}$$

Importantly, z_j follows a t-distribution, $z_j \sim t(N-p-1)$ where N denotes the number of data points and p the number of explanatory variables. This is used to get the p-value of the above null hypothesis tested against the alternative hypothesis that $\beta_j \neq 0$.

It is also of interest to test the whole model. For this the F-statistic is used. It is calculated as a comparison between the residual sum of squares for the model under the null hypothesis, RSS_0, where all coefficients $\beta_1, ..., \beta_p$ are assumed to be zero, and the residual sum of squares of the alternative model where all coefficients are assumed to be non-zero, RSS_1. For this set of hypotheses it is defined by

$$F = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(N-p-1)}. \tag{9}$$

The F-statistic follows the F-distribution, $F \sim F(p, N-p-1)$. This value is used to test if the entirety of the model is a significant result given the data,

or if it is probable to observe the values given a model that only includes the intercept.

2 Data

In this section the data and how it is processed before the modeling is described.

2.1 Data background

The books have been downloaded as text files from Wikisource and Project Gutenberg and are all freely available literary classics, whose authors have passed away at least 100 years ago. This is the criterion for a book to be in the public domain in the United States, which guarantees availability of books. It also ensures that the compared literature occupies a similar space in the culture, as to minimize differences in factors which are hard to account for. It should be noted that the word "classics" in "literary classics" is not used in accordance with any theory or canon from the humanities. Here it refers to the book being sufficiently old and sufficiently popular. The latter is measured in the number of ratings on Goodreads, with a lower bound of 10,000.

	title	author	gender	score	year	popularity	lang
26	Anna_Karenina	Tolstoy	0	4.09	1878	1728285	0
27	Scarlet_Letter	Hawthorne	0	3.43	1850	1030417	1
28	Don_Quixote	Cervantes	0	3.90	1605	338205	0
29	The_Three_Musketeers	Dumas	0	4.09	1844	1560773	0
30	The_Adventures_of_Tom_Sawyer	Twain	0	3.92	1876	2793915	1
31	Mansfield_Park	Austen	1	3.86	1814	8475100	1
32	Moby_Dick	Melville	0	3.55	1851	735981	1

Figure 1: sample of books with the properties gender of author, average rating ("score"), year of first publication, author's total number of ratings on Goodreads ("popularity") and if the work was originally written in English or not.

General information about the books and their authors has been collected on Goodreads and Wikisource. This includes gender, which year the book was first published, who the author is, if it was originally written in English or if it has been translated, and the metrics based on numbers found on

Goodreads. A small sample of this data is displayed in figure 1, a larger list is included in the appendix 6.1. The sample seen in figure 1 is typical for the data in general. Most of the literature in the data comes from the 19th century, there are more male authors than female authors and the popularity of the authors vary greatly. The main metrics gathered online are average score and number of ratings the author has on all their books. The number of ratings serves as the measure of the author's popularity shown in figure 1. This is the only way the authors are identified in the quantitative data. Another method of identifying authors would have been to add one binary variable for each author, mapped onto 1 if the book was written by that author and 0 otherwise, but this would have required too many variables.

There are further restrictions regarding properties of the books used in this project: the exclusion of poetry, plays, children's literature and books that are a part of a series. There are no exact definitions of what constitutes these types, so the decision has been made for each book by itself. Some books in the data set might be considered exceptions: for example Mark Twain's two famous *The adventures of...* may be considered as being both children's literature and as being part of a series. They are still included because they follow different characters, are readable in any order and do not follow the standard progression of a series, where the first book is the most popular.

It should also be noted that the books that make up the data are almost exclusively from western countries and part of a western tradition. This is a consequence of the analysis being of works in the English language and measuring popularity by the ratings of a primarily English-spoken community.

Plays and poetry are strictly barred. This work only analyzes prose. There are papers on distant reading (quantitative text analysis) of poetry, like the forementioned paper by Ganjigunte Ashok et al.[1]. They also set a restraint of a maximum of two books written by the same author in their training data, to alleviate the problems of the model favoring writing styles of authors who might be popular for reasons outside the text. The data used here has no such limit and many of the authors have multiple books. Instead their popularity is taken into account, as mentioned above.

2.2 Processing

For the general information about the books, such as gender, popularity, if it was originally written in English or if it has been translated and so on, all numeric values are kept as they are, while the non numeric variables are

transformed into ones and zeroes. Categorical data is kept on a binary scale as to avoid the difficulties of handling nominal data in linear regression.

To make the books readable by programs, Regex in R has been used to transform them into a vector consisting of all words in lowercase as string elements and a vector of all punctuation marks, also as string elements.

The bigger task in the processing of data has been the transformation of text into quantitative information. In this process the main challange has been restriction; there are too many words and patterns in language for us to be able to consider them all.

This analysis focuses especially on stop words, which here is used to refer to words that carry little information by themselves and are the words that are most commonly used. To decide which stop words that are to be considered we have selected some from the list of most common words in the Oxford English Corpus (Wikimedia Foundation n.d. [10]). In the data set these are included as frequencies: the number of uses devided by the total number of words in the book.

It is reasonable to believe that how stop words are used has a great impact on the experience provided by the book. Examples of stop words are "and", "to", and "you". Some tangential papers remove stop words entirely, Schmidt et al.[9], but there are elements of stop words which seem to carry great weight in computational text analysis. Mohseni et al.[7] summarizes the importance of pronouns and shows that the distribution of "I" and "you", words usually being more frequently used in dialogue than outside dialogue, and of "she", "he" and "they" are telling signs of a text being fictional or non fictional. They are thus informative about textual structure.

Another use of the Oxford English Corpus has been assisting the choice of verbs to account for. The distribution of word classes, and especially verbs, has been shown to have a great impact in classification of successful/unsuccessful literature by Ganjigunte Ashok et al.[1] (p. 1757). It is in general harder to predict how these distributions develop given a previous part of the text in fictional classics than in fiction that has not achieved the status of a classic as shown by Mohseni et al.[7] (p. 11). Because of this project's special focus on stop words the frequencies of the verbs, including their non-archaic conjugations, on the top 100 list of commonly used words in the Oxford English Corpus are included.

Other, more general textual factors that have been developed are the number of characters as a measure of length, number of words that occur less than 10 times, and number of unique words. All of these turn out to be somewhat correlated. The absolute value of the Pearson's correlation coefficient is

above 0.8 for each of the three possible pairs, which can be explained by the fact that if a book is long there are more opportunities to use rare and unique words. To avoid problems with collinear variables only the length of the book in number of characters is carried forward into the statistical modeling. As mentioned in section 1.2 there are multiple legitimate least-squares estimators for β when the design matrix does not have full rank, Hastie et al.[2].

