UTRECHT UNIVERSITY

MASTER THESIS

Inequality in Dutch Literary Prizes

Analysing word use in Dutch nominated novels using NLP techniques

Author:
Noa VISSER

Student number:
6979688

Supervisor:
Dr. Dong NGUYEN
Dr. Andreas van CRANENBURGH
Prof. dr. Kees van DEEMTER
Dr. Almila AKDAG

Master Artificial Intelligence

April 12, 2022

Universiteit Utrecht

UTRECHT UNIVERSITY

Abstract

Master Artificial Intelligence

Inequality in Dutch Literary Prizes
Analysing word use in Dutch nominated novels using NLP techniques

by Noa VISSER

Dutch authors have been criticising the homogeneity and the dominance of white men authors in Dutch literary prize nominations and the Dutch literary scene (Ramdas, 1997; Amatmoekrim, 2015; Rouw, 2015; Weijers, 2014). This homogeneity is clearly seen in the Dutch literary prizes. In general fiction the two most important prizes are the *Boekenbon Literatuurprijs* and the *Libris Literatuur Prijs*. For these two prizes, about 80 % of the nominated books from 1987 to 2020 were written by men. Such a discrepancy is quite remarkable, considering that an equal number of women and men writers publish novels in the Netherlands (Koolen, 2020).

Given that there is an overrepresentation of men in Dutch literary nominations, this inequality may be visible in the word use of the authors, as people tend to use similar language as their peers (Eckert, 2012). Therefore, this research investigates whether it is possible to identify author gender inequality in Dutch literary prizes using quantifiable literary qualities. I hypothesise that nominated and not nominated novels can be identified based on word use. I also hypothesise that, due to the dominance of men authors in literary nominations, nominated novels written by men will be easier to classify compared to nominated novels by women; and vice versa for not nominated novels. I have used logistic regression classification, LDA topic modelling and cosine delta, to identify author gender in equality in Dutch literary prizes.

I collected a corpus of 300 original Dutch novels from 1989–2012, consisting of three subcorpora: **NomNov**: nominated novels, **NomAut**: not nominated novels by nominated authors, and **NotNom**: not nominated novels by not nominated authors.

The results show that it is possible to investigate author gender inequality in Dutch literary prizes with quantifiable literary qualities, but it also indicates that the inequality in Dutch literary prizes is rooted in a homogeneous writing style that is related to the writing style of men. The results clearly show that nominated and not nominated novels are distinguishable, both for men and women writers, thus indicating that a particular word use exists that identifies literary quality. However, this word use seems to be further removed from women writers, even from their word use in nominated novels, as the classification of novels written by women consistently have the lowest performance. The analysis of the topics in nominated and not nominated novels indicate that the relation between nominated and not nominated novels and author gender is rather complex, and highly depends on the topic which is investigated. The difference in writing style of nominated and not nominated novels cannot be clearly defined, but the results do indicate that the writing style of Harry Mulisch has is related to writing styles that are perceived to be of literary quality.

Acknowledgements

I would like to thank Andreas van Cranenburgh, Dong Nguyen and Kees van Deemter for their guidance and feedback during this project. Without their feedback and aid, I would not have been able to have completed this thesis as successfully as I have done now.

I am very grateful for the unconditional love, support and encouragement of my family and friends during this journey. I would particularly like to thank Alba, Esther, Lorenzo, Fay, Ghislaine, Laurens and Mara for proofreading my thesis and Arja for helping me statistically analyse author gender inequality in the selection process for nomination. I would also like to thank my wonderful housemates, Kofi and Cloudy, for giving me a home where I could come back to and relax, after long days of working on this thesis.

Lastly, I would like to thank my nephew who will be welcomed in this world very soon, and with that setting a strict deadline for me to graduate. Thank you for this extra motivation in the final part of my thesis.

Contents

Chapter 1

Introduction

In the Netherlands, men win substantially more literary prizes than women. In general fiction the two most important prizes are the *Boekenbon Literatuurprijs* and the *Libris Literatuur Prijs*. For suspense novels, the most important prize is the *Gouden Strop*. For these three prizes, 80% of the nominated books from 1987 to 2020 are written by men. Considering that an equal number of women and men writers publish novels in the Netherlands (Koolen, 2020), such a discrepancy is quite remarkable.[1]

Not only is the percentage of nominated novels by men much larger than the percentage of novels by women, but the percentage of men writers with multiple nominated novels is also higher than for women writers. This imbalance is seen in the nominations for all three prizes mentioned above, but it is particularly strong for the *Gouden Strop*. From all the nominations from 1978 until now, there are three authors with more than eight nominated novels, namely Charles den Tex, René Appel and Tomas Ross. Overall, 155 novels have been nominated for the *Gouden Strop*, thus the works of these three authors form a large portion of all nominated works.

In general fiction, Arnon Grunberg is the author who has been most often nominated for the *Libris Literatuur Prijs*, with six nominations. The second and third most often nominated authors are A.F.Th. Van der Heijden (4) and Frank Westerman (5), both men. Arnon Grunberg is also the most often nominated author for the *Boekenbon Literatuurprijs*, with seven nominations. The second most often nominated author is Esther Gerritsen, a woman, who has been nominated four times. Thus, also for the *Libris Literatuur Prijs* and the *Boekenbon Literatuurprijs*, the majority of the multiple nominated authors are men. Therefore, it is interesting to research author gender distribution in nominations for Dutch literary prizes. This thesis will use word features to analyse nominated and not nominated novels, in order to identify literary qualities, such as word use and topics. These literary qualities will be used to investigate whether quantifiable literary qualities can be used to identify author gender inequality in Dutch literary prizes.

Although this thesis will focus on the author gender inequality, it is important to note that all the authors mentioned concerning the *Libris Literatuur Prijs*, *Boekenbon Literatuurprijs* and *Gouden Strop* are white. This thesis will only focus on author gender inequality, but other forms of inequality, such as ethnic and cultural background also lead to a form of homogeneity in the Dutch literary scene (see Section 1.2). Due to limitations of the corpus available, other forms of inequality of research, beside author gender, could not be investigated. Additionally, the analysis of author gender will only focus on men and women, also due to limitations of the dataset.

[1] I have chosen to use *women* and *men* writers instead of *female* and *male* writers, as the words female and male describe sex and this thesis is focused on gender.

Before the homogeneity in the Dutch literary scene is discussed, the author gender distribution of the *Libris Literatuur Prijs* will be analysed (see Section 1.1) to discover if statistically a bias against women can be identified in the nomination process. Then, a background on homogeneity in the Dutch literary scene will be given. Lastly, the research questions and hypotheses will be introduced in Section 1.3.

1.1 Libris Literatuur Prijs

As stated above, the goal of this section is to provide a short statistical analysis, to determine whether a bias against women can be found. The *Libris Literatuur Prijs* is chosen, as the organisation made an analysis on the author gender distribution of the grosslist, longlist, shortlist and winners and the gender distribution of the jury members (Stichting Literatuur Prijs, 2021) due to the critique on the limited number of women winning the *Libris Literatuur Prijs*. The grosslist is the list of all novels submitted to be considered for nomination, the longlist is the selection of 18 books, which can be nominated for the shortlist. The shortlist consists of six novels, of which one is the winner.

The overall author gender distribution from 1987 to 2013 shows that the gender inequality starts with the novels that are submitted by publishers, with 68.4% works of men and 31.6% works by women. Note that it only includes men and women authors, as the first openly non-binary authors got nominated for a prize after 2013. A two sided hypothesis test shows that the inequality of novels written by women being selected for the longlist is significant[2]. This inequality increases when the 18 books for the longlist are selected, as now 72.2% of the total number of books that have been selected for the longlist are written by men, opposed to merely 27.8% by women. Moving on to selection for the shortlist, about a quarter of the books has been written by women. The biggest increase in author gender inequality is seen in the winners of the *Libris Literatuur Prijs*. Only 10.5% of the winners are women.

Thus, in every step of the process fewer women are selected. Despite the percentages of women decreasing in every step closer to the selection of a winner, the biggest indication that there is a bias in the selection procedure is the number of novels written by women being selected for the longlist. This is significantly smaller than the percentage to be expected if novels are randomly selected for the longlist. Even though it cannot be assumed that all novels on the grosslist are of the same quality, and therefore should not have an equal chance to be selected, it does clearly show gender inequality on the chance of being selected for the longlist.

1.2 Homogeneity in the Dutch Literary Scene

The homogeneity in literary prize nominations and the literary scene has been critiqued by writers for a long time (Ramdas, 1997). Amatmoekrim (2015) and Rouw (2015) argue that this homogeneity, in particular the lack of acknowledged non-white authors, are due to the homogeneity of the literary environment, with several established publishers, similar authors being nominated and a homogeneous notion of literariness. Weijers (2014) also argues that works written by men determine the norm of Dutch literature, and the works of women are seen as a derivative

[2]I would like to thank Arja Rydin for calculating the chance that a novel written by a woman is selected for the longlist, shortlist and is chosen as winner after being submitted by the publisher. It is assumed that all novels on the grosslist have an equal chance to be selected, which is a simplified reality.

of the works by men. She also notes the difference in promotion of novels by men and novels by women. These experiences and arguments of Anil Ramdas, Karin Amoetmoekrim, Ebisse Rouw and Niña Weijers show that the homogeneity in literary prizes is recognised and critiqued within the Dutch literary scene and that this inequality is complex and multifaceted.

Homogeneity in the literary scene does not only exists in the Dutch literary scene. Several projects have been set up worldwide to quantify the gender breakdown of major literary works and book reviews, such as Stella[3], focused on Australian writers, the VIDA count[4], focused on the United States of America, and Frauen Zahlen[5] and Literaturkritik in Zahlen[6], both focused on books written in German. These projects focus on publications and book reviews, as book reviews, particularly in major news papers, have a large influence on the popularity of a novel. Unfortunately, the outcomes from all projects show that novels written by women do not receive the same attention in major news papers and books reviews, as novels written by men.