The number of unique words is also a problematic measure because some books have an inflated number of unique words. As an example there is one dramatic outlier: *Ulysses* by Joyce. It has approximately 29000 different words. Which is about 10000 more than *War and Peace* by Tolstoy (translated by Louise and Aylmer Maude), a book that is more than double the length of *Ulysses*, measured in characters. The reason for an inflated number of unique words can sometimes be attributed to the author writing conversations in dialect, including passages in other languages or, which is partly a reason for the large number of different words in *Ulysses*, the author inventing new words.

In an attempt to measure readability the Automated Readability Index was calculated for each book using the formula

$$4.71\frac{\#char.}{\#words} + 0.5\frac{\#words}{\#sent.} - 21.43 \tag{10}$$

where "#" signifies "number of". A high value indicates that the text is difficult to read, either because of long words, long sentences or a combination of both. The coefficients are in place to balance the values of the quotients, and the subtracted value puts most regular texts on an interpretable scale, usually between 1 and 14, which is of lesser importance to this project. It is one of the simpler measures of readability according to Pitler et al.[8] (p. 187), which in part is why it is interesting; if it carries an important role in the statistical analysis it shows clearly what factors are important, and if it does not it shows these factors to be of less importance.

Lastly, for the processing of words, the NRC emotion lexicon [5, 6] is used to map words onto the emotions anger, anticipation, disgust, fear, joy, sadness, surprise and trust. If the word is associated with an emotion it has the value one, otherwise zero. Note that a single word can be mapped to multiple emotions. This is done for each word, then the results are summed and divided by the number of words to give the frequency of words that are connected to a certain emotion for each book.

In their original form the explanatory variables differ greatly in value. Popularity is measured in thousands to millions ratings, while the frequencies are all on the interval $[0, 1]$. This is especially problematic because one of the methods that is implemented is the Lasso, which involves setting a maximum allowed size of estimated coefficients. If we would try to apply the Lasso on the data as it originally was, then the method would favor explanatory variables with larger values because they generally have smaller coefficients. To counteract this all explanatory variables are standardized, by subtracting each value with the mean of the variable which it belongs to and then dividing the difference with an estimate of the standard deviation.

2.3 Data overview

The resulting data set has 55 variables, of which 54 are explanatory, which is a little less than a third of the number of books used to train the models: 184. This stands as a challenge to the application of models; there are preferably more observations in relationship to the number of variables in the training data. There is also a smaller set of test data, consisting of 15 books. This set is not large enough to conduct a proper test, but is used to corroborate the findings of other testing methods.

To get an overview of the data a few plots of some important variables are displayed. The first one, figure 2, shows the values of the response variable in the data. For those interested: the three highest rated books in the data are *Martin Eden* by London, *The Brothers Karamazov* by Dostoevsky, and *The Count of Monte Cristo* by Dumas. While the three lowest rated books are *The Castle of Otranto* by Walpole, *Fanny Hill* by Cleland, and *The Red Badge of Courage* by Crane.

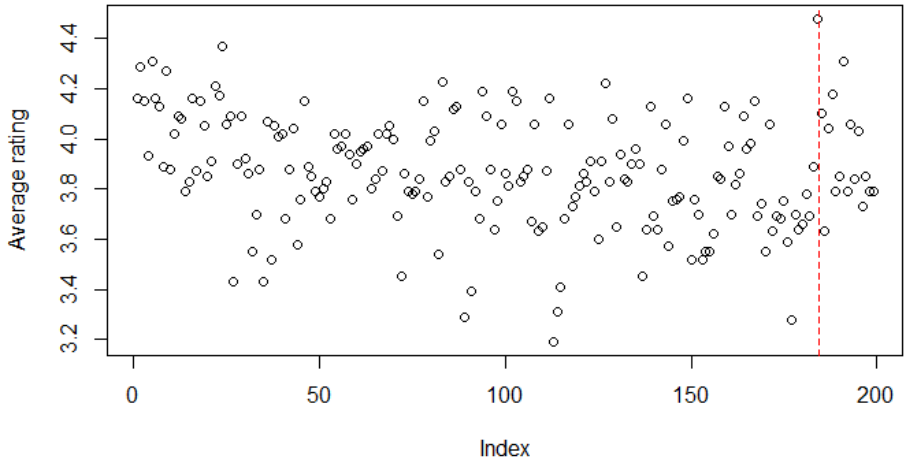

Figure 2: average rating of the books over index, with a red line to indicate where test data begins.

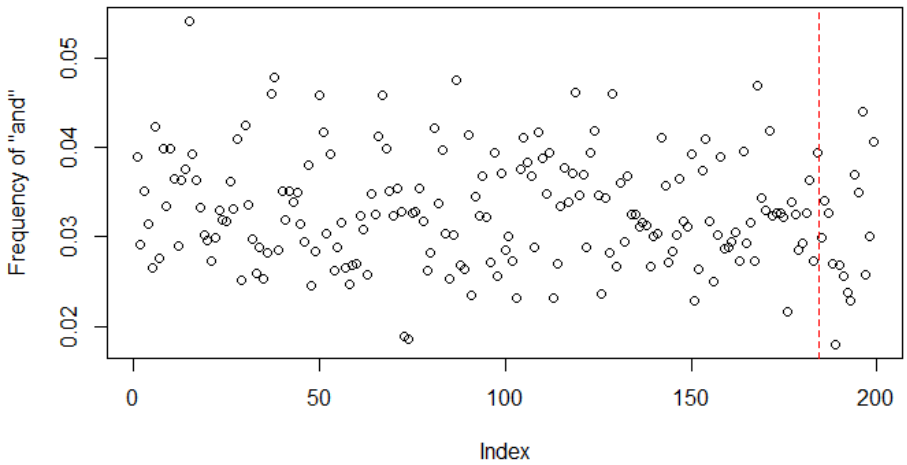

Figure 3: frequency of the word "and" over index, with a red line to indicate where test data begins.

In the end of figure 2 there is a red line to indicate where the test data begins. As seen in the figure the test data seems to have generally higher ratings than the training data. The two data sets' respective means are approximately 3.86 and 3.92, confirming that the testing data has a tendency of higher ratings. Another thing we may note is that there are no dramatic outliers in the testing data, but there is one book that shares the rating of Dumas' *The Count of Monte Cristo*, that is *The Blue Castle* by Montgomery, which is one of the highest values in the training data.

To show an example of what the word frequencies may look like we display the unstandardized values of the frequency of the word "and" in figure 3.

As seen in figure 4 the variables generally have low correlation between each other. A little less than 1% of the possible pairs have a Pearson correlation coefficient with an absolute value above 0.7. And of these 12 pairs 10 are between different NRC emotion frequencies. In the heat map these NRC-pairs mark the almost checkered area at bottom left. The last two high correlation pairs are between the sum of full stops, exclamation marks and question marks with *ARI*, which is reasonable because *ARI* depends on the number of sentences. There is also high correlation between the verbs "come" and "go", including their conjugations.