Thus, (white) men seem to dominate the literary scene of several Western countries, as these are the novels that are mostly read, mostly reviewed and mostly nominated. Changing the lack of diversity in institutions, such as literary awards, is a complex matter (Ahmed, 2012). Over the last decade, more women have made up the majority of the jury members in the Dutch literary award scene (Boudewijn, 2020). However, as Dijkgraaf and Appel (2013) also show, the inclusion of more women on the jury, does not necessarily lead to less homogeneous nominations.

Factors inequality The homogeneity in Dutch literary awards seems to not only be caused by the lack of diversity of the juries, nor does it seem to be a phenomenon specific to the Dutch literary seen. The effect of this homogeneous norm, is that everyone that is positioned outside this norm is seen as the 'other', which often compromises the identity of the person who is positioned as the 'other' (Beauvoir, 1953; Fanon, 2008).

In the Dutch literary scene, the positioning of non-white men is seen and enforced by several factors. Literary publishers and other professionals, value formal aspects of literary works, in which prestigious novels are 'literary' and 'universal' (Koren and Delhaye, 2019). They often place white writers in the framework of 'literary' and 'universal' works. Contrarily, non-white writers and publishers are placed in frameworks based on their identity. For example, book reviews in Dutch news articles stress the ethnic and cultural background of non-white writers more, in comparison to German newspapers and newspapers from the USA (Berkers, 2009). This emphasis creates the idea that novels written by non-white authors are different from the Dutch norm of literary quality, positioning these works outside of the norm (Staszak, 2008). Due to this framework, the work of non-white writers is perceived as 'political' and 'subjective' and therefore less prestigious (Koren and Delhaye, 2019). Another factor that is likely to influence the inequality in the nominations of novels, is the influence prestige of the genre, the author and the novels. Literary ratings by amateur readers are influenced by genre, prestige of the author and prestige of the novel itself (Koolen et al., 2020). Online literary communities show a clear gender bias, where popular literature, romantic novels, chick lit or thrillers are more

[3] https://stella.org.au/initiatives/research/
[4] http://www.vidaweb.org/the-count/
[5] http://www.frauenzählen.de/
[6] https://www.uibk.ac.at/iza/literaturkritik-in-zahlen/

often associated with female authors and are segregated from mostly male authors of literary quality (Deijl, Smeets, and Bosch, 2019).

Lastly, the homogeneous idea of literary quality is maintained by the Dutch school curriculum. In the Netherlands, secondary school students are obligated to read a certain number of literary novels (differing per education level), the so-called '*leeslijst*'. The manner in which subjects are structured and topics are discussed, influence what students perceive as the standard and norm of knowledge (Wekker, Slootman, Icaza Garza, Jansen, and Vázquez, 2016). Dera (2021) shows that the majority of the works read by students are white men. Women and non-Western authors are the most underrepresented, but Flemish authors are also considerably less represented than Dutch authors. As the goal of the 'leeslijst' is to teach students what literature is, these structural under representations uphold the idea that the norm of literary quality is associated with white, Western men writers (Dera, 2020).

As can be seen, the association between literary quality and white, Western men writers is uphold by multiple factors, such as the identities emphasised in books reviews and the manner in which school curricula teach students what literary quality is. As literary awards are supposed to award the 'best' novel, it is interesting to further investigate how the texts itself relate to this homogeneous norm of literary quality, by researching the word use and topics within nominated and not nominated novels.

1.3 Research Question

Computational techniques are suitable to research literary quality using word use and topics (Koolen, 2018; Cranenburgh and Bod, 2017). The benefit of using computational techniques to analyse literature, is that large corpora can be easily analysed and word use that relates to a specific author or genre can be statistically identified (Herrmann, Jacobs, and Piper, 2021). A well-established method to analyse text in relation to writing style and topics, is to analyse the distinctive textual features identified by logistic regression using word frequencies (Herring and Paolillo, 2006; Bamman, Eisenstein, and Schnoebelen, 2014; Fast, Vachovsky, and Bernstein, 2016; Koolen and Cranenburgh, 2017). In this method, the distinctive textual features are analysed to identify the relation between writing styles, topics and the predicted variable, which is very applicable to analyse the author gender inequality in literary nominations and its relation to differences in writing style and topics.

The homogeneity of the nominations of literary awards could also be related to the topics of a novel. As amateur readers relate genre to literary quality, in which genres of 'less' literary quality are also more strongly related to women authors (Deijl, Smeets, and Bosch, 2019; Koolen et al., 2020), it could be expected that this relation can be found in the text of nominated and not nominated novels as well. Therefore, I will explore the difference in word use in nominated and not nominated Dutch novels.

Writing style and topics are interesting to investigate, as people tend to use similar language as their peers. For example, as peer groups are often homogeneous in gender and age, people of the same gender and age have a language use that is more closely related to each other (Eckert, 2012). Since the majority of the (multiple nominated) writers are white men, and the Dutch literary scene is homogeneous as well (Boudewijn, 2020; Koren and Delhaye, 2019; Dera, 2020; Ramdas, 1997; Amatmoekrim, 2015), the peer group of writers of high literary quality novels could have a language use that is specific to this scene. Due to the make up of the Dutch literary

scene, this specific word use could relate more to the word use of white men, than to other people.