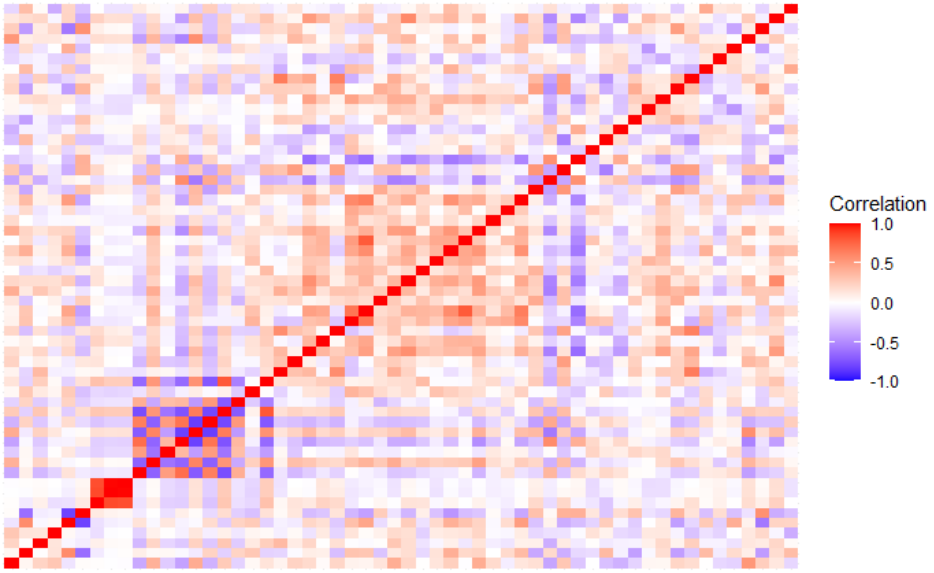

Figure 4: Heat map of all of the variables in the processed data. A stronger color indicates a higher absolute value of the Pearson correlation coefficient.

Pairs that have a correlation coefficient with an absolute value above 0.5 are also somewhat rare; approximately 5% of the pairs meet this criterion. Some of these might be expected, like the use of "her" and the author being female. Pairs of words at this level of correlation often form normal phrases: the verbs "will" and "be", "come" and "get", "to" and "have", and so on. Other pairs are seemingly random, while a few are interesting; as an example ARI has a negative correlation with year, meaning that texts have become more readable over time by that measure. Another interesting pair is between ARI and the word "as"; "as" is often used to begin a subordinate clause and therefore has a positive correlation with ARI: increasing the difficulty of the text. Other than that the vast majority, about 80% of the pairs, have a correlation coefficient with an absolute value below 0.3.

3 Statistical Modeling

First we get an overview of the explanatory variables from the perspective of linear regression. Before attempting to construct finalized models it is necessary to address and overcome a challenge in the data: there are few observations compared to the number of explanatory variables. To this end, we apply best subset selection and Lasso regression.

Finally, the models are evaluated. This is done by leave one out cross validation and Bayesian information criterion. In addition the models are applied on a small test set.

3.1 Best subset selection

Best subset selection is, Hastie et al.[2] (p. 57), a method that determines the best subset of explanatory variables that consists of k variables, for each $1 \leq k \leq p$, where p is the total number of variables. The models of the different subsets are compared by their residual sum of squares.

The number of variables is also a challenge for best subset selection. A set with n variables has $2^n - 1$ subsets that are non-empty, which means that the number of models that have to be estimated grows exponentially as the number of variables increases. Best subset selection is implemented by the use of the leaps and bounds algorithm, that does not perform the calculation for each subset, and in theory allows for up to 40 variables according to Hastie et al.[2] (p. 57). With the computational power available for this project, that number is realistically smaller.

To work around this limitation the variables are split into groups: general and external factors, non-verb stop words, verb stop words, and emotion lexicon variables. The method is not applied on the general factors by themselves, but on the general factors in combination with all of the latter three groups. It is supposed, as a working hypothesis, that the general properties have an overarching importance that is separate from the other factors.

The results are then compared using the Bayesian Information Criterion (BIC), which is provided by the method used to implement best subset selection: the package "leaps" in R. BIC is defined by the formula

$$BIC = -2\log(L) + \log(N)(p+1).$$

Where L is the likelihood of the model, N is the number of observations, and $p+1$ is the total number of estimated coefficients, which includes the intercept. To achieve a final group of approximately 20 variables, the variables that are present in one of the six differently sized best subsets with the smallest BIC in each group are carried forward.

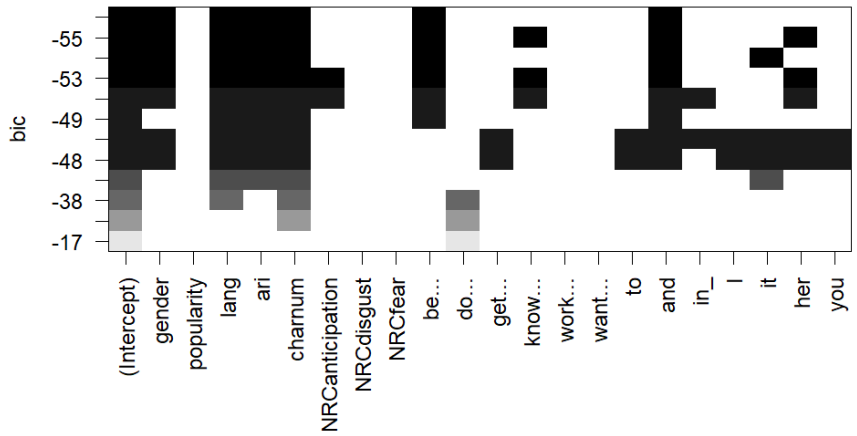

Figure 5: Plot of the best subsets in the final group containing 1 to 12 variables, each row represents a model where a colored box indicates the presence of a explanatory variable, they are ordered by BIC instead of size. The color grading is there to help appreciate the bic-scale and does not provide additional information.

Figure 5 displays all of the best subset models with 1 to 12 explanatory variables (not counting intercept) from the final group ranked by their *BIC* values. The colored tiles signify that a explanatory variable is included in that model. The assumption that the general factors are the most important to the model seems to be well founded; all of the best 5 models, measured in *BIC*, include gender, if it was originally written in English or not ("lang"), the automated readability index ("ari") and the number of characters in the book ("charnum"), but not popularity. Beyond that it is noted that many of the variables associated with the text hop in and out of the top models. Measured by BIC the best model only contains 6 variables: the general factors mentioned, and the frequencies of word "be" with conjugations as well as "and".

Accounting for the fact that BIC favors simplicity by adding the natural logarithm of the number of observations times the number of variables, n, in this case $n \cdot ln(184) \approx 5.22n$, models with more variables such as the one with 10 that has a *BIC* of 50, seem to fare comparatively well when simplicity is not a priority.