A logistic regression model based on textual features will be trained to classify which novels have been nominated for a prize, and which have not. The output is a prediction of which novels have been nominated in the past. Unfortunately, insufficient data was available to properly analyse other intersections of identities using Natural Language Processing (NLP) techniques. I have chosen to focus on author gender only, and not include other intersections of identity, such as cultural or ethnic background, as the data available consists of mainly Western white authors. In order to investigate whether the word use in literary novels can be related to the homogeneity in the Dutch literary scene, I will analyse my results concerning author gender. I hypothesise that nominated and not nominated novels can be identified based on word use. I also hypothesise that, due to the dominance of men authors in literary nominations, nominated novels written by men will be easier to classify compared to nominated novels by women; and vice versa for not nominated novels. Thus, I expect that author gender inequality in Dutch literary prizes can be identified using quantifiable literary qualities.

In order to research this hypothesis, the following three research questions will be answered:

1. RQ1: Can nominated and not nominated novels be identified based on textual features only?

2. RQ2: Is there a relation between classifications on nominated and not nominated novels and author gender, where both classifications are based on textual features?

3. RQ3: Are the differences in topics/writing styles between books that are nominated for literary prizes and those that are not, related to author gender?

The goal of the first question is to investigate whether it is possible to identify nominated and not nominated novels based on textual features using logistic regression. It is assumed that there is a difference in word use between books that have been nominated for literary prizes, and books that have not. However, it has not been researched if a logistic regression model can be trained to classify which books have been nominated, and which books have not, using bag-of-words word features only. I hypothesise that this is possible, and this hypothesis will be confirmed when the classification task surpasses chance. This means that the predicted classes are more accurate than when the model makes random guesses, and thus the model must have made generalisations over the textual features that distinguish nominated and not nominated novels.

The second question aims to explore if a relation between nominated novels and author gender can be identified using textual features. To answer this question, the results of the model trained to classify which novels have been nominated or not, will be analysed on author gender. The goal is to identify author gender specific patters in the results of the classification task. Also, a logistic regression model will be trained to classify author gender, of which the results will be analysed focusing on whether a novel has been nominated or not. The goal is to relate the results of the author gender classification to nominated and not nominated novels, and to analyse how these patterns relate to the results of the model trained to classify nominated novels. Lastly, the results of a logistic regression model trained to classify

nominated novels, and a model trained on author gender will be compared. I hypothesise that such a relation between nominated novels and author gender can be identified, namely that nominated novels relate to novels written by men, and not nominated novels relate to novels written by women.

The last question aims to identify the topics in nominated and not nominated novels. The goal is to explore which topics occur more in nominated novels, and are therefore probably associated with higher literary quality. The topics will be identified using LDA clustering. To find out how the topics that strongly relate to nominated and not nominated novels relate to author gender, the results will be analysed concerning author gender as well. I expect that topics can be identified that relate more to nominated novels than to not nominated novels and vice versa. I also expect that certain topics will relate more to novels written by women than to novels written by men and vice versa.

For the writing styles, cosine delta will be used to identify specific word use that relates to nominated novels and to not nominated novels. These results will be used to research the relationship with author gender as well. I hypothesise that a specific writing style can be found that relates to nominated novels, and that the novels written by men most strongly relate to this writing style. For the not nominated novels, I hypothesise that a writing style can be identified that relates to not nominated novels, which most strongly relates to novels written by women. As the techniques used to answer this question are unsupervised algorithms, these results will be used to give a more interpretable insight on the relation between nominated and not nominated novels and author gender.

Gender As this thesis focuses on the relationship between author gender and nominations for literary prizes, it is important to clearly define how the variable gender will be used throughout this research. Gender is an ethically complex feature to use in AI research, as it is a social construct (Butler, 1998). It is often implemented as a binary variable, whereas more than two gender identities exist. As a binary view on gender is a Western and colonial categorisation (Oyewumi, 2002), it is important to be very critical on the use of gender in AI, as it can unintentionally reproduce and reinforce a very limiting perspective. Keyes, May, and Carrell (2021) argue that it is important to treat gender as 'multiplicitous': a concept which has many meanings and relations to individuals and communities.

Even though researching original Dutch novels, I will work from this point of view throughout my research. This means that I will propose a method which enables me to analyse the influence of author gender on literary nominations, without reinforcing an overly generalised, binary interpretation of gender. I will also analyse my results in a context of gender which is multiplicitous. This is not only important due to the complexity of the concept gender, but also to analyse the results of this research as a 'bias transforming' metric. A bias transforming metric, is a metric that does not blindly accept the social bias perceived in data (Wachter, Mittelstadt, and Russell, 2021). As gender bias in society is based on a Western binary view, I find it important to evaluate and analyse my research in a manner that does not blindly reproduce the gender stereotypes and biases in society. My goal for this research, is to analyse the inequality in literary nominations in such a manner, that it gives insight in this gender inequality and identifies the related bias and stereotypes. Then, this research can be used to challenge gender stereotypes and biases.

Chapter 2

Related Work

In this chapter, an overview of relevant related work will be given. First, I will discuss various approaches and techniques in computational stylistics & literature, as well as the usage of textual features to investigate literary quality in Dutch literature. Then, I will discuss specifically how author gender can be researched using NLP techniques. This is important, as it gives an insight in how computational stylistics and NLP techniques can be used to research the relation between author gender and written text in a nuanced manner.

2.1 Computational stylistics & literature

Computational stylistics is a field that focuses on modelling 'literary discourse' using computational and statistical methods (Herrmann, Jacobs, and Piper, 2021). The goal of this section is to provide an overview of computational stylistics, and of previous research on Dutch literature and literary quality using computational stylistics.

2.1.1 Computational stylistics

Herrmann, Jacobs, and Piper (2021) give an overview of the field of computational stylistics, grouping the field into three categories: formalist, social and cognitive approaches. Formalist approaches focus on understanding the distinctive features and structures of literary works, including the manner of writing that constitutes literariness, the nature of genres, literary quality or authorial style. Social approaches investigate social practices across communities, such as 'canonicity' and 'prestige'. Cognitive approaches research the 'cognitive' side of aesthetics and stylistics, such as the psychology of literature and reader response.

I will discuss the formalist and social approaches in this section, as formalist approaches focus on distinctive textual differences that constitute literariness and social approaches research reading communities and larger social fields of interaction, which is relevant when investigating author gender inequality in literary prizes. Thus, these two categories are deemed as the most relevant for this thesis.

Formalist approaches This thesis aims to answer the question 'Can quantifiable literary qualities be used to investigate author gender inequality in Dutch literary prizes?' using computational techniques applied on bag-of-words. Therefore, computational stylistics research on style, authorial signal, literariness and fictionality will be discussed in this section.

Writing style can be seen as a complex system of combinations of formal features (Herrmann, Jacobs, and Piper, 2021), in which formal features are linguistic features

on character, lexicon, syntax and semantic level. The most reliable features for measuring stylistic similarity and distinction are function words, such as *the, of* and *in* (Burrows, 2002). Computational stylistics is based on the assumption that individuals have idiosyncratic and largely unconscious habits of language use, leading to stylistic similarities between texts written by the same person (Evert, Proisl, Jannidis, Reger, Pielström, Schöch, and Vitt, 2017). Therefore, computational techniques can determine authorship, due to the relative frequency of function words, parts of speech, degrees of vocabulary richness or syntactic complexity.

Computational techniques has been successfully used to determine probabilisticly authorship, across journalistic and literary texts, and across different languages such as English, Ukrainian, Portugese, Spanish, French, German and Italian (Marsden, Budden, Craig, and Moscato, 2013; Lupei, Mitsa, Repariuk, and Sharkan, 2020; Varela, Justino, Britto, and Bortolozzi, 2016; Tuzzi and Cortelazzo, 2018). Different authors use measurably distinct styles by over-utilising or avoid particular common words and phrasing, despite using the same structural and grammatical bounds of a common language (Marsden et al., 2013). Writers favour (or filter) certain words in a manner which goes beyond the use (and avoidance) of common phrases due to word use in social groups. This word preference creates an individualistic style which can be probabilisticly identified. In authorship research, research focuses on writer-dependent approaches, in which models are built for a specific author, or writer independent approaches, in which models are built to determine if a given text is authentic or false (Varela, Albonico, Justino, and Bortolozzi, 2018). Models can also be used to verify or identify authorship. In authorship verification, texts are compared to samples of the same author, to see if the model can correctly verify the author. In authorship identification, a text is compared to multiple possible authors, from which is probabilistically determined which author the text most likely belongs to. For example, Tuzzi and Cortelazzo (2018) compared the works of Elena Farante to several Italian writers, using both the entire vocabulary as well as only grammar words. The results of several models were used to determine the most likely ghost writer of Elena Farantes work.

The amount of research that shows that it is possible to identify and verify authorship using computational techniques, and that computational techniques lend themselves for identifying distinctive writing styles. However, identifying literary writing styles across authors is difficult, due to internal variation of literary genres (Underwood and Sellers, 2012). The goal of studying 'literariness' with computational stylistics is to test various stylistic features that distinguish literary/fictional from non-literary/non-fictional discourse (Herrmann, Jacobs, and Piper, 2021). This type of research is still an emerging field, and therefore more research with more data and across more languages is needed before new theories and hypothesis can be derived on the quantitative stylistics features that contribute to literariness.

Social approaches One of the key areas of the socially-oriented frameworks in computational stylistics is examining the relationship between representation and inequality, by examining inequalities and biases of representation in literary and other cultural documents (Herrmann, Jacobs, and Piper, 2021). Representations are explored on two levels, namely on the level of agents, such as authors, characters, publishers and editors, and level of form, such as style and semantics. An example of representation on agent level is Underwood, Bamman, and Lee (2018), which shows a massive decline of women authors in English fiction in twentieth century. The form level of representation is also explored in this study, as the historical investigation of English fiction in the twentieth century also showed that the gender

division between characters becomes less sharply marked over this period of time, suggesting a growing equality in gender representation in characters.