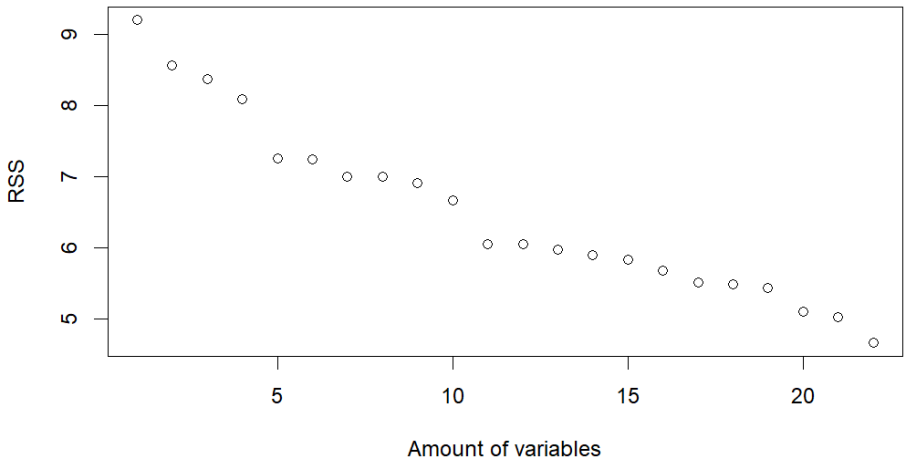

Figure 6: plot of residual sum of squares of all the best subsets from the final group.

This model includes the general factors mentioned above, and in addition the frequencies of words associated with anticipation, how often the verbs "be", "know" and their conjugations occur, and the frequencies of the stop

words “and”, “in”, and “her”. Figure 6 shows that the residual sum of squares decreases with a jump from 4 to 5 and from 10 to 11 variables. The latter jump marks a great shift in the model; models with 5 to 10 variables are based on a variety from all groups, where “be” and “and” dominate with their presence, while the models with 11 and 12 mainly contain stop words. This trend continues beyond what figure 5 shows; from 12 to 14 variables, verb stop words are incorporated.

With a few exceptions all of the variables in all of the above mentioned models are significant at the 5%-level. The model with 12 variables is the largest to have all its variables be significant. Also, all of these models are significant at the 5%-level; the highest p-value in the F-tests is $2.213 \cdot 10^{-15}$.

3.2 Lasso

Another apt method is the Lasso, described by Hastie et al.[2] (p.68). The reason for choosing Lasso over for example Ridge regression is that it, in certain situations, forces coefficients to be zero, and thus lessen the number of variables that the model depends on. It therefore functions like a combination of a subset selection method and a method that shrinks coefficients.

The Lasso minimizes the same residual sum of squares as the regular least-squares method, described in (2), but the Lasso coefficients have an additional constraint:

$$\sum_{j=1}^{p} |\beta_j| \leq t, \tag{11}$$

for some constant $t > 0$. This imposes a penalty on the L^1-norm of β in the linear regression model. Because of the slopes most often being smaller than in the results produced by the least-squares method, the Lasso model generally gives smaller changes in the predicted response variable for the same change in explanatory variables. Variance is thus reduced at the cost of introducing some bias, which is the result of the model fitting the data in a less strict manner. A smaller t results in a lower roof for the allowed sum of slopes and thus makes the model even less sensitive to the explanatory variables, which might lower variance even further.

A special property of the Lasso is that it, in contrast to Ridge regression, often sets coefficients to 0. Because there are many explanatory variables all of them are probably not relevant in a predictive model. But if many of the coefficients are non-zero in the model, it suggests that some of them are

useful for describing and predicting the response variable. It is harder to specify which individual coefficients are important in a Lasso model. Some coefficients might only be relevant in the context of the penalty on the L^1-norm of β.

The two simultaneous requirements, that of minimizing (3) as well as adhering to (11), can be summarized in what is called a Lagrangian form:

$$\hat{\beta}_{\lambda}^{Lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{12}$$

This formulation is equivalent to the two previous criterion. To find the best λ-value 10-fold cross validation is used in R via the package *glmnet*, which also serves the purpose of minimizing the above sum to find $\hat{\beta}_{\lambda}^{Lasso}$. 10-fold cross validation works by separating the data randomly into 10 parts, that are of the same size, then by training them on the remaining 9 parts for each part, and calculating the mean squared error of the trained model on that single part. The mean squared errors are summed and divided by 10, leading to the resulting estimate, as described by Hastie et al.[2] (p. 242).

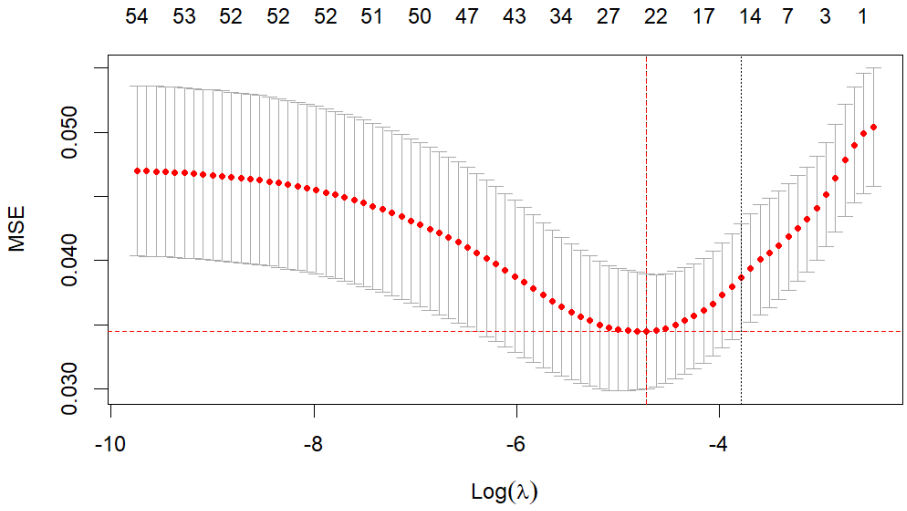

Figure 7: Mean squared error of Lasso-model over the logarithm of λ, the numbers on top of the figure tell how many non-zero variables there are in the model, and the intersect of the red lines indicates the λ that gives the minimal MSE.

Figure 7 shows the results of the 10-fold cross validation, where the minimal MSE (the residual sum of squares divided by the number of observation) is the product of the λ at the intersection of the two red lines. The resulting model has 26 explanatory variables, and thus depends on less than half of the original variables.

3.3 Evaluation of models

Before we evaluate the models' qualities, the assumptions of the standard linear regression models are revisited, which was formulated in (2). We inspect the residuals of the models; under the assumptions regarding the error term in (2) they are the realizations of the error terms and should therefore follow a normal distribution with constant variance and should not follow any pattern, which would indicate dependence.