Lejun, Xiangyu, and Huakang (2021) is one of the few studies on the relation between the representation of characters and literary prizes on the form level. They found that a high concentration of characters and emotion fluctuation are common characteristics in the works of the Nobel Prize in Literature in 2012 and 2013. This suggests that the manner in which characters are portrayed in novels can have an influence in the perception of literary quality. Thus, it relevant for this thesis to have an overview on research on computational stylistics from a social approach, as this approach can be used to explore inequalities and biases of representations within novels.

The research focusing on social approaches show that there is a complex influence of the perception of gender on the valuation of novels written by authors of different genders. Authors seem to write predominantly about characters close to their daily life experience (Van Der Deijl et al., 2016). As the authors are predominantly men, this results in an overrepresentation of men as main characters in Dutch literature. Van Der Deijl et al. (2016) also show that the narrating main characters are predominantly highly educated men from Western descent, similar to the majority of the authors in the corpus. Authors also appear to portray the characters of different genders in very different professional settings. In the corpus students of primary, secondary or higher education are the most common professions for both men and women characters. However, for men the third and fourth most common professions are entrepreneur and teachers, whereas for women those are sex worker and housewife. These difference in types of profession, influence the plot and topic of a novel. Thus, the position of men and women in society, as well as the homogeneity of author gender in Dutch literature, seem to influence the way men and women characters are described. Also, the homogeneity in the characters of literary novels, and the manner in which the characters are portrayed, could have an effect on whether a novel is perceived as literary or not. Smeets, Sanders, and Bosch (2019) nuances these results, as their social network analysis on Dutch characters in contemporary novels show that women and immigrant characters statistically take up a more central position these novels than men and non-immigrant characters. Due to the limitations of their corpus, these results could be skewed. Therefore, they argue for the urge to strongly connect qualitative and quantitative strands in further research.

In contemporary English fiction, gender bias and heteronomativity, in particular heteronormative pairings and interactions, seem to occur highly across genres (Kraicer and Piper, 2019). Women writers reduce these biases, but nonetheless include more man characters in their narratives than characters of other genders, and are more likely to create heteronormative social networks within their novels. Thus, women characters are less visible than men characters in English contemporary novels, and author gender can reduce, but not erase, these inequalities (Kraicer and Piper, 2019). Additionally, in an analysis of 18th- and early 19th-century English fiction, Rybicki (2016) found that women tend to become part of the canon if/when they write more like men. Novels by women become canon when their most frequent words are closer to the most frequent words men use. The gendered word use, however, seems to be consistent with the social values and gender stereotypes of the period in which the books were published. In order to become part of the canon, women writers seem to have to use words more similar to men (Rybicki, 2016), despite portraying men and women according to gender stereotypes.

2.1.2 Textual features & literary quality in Dutch literature

In this thesis, Dutch novels will be used, and therefore a more in depth overview of computational stylistics research on literary quality in Dutch literature will be given. Koolen et al. (2020), Cranenburgh and Bod (2017) and Cranenburgh and Koolen (2020) show that features that distinguish literariness in texts can be identified. Thus, this supports that quantifiable literary features can be used to investigate the author gender inequality in Dutch literary prizes.

National Reader Survey In order to investigate author gender inequality in Dutch literary prizes, it is important to understand how literary judgements on Dutch literature can be predicted using computational techniques. To do so, Cranenburgh and Bod (2017) used the results of the National Reader Survey, which measures the perception of literary quality by Dutch readers. In this survey, readers could rate books on literary quality, both on books that the respondents have read and books that they have not read (Koolen et al., 2020). For the books that the readers did not read, respondents could fill in the rating of literary quality they expected the novel to have. The participants could also motivate their rating. Two main motivations were given by the respondents to rate literary quality, namely genre and the text itself, which includes style, structure, plot and layers.

The results show that respondents base their expectations of literary quality on literary quality from the 'genre' of the book, such as suspense and chick lit. Detectives, thrillers and chick lit are not perceived to be of high literary quality, whereas literary novels are mainly perceived to be of high literary quality. This influences the rating of women authors, as these books are more often marketed within a particular genre. This relation between author gender and genre is in line with the findings of Deijl, Smeets, and Bosch (2019), which show a clear relation between certain genres and author gender in online literary communities. Furthermore, respondents often compare low-rated novels to "women's novels", whereas "men's novels" are barely mentioned. From the difference in literary ratings between novels of different genres, as well as the motivations given by the respondents Koolen et al. (2020) conclude that a shared consensus of literary quality exist amongst Dutch readers, which is grounded in textual features, such as writing style which includes sentence length and word usage.