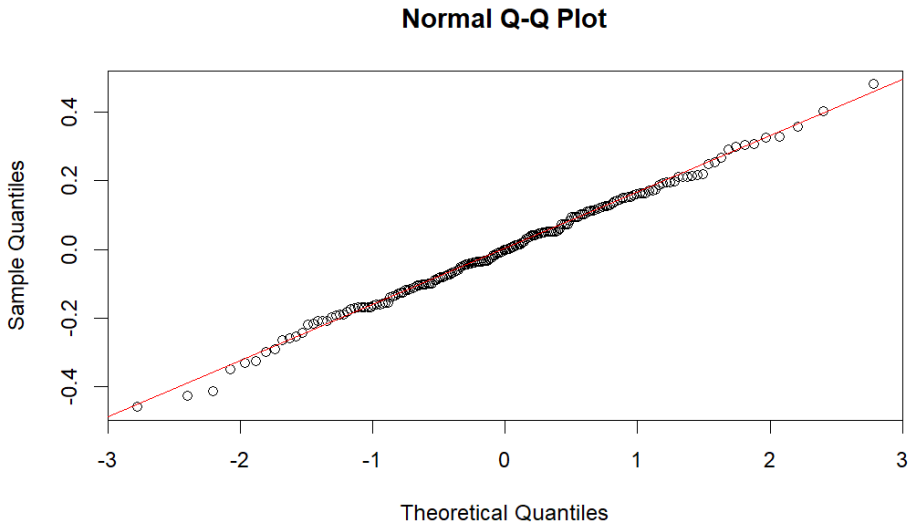

Figure 8: QQ-plot of the residuals of the best subset model with 12 variables. The red line indicates how the sample quantiles should lie in relation to the theoretical quantiles of the normal distribution. These residuals are plotted over index in the appendix 6.2, in figure 16.

As can be seen in figure 8, which shows a QQ plot for the best subset model with 12 variables, the quantiles seem to line up, and the observations follow the guiding line. The tails deviate slightly but there is no apparent curvature away from the line.

QQ-plots have been produced for the models with 5 to 15 variables and in general the smaller models, with less variables, do not follow the line as well as the larger models; chunks of observations deviate, which could be an indication that there is variance which is not taken into consideration. The model with 12 variables is the first one which is stable around the line. Moreover the plots of the residuals over index, which have also been produced but are omitted here (some are included in the appendix 6.2), show almost no sign of dependency; it looks like white noise.

Because of the forementioned problems regarding lack of data, the application of the models on a test set is only done to support or challenge the results of other evaluation methods. One of these, which has already been introduced, is *BIC*. The results of this is described as a way to compare the best subset selection models. What remains for these models is the discussion about their coefficient of determination, R^2.

R^2 measures how much of the variance in the data can be accounted for by the explanatory variables. It is defined by

$$R^2 = 1 - \frac{RSS(\hat{\beta})}{SS_{total}},$$

where $SS_{total} = \sum_{i=1}^{N}(y_i - \frac{1}{N}\sum_{j=1}^{N} y_j)^2$ – the sum of squared differences between the observations and their mean, James et al.[3]. Of greater interest to this analysis is the adjusted.

$$R^2_{adj.} = 1 - \frac{RSS(\hat{\beta})/(N-p-1)}{SS_{total}/(N-1)}, \tag{13}$$

The division of the sum of squares with the models respective degrees of freedom is done in an attempt to counteract the fact that more variables will always increase the standard R^2.

Variables	9	10	11	12	13	14	15
$R^2_{adj.}$	0.4068	0.4113	0.4172	0.4317	0.4359	0.4389	0.4414

Table 1: number of variables and corresponding $R^2_{adj.}$

The models from best subset selection with 5 to 15 explanatory variables cover a range of $R^2_{adj.}$ from 0.3343 to 0.4414. These values grow as more variables are added. The three largest subsets have the highest $R^2_{adj.}$ as can

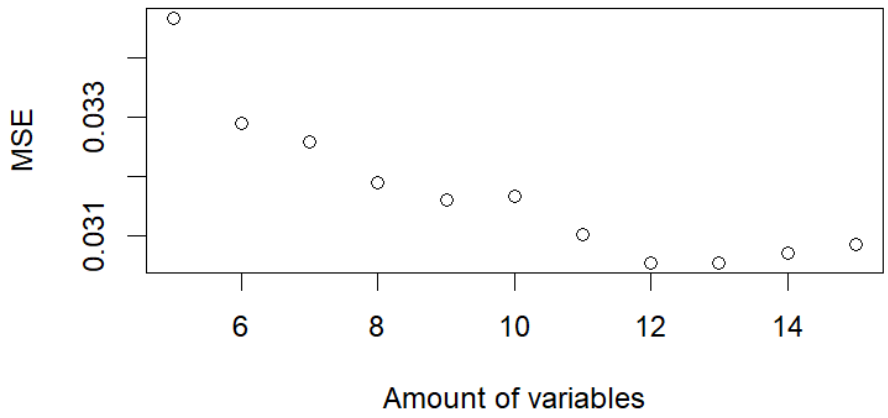

Figure 9: MSE from LOOCV for best subset selection models.

be seen in table 1 but all of them include variables that are not significant in the model. With 12 variables all of them are significant, and the model marks a jump in the value of $R^2_{adj.}$.

These large models give greater possibility of interpretation, while the smaller models, the ones with the best *BIC*, might be more useful for prediction.

To measure the prediction capabilities of a model without a testing- or validation set we use leave one out cross validation (LOOCV), which results in an almost unbiased estimate of the theoretical expected prediction error, but with high variance – Hastie et al.[2] (p. 242). It works in the same way as 10-fold cross validation, but it divides the training set into N, the number of observations, groups instead of 10. It tries to predict the value of each data point based on a model trained on all other data points, the errors are summed and divided by N. The results for the best subsets from 5 to 15 variables are shown if figure 9. The MSE reaches its lowest point at 13 variables.

It turns out that all of these have a lower MSE value than the Lasso model, which has a MSE of approximately 0.035.

The final step in the evaluation of the models is a prediction test on real books. For this an additional 15 books and their properties have been collected and processed in the same way as the training data. The values have

been standardized using the mean and estimated standard deviation from the training data.

Since we choose to have relatively few observations in the test set, approximately 7.5% of the total data, it might not be representative of the models actual predictive power. There is no exact measure of how much of the total data should be allocated to the test set, but 25% is the suggestion of Hastie et al.[2] (p. 222). And as they also mention: the ideal order of data collection is to collect all of it at the same time, randomly select your test set and train on the remaining data. For this project's test the data has been collected after the fact.

As mentioned previously this test is included primarily to challenge or corroborate the results of leave one out cross validation, BIC and of $R^2_{adj.}$.

The models that have been carried forward to this test are the Lasso model and the best subset selection model with 12 variables. The latter is chosen because it has performed well across all measures and balances the interests of simplicity, relevance and performance. Figure 10 displays the real and the predicted values of this model for the test data. The mean squared error is approximately 0.0405, which is a bit higher than the mean squared error produced by the leave one out cross validation, but still in the same range of values.

Figure 11 shows the same thing but for the predicted values that the Lasso model gives. The mean squared error of these predictions are approximately 0.036, a tiny bit higher than the value from the leave one out cross validation.

When comparing figures 10 and 11 the similarities are striking; the two models almost predict the same values, with a few exceptions: books 8 to 10 and 13 to 15 have similar prediction error sizes but differ in direction. Other than that all predictions are almost equal between the two models.

The mean squared errors measured in this test, and in leave one out cross validation for both models, are all lower values than that achieved by Maharjan et al.[4] and their regression model. But it should be noted that their test set constituted 30% of a total of 800 observations, which makes their finding more stable.

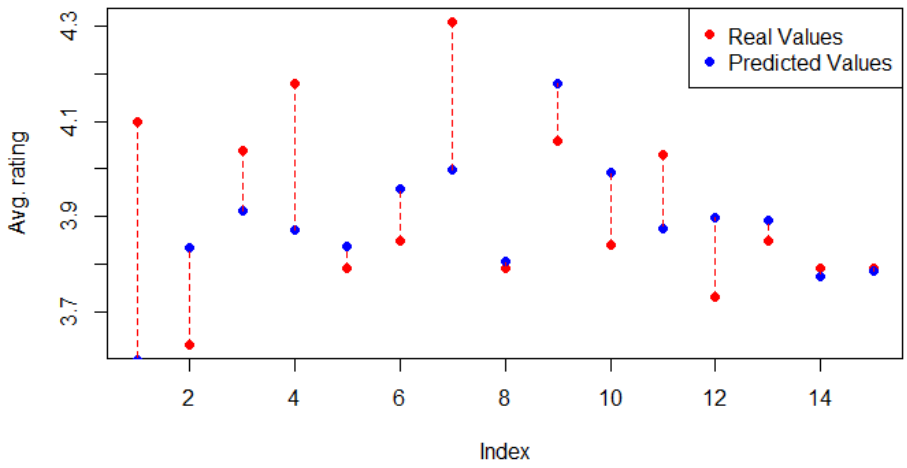

Figure 10: predicted vs. real values with highlighted differences of the best subset selection model with 12 variables.

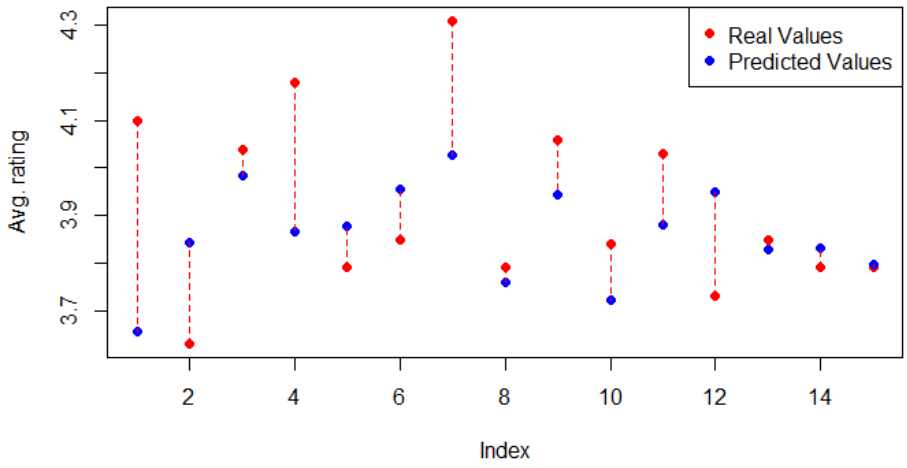

Figure 11: predicted vs. real values with highlighted differences of the Lasso model.

4 Conclusion

4.1 Interpretation of the models

There are multiple indications that analyzing text with linear regression is a fruitful endeavor; the top models put forth by the best subset selection used many different explanatory variables, that were almost all significant, showing that many of the verb and non-verb stop words might have important connections with how the text is read and perceived.

A more constant force in the models were the non-textual factors, especially the authors gender, if the book was originally written in English or if it had been translated, automated readability index, and the length of the book measured in characters. As can be seen in figure 12, which displays part of the summary of the best subset model with 12 variables, the strongest of these effects is of *ARI*. The more words there are per sentence, and the more characters there are per word, the lesser the model estimates the average score. It should be noted that the number of sentences were never significant, which puts emphasis on having generally shorter words as being impactful. Language has the second largest impact. The model prefers books that are not of English origin. This can perhaps be attributed to the model being based on the scores from an English community. It is not unlikely that this community is mostly exposed to English literature, such that the works from foreign cultures that reach the community have to be more excellent to bridge the gap. Gender is also important in the model; female authors have a higher level of appreciation. There are several possible explanations for this. It might be the case that a similar reason to the one for translated works being more popular is at play. Since the books in this project are old they stem from a time of great oppression, when it required much more of a female to become published than of a male, so that classics written by females might generally be of higher quality. The number of characters is the last general factor, it has a strong impact on the model; the longer the book the more it is appreciated on average.

Notice that none of the frequencies of words associated with NRC-lexicon's emotions are present in this model. Anticipation was in the best subset models with 9 and 10 variables, but was not significant at the 5%-level.

As for verbs, this model only includes "get" with its conjugations. The verbs "be" and "know" with conjugations were also present in the models with the greatest *BIC*. The larger models, with 14 and 15 variables include "want" and "work". In general the larger models are dominated by non-verb stop words, and include verbs only as they increase in size, while the