Predicting literary quality computationally The results of The National Reader Survey have also been analysed computationally. Cranenburgh and Bod (2017) use the results of The National Reader Survey, to predict the average literary rating in the survey, based on textual features. A combination of textual features were used: sentence length, direct speech, vocabulary richness, cliches, topics, character and n-grams, and tree fragments. Meta data features such as genre, author gender and whether the work was translated or not, were used as well. Interestingly, author gender and translation only increase the score when they are both present. This increase could be due to a bias in the dataset, as the dataset contained more translated novels than originally Dutch novels by women and more original works, than translated novels by men. Thus, it seems that it is important to use novels in original language only when analysing author gender. An ensemble of the features reaches an accuracy of 76% when predicting literary scores that were abstracted from The National Reader Survey. Therefore, Cranenburgh and Bod (2017) shows that it is possible to computationally predict the literary quality of Dutch novels.

To further analyse the influence of prestige on literary quality, Cranenburgh and Koolen (2020) conducted an exploratory analysis in which readers rated anonymised fragments of 250 words of 8 novels from The National Reader Survey, which results confirm that the prestige of an author and/or novel influences literary judgements.

Uneven author gender distribution A drawback of the National Reader Survey, and the research based upon its results, is that, in the corpus, author gender is not evenly distributed across genres. Despite that this corpus of 401 Dutch novels has an almost equal percentage of men and women writers, this is not seen in the subset of general fiction. In this genre, there are more originally Dutch works by men, and more translated works by women. As genre and author gender both influence literary ratings, Koolen and Cranenburgh (2017) analysed a second corpus of general fiction, consisting of an equal amount of works from women and men. They trained a support vector classifier on both datasets to classify on author gender. As the features with the highest weights could not be interpret or related to author genders, a topic model was applied. This resulted in topic clustering which could clearly be related to author gender. For example, the topic military is strongly related to works by men, whereas the topic settling down is strongly related to novels by women. Thus, Koolen and Cranenburgh (2017) shows that it is possible to use topic modelling to investigate and interpret how topics in novels relate to author gender.

Thus, previous research suggest that it is possible to distinguish nominated and not nominated novels based on textual features, as it seems possible to predict literary quality based on textual features (Cranenburgh and Bod, 2017; Cranenburgh and Koolen, 2020). It is important to take genre and author gender in account in this type of research, and to use interpretable models to draw nuanced conclusions and limit the reproduction of stereotypes (Koolen and Cranenburgh, 2017).

In conclusion, computational stylistics is suitable to analyse literary quality and the influence of author gender on literary nominations. Even though no computational stylistic research has been done on Dutch literary nominations, previous research on literary quality and authorship attribution suggests that it could be possible to computationally distinguish the writing style of nominated novels from the one of not nominated novels.

2.2 Gender as a social variable

In this section, I will discuss NLP research that focuses on the relationship between gender and language. The goal of this section is to show the nuance required when using NLP techniques to analyse author gender. It is important to provide an overview of such research, as computational stylistic techniques in itself does not guarantee a nuanced analyses of author gender. Therefore, this section provides an overview of NLP research that uses gender in a not strictly binary manner. Furthermore, the research discussed in this section shows the importance and different manners in which author gender can be researched in a nuanced way using computational analysis.

I will begin with a short explanation on gender as a social variable. Then I will discuss Argamon, Koppel, Fine, and Shimoni (2003) in depth, as this research is the basis of almost all other research discussed in this section, despite making very generalising conclusions. To conclude, I will discuss a few papers treating gender as a social variable.

In NLP, gender is often treated as a biological characteristic, which is can be a very limiting view on gender, as it ignores the agency of a speaker, while going against gender theory and social science where it is considered that gender is something that someone *does* instead of *is* (Nguyen, Doğruöz, Rosé, and Jong, 2016). Additionally, individual language use varies due to the social group someone is situated in or communicates with (Eckert, 2012). As peer groups are often homogeneous in gender and age, people of the same gender and age have a language use that is more closely related to each other. Thus, the relation between gender and language is social.

Recent NLP research has also confirmed that gender should be approached as a social variable, rather than a static biological one (Bamman, Eisenstein, and Schnoebelen, 2014; Nguyen, Trieschnigg, Dŏgrüoz, Gravel, Theune, Meder, and De Jong, 2014). As language is inherently social, individual speakers often diverge from the gender stereotypes that are found in many studies (Nguyen et al., 2016). Even though certain language features are used more by a certain gender on average, NLP research should refrain from drawing generalising conclusions. Furthermore, gender varies in different cultures and languages, and linguistic variation can also be identified within speakers of the same gender. Thus, in NLP gender should be treated and analysed as a social variable, rather than a biological characteristic.

2.2.1 Do women write more involved and men write more informative?

Argamon et al. (2003) is an influential research that investigated the difference between writing of men and women, in English fiction and nonfiction. This research shows that differences in writing style is seen between authors or different genders, and that these differences are strongly related to genre. They find that women writing style is related to fiction, whereas men writing style is more related to nonfiction. Based on the distinctive features found, they conclude that women write more 'involved' and men write more 'informative'. Despite the thorough analysis of Argamon et al. (2003), the conclusions draw by them are rather generalising.