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.86163    0.01247 309.656  < 2e-16 ***
gender       0.06918    0.01664   4.157 5.09e-05 ***
lang        -0.07030    0.01386  -5.072 1.02e-06 ***
ari         -0.09668    0.01682  -5.749 4.04e-08 ***
charnum      0.06347    0.01372   4.625 7.37e-06 ***
get...      -0.03769    0.01582  -2.382 0.018298 *
to           0.05999    0.01685   3.560 0.000481 ***
and          0.06670    0.01529   4.362 2.22e-05 ***
in_          0.03518    0.01514   2.323 0.021333 *
I           -0.04188    0.01515  -2.764 0.006344 **
it           0.06065    0.01490   4.070 7.17e-05 ***
her         -0.05920    0.01725  -3.432 0.000751 ***
you          0.05196    0.01719   3.023 0.002888 **
```

Figure 12: part of summary of best subset model with 12 variables.

smaller models have a few words from both the verb- and non-verb groups. This may be an indication of non-verb stop words having less impact by themselves but a greater impact together.

Some notable stop words are the pronouns. The frequency of the word "her" has a negative impact on general likability according to the model. This might be an indication against too many female characters, or against having a female main character. On the other hand, the word "you" has a positive impact. In the context of classics this is probably a sign of there being more dialogue. Although "I", which is also a word that is frequently used in dialogue, has a negative coefficient.

Another interesting stop word is "and", which also has a positive connection with general appreciation. Frequent use of "and" makes text less dense by having words with meaning appear less often, and may improve rhythm.

The Lasso model used an even greater number of variables: a total of 22, excluding intercept. For the general properties this model includes the same as the best subset one with 12 variables, but also year and popularity. A heavier emphasis is put on the work being originally written in English by the Lasso model; it has the coefficient with the largest absolute value, again preferring works written in a foreign language. The second largest is the number of characters. *ARI* on the other hand is not as important as in the best subset model, and approximately shares coefficient size with the other general factors. The frequency of words connected with sadness is the

sole member of the emotions that are in the Lasso model, with a negative coefficient. When it comes to words the Lasso model differ from the best subset selection one, which only included the verb "get" with conjunctions. Instead the Lasso model included the other verbs that were in the final group of the best subset selection, in addition to "use" and "have", with conjunctions. For non-verb stop words they were almost the same, but the Lasso model did not include "to" and had a few, previously unused, with small coefficients.

Notably "and" has the coefficient with the third largest value. Indicating, as seen in the best subset model, that its use has a connection with appreciation of books.

The value of the coefficient of determination, $R^2_{adj.}$, for the best subset models tell of there being a lot of variance which is not accounted for in the model. Yet, there are many signs of the models working rather well in a predictive setting, and thus of the explanatory variables being pertinent. Perhaps a model that takes more important words and patterns into consideration could achieve higher values of $R^2_{adj.}$ and perform even better in a predictive context.

The frequencies of words associated with emotions have not been strong contenders in the models. Some of the larger models have included at most one of the emotion frequencies. Perhaps there is interest in a wide field of emotions and connotations in literature, such that no single one has a strong connection with appreciation. Another possibility, discussed in a paper by Schmidt et al.[9], is that some stop words, especially in older literature, are associated with emotions, and thus inflate the frequency of some emotion without actually having a strong connection with that emotion.

4.2 Prediction

The values of MSE that was calculated with leave one out cross validation gave promising signs for all of the models' predictive ability. This was corroborated by the models performing similarly well when applied to previously unseen data. The fact that these models possibly function for predicting the average score of classics demonstrate that there might be a connection between general appreciation and language-properties. This invites further questions: if the most improvement can be gained by fleshing out the models with more features, or if it is the choice of model that is the limiting factor.

As mentioned, the works, by Ashok et al.[1], and by Maharjan et al.[4], primarily focused on support vector machines for the classification of suc-

cessful/unsuccessful and liked/disliked literature. Both this project and the latter of the two references find signs of regression being an useful tool in computational text analysis. On the other, hand classification might be better in the sense that it more directly answers certain questions of publishers, authors and readers want answered. Will a book be successful? Will a reader like a certain book? Is a certain style appreciated? And so on.

5 Discussion

A challenge throughout this process has been the lack of data. Although there have been ways to navigate around this deficiency, more data would allow for experimentation with larger models and more refined validation.

Another aspect to consider is the order of data collection; when the test data was collected all of the models had already been developed and trained. The test data was collected in the same way as the training data, by filtering for popular public domain works on Goodreads. This also had the consequence of the books in the test data being less popular, in general, than the books in the training data.

It is also interesting to investigate how much one can extrapolate the results of this study to modern situations; literary classics are different from modern literature. On the other hand it is not unlikely that similar but different patterns exist in modern literature, and that they can be studied using the same method.

The author of this project does not have any previous familiarity with the field of linguistics. It can be assumed that this project would be able to reach further if it was supported by greater experience in the field of linguistics.

6 Appendix

6.1 A larger sample of the books in the data

title	author	gender	score	year	popularity	lang
War_and_Peace	Tolstoy	0	4.16	1869	1728285	0
Pride_and_Prejudice	Austen	1	4.29	1813	8475100	1
Jane_Eyre	Charlotte_Bronte	1	4.15	1847	3979264	1
The_Great_Gatsby	Fitzgerald	0	3.93	1925	5925712	1
The_Count_of_Monte_Cristo	Dumas	0	4.31	1844	1560773	0
Little_Women	Alcott	1	4.16	1868	2688047	1
The_Picture_of_Dorian_Gray	Wilde	0	4.13	1890	2616810	1
Wuthering_Heights	Emily_Bronte	1	3.89	1847	1917756	1
Crime_and_Punishment	Dostoyevsky	0	4.27	1866	2478385	0
Frankenstein_or_the_Modern_Prometheus	Shelley	1	3.88	1818	1728406	1
Dracula	Stoker	0	4.02	1897	1399934	1
Sense_and_Sensibility	Austen	1	4.09	1811	8475100	1
A_Christmas_Carol	Dickens	0	4.08	1843	3984047	1
Great_Expectations	Dickens	0	3.79	1860	3984047	1
The_Adventures_of_Huckleberry_Finn	Twain	0	3.83	1884	2793915	1
The_Secret_Garden	Burnett	1	4.16	1911	1595644	1
A_Tale_of_Two_Cities	Dickens	0	3.87	1859	3984047	1
Persuasion	Austen	1	4.15	1817	8475100	1
Emma	Austen	1	4.05	1815	8475100	1
Northanger_Abbey	Austen	1	3.85	1817	8475100	1
The_Gambler	Dostoyevsky	0	3.91	1866	2478385	0
The_Idiot	Dostoyevsky	0	4.21	1869	2478385	0
Notes_from_the_Underground	Dostoyevsky	0	4.17	1864	2478385	0
The_Brothers_Karamazov	Dostoyevsky	0	4.37	1880	2478385	0
Alice's_Adventures_in_Wonderland	Carroll	1	4.06	1871	1661942	1
Anna_Karenina	Tolstoy	0	4.09	1878	1728285	0
Scarlet_Letter	Hawthorne	0	3.43	1850	1030417	1
Don_Quixote	Cervantes	0	3.90	1605	338205	0
The_Three_Musketeers	Dumas	0	4.09	1844	1560773	0
The_Adventures_of_Tom_Sawyer	Twain	0	3.92	1876	2793915	1
Mansfield_Park	Austen	1	3.86	1814	8475100	1
Moby_Dick	Melville	0	3.55	1851	735981	1

Figure 13: A table of 32 books that are included, in addition to some general information about them: gender (of author) with 0 for male and 1 for female, average score, year of first publication, popularity measured in number of ratings on Goodreads.com, and language with 0 for translated and 1 for written in English.

6.2 Plots of residuals

Below are four plots of residuals discussed in section 3.3.

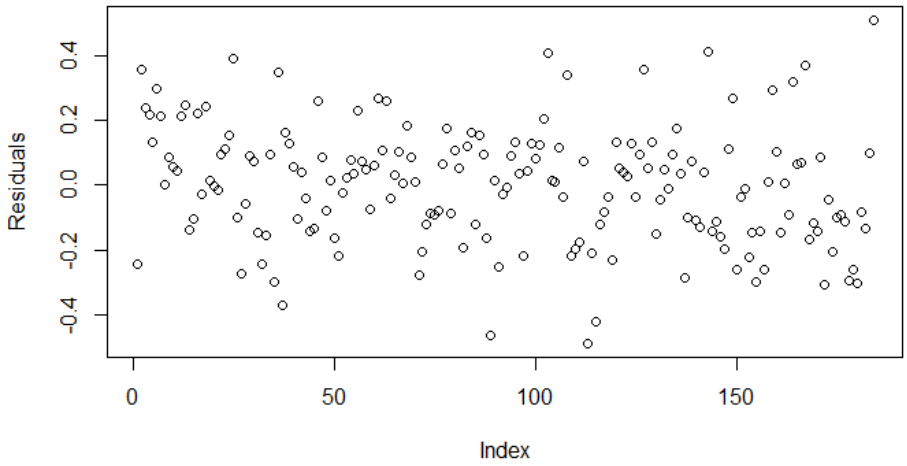

Figure 14: Residuals from the best subset model with 5 explanatory variables over index.

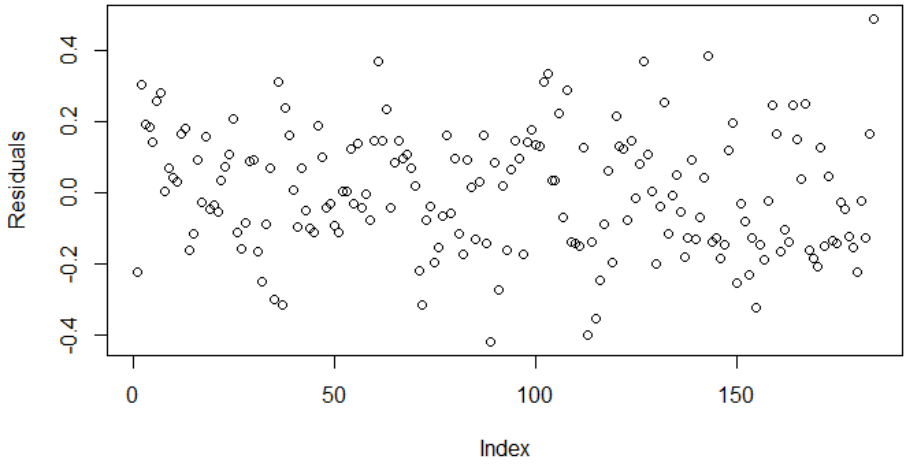

Figure 15: Residuals from the best subset model with 10 explanatory variables over index.

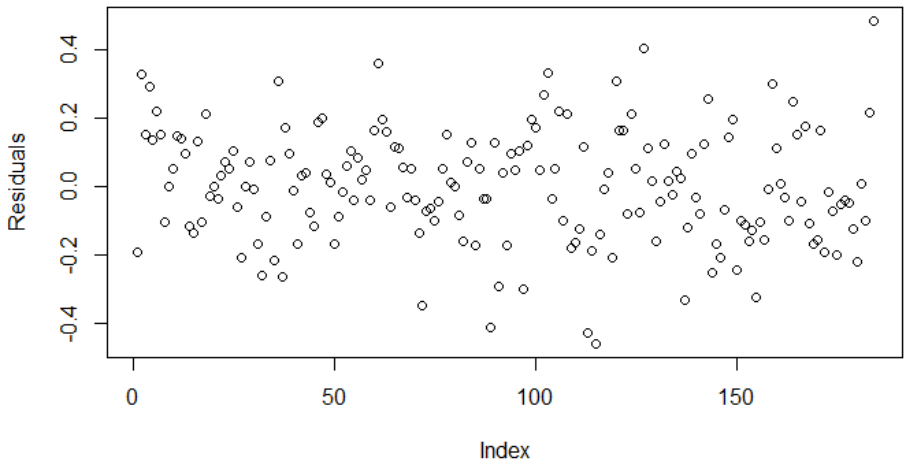

Figure 16: Residuals from the best subset model with 12 explanatory variables (which are also shown in the QQ-plot in figure 8) over index.

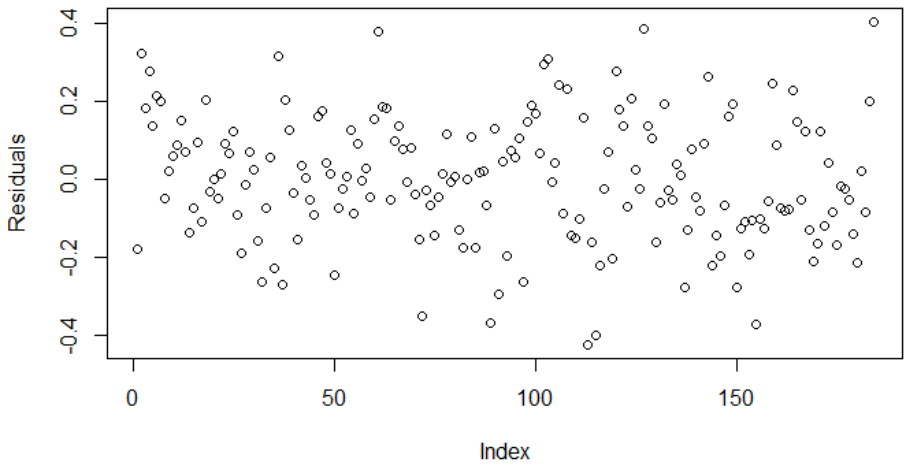

Figure 17: Residuals from the best subset model with 5 explanatory variables over index.

References

[1] Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1753–1764, 2013.

[2] Trevor Hastie. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[3] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[4] Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A González, and Thamar Solorio. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, 2017.

[5] Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34,

Los Angeles, CA, June 2010. Association for Computational Linguistics.

[6] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[7] Mahdi Mohseni, Christoph Redies, and Volker Gast. Approximate entropy in canonical and non-canonical fiction. *Entropy*, 24(2):278, 2022.

[8] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195, 2008.

[9] Thomas Schmidt and Manuel Burghardt. An evaluation of lexicon-based sentiment analysis techniques for the plays of gotthold ephraim lessing. Association for Computational Linguistics, 2018.

[10] Wikipedia contributors. Most common words in english — Wikipedia, the free encyclopedia, 2024. [Online; accessed 8-January-2025].