To identify the differences in writing styles between men and women, an Exponantiated Gradient (EG) algorithm selected the most useful features for categorising a document. Determiners and quantifiers were identified as indicators of man authors and pronouns as indicators of woman authors. Argamon et al. (2003) relate these features with a 'involved' and 'informative' writing. They establish that women writers use more pronouns that are related to the relationship between the writer and the reader, such as first person singular and second person pronouns. Men tend to use more generic pronouns. They argue that these results indicate that woman 'personalise' text more than men. Additionally, they claim that men indicate or specify the things that they write about more frequently, as they use determiners more often. They also find that the use of dialogue in texts is characteristic for woman authors, especially in fiction. In nonfiction, women tend to use quotation marks more often, suggesting that women cite other people's words more than men do. Another important result is that they find a strong correlation between texts written by men and nonfiction and texts written by women and fiction. To conclude, Argamon et al. (2003) argue that the gendered difference between 'involved' and 'informative' writing is due to the differences in socialisation of people of different genders. They also argue that the significant relation between gender and fiction and non-fiction is related to the cultural situation that the genres are placed in, as different situations require different forms of communication. However, it is important to note that Argamon et al. (2003) do not specify in what way the cultural situation

of the texts used are gendered. For example, from social sciences 60 texts by men and 60 text by women were included in the dataset. They do not reflect whether people of a certain gender publish more texts in this field. This could influence the expected writing style and form of communication of a particular genre.

Influence of genre on writing style To isolate the influence of genre on (gendered) writing style, Herring and Paolillo (2006) investigate the influence of gender in language, when the genre of text is constant. They analysed weblogs of two different genres: diary and filter, in which diary blogs report on the author's life, and filter report on events external to the author's life. They divide the gendered features of Argamon et al. (2003) into preferential features by women, such as personal pronouns, and preferential features by men, such as determiners. For both of these features logistic regression models were used to confirm which features interacted with gender significantly. Surprisingly, no significant correlation between the stylistic features and gender was found. Significant correlations were found between woman preferential features and personal blogs and man preferential features and filter blogs. Thus, they conclude that genre is a stronger predictor than author gender of the 'gendered' stylistic features found by Argamon et al. (2003). They argue that genres appear to be gendered, due to the topics discussed, since diary writing is traditionally associated with women. On the contrary, politics is one of the most common topics in the filter blogs, and is also traditionally associated with men. They also hypothesise that men and women use similar language within a genre, and that therefore influence of gender on language is not identified within one genre. This shows that it is hard to identify the difference between features that are related to gendered language use, and features that are related to gendered (sub)genres. It also stressed the importance of carefully drawing conclusions on gendered language use, as gendered language use can be strongly related to the topics within a text (Herring and Paolillo, 2006; Koolen and Cranenburgh, 2017).

2.2.2 Approaching gender in NLP in a more social manner

As shown in Section 2.2.1, it is important to use and analyse gender as a nuanced, social variable. In this Section, Bamman, Eisenstein, and Schnoebelen (2014) and Nguyen et al. (2014) will be discussed. Both research strongly connect the relation between gender and language use with the perception of language and gender in society.

The relationship between gendered word use and social networks Bamman, Eisenstein, and Schnoebelen (2014) used clustering to analyse how difference in gender use relate to topics of Tweets and to the social network individual Twitter users have. They used a corpus of 14.000 Twitter users to identify the relationship with gender and writing style. As gender is a non-binary social variable (Oyewumi, 2002; Butler, 1998), they implemented a more nuanced approach to analyse author gender. In addition to a binary author gender classification, they clustered the Twitter users based on their tweet to find a more natural grouping of writing styles and topics. They found that users who have a social network that includes fewer same-gender social connections, use language that is not matched with the classifier's model for their gender. Also, they identified multiple clusters containing authors of different genders. Thus, writing style and topics in tweets appear not be as strongly related to gender.

Firstly, Bamman, Eisenstein, and Schnoebelen (2014) trained a logistic regression classifier predicting author gender. Instead of using gender as a independent

variable, they have chosen to use author gender as a dependent variable. As independent variables, the counts of the 10.000 most frequent lexical items in the corpus were used. The logistic regression classifier reaches an overall accuracy of 88% in binary gender prediction, thus the features selected capture language which can distinguish and predict gender.

To analyse their results and to further compare them to Argamon et al. (2003), probabilistic clusters are created using the Expectation Maximization framework. This framework is designed to iteratively group authors together by similarities in word usage. Fourteen out of the twenty clusters show a clear gender orientation, containing at least 60% women or men authors. The clusters also show multiple expressions of gender, such as interactions between gender and age or race, underlining the importance of intersectionality. The clusters are also related to certain topics, such as athletes and sport-related organisations and politics. From these topics, Bamman, Eisenstein, and Schnoebelen (2014) conclude that in their data, men are more likely to write about hobbies and careers. As these topics are related to large numbers of named entities, men use more named entities in their language. They state that these specific topics are the most probable explanation for the usage of named entities by men, and not 'informativity' or 'explicitness' as Argamon et al. (2003) argue.

Lastly, Bamman, Eisenstein, and Schnoebelen (2014) analysed the relation between author gender and writing style using the social network of their data, which was found through the direct conversation between the Twitter users. They find that the more gendered the language of an author is, the more gendered their Twitter network is. For example, the segment of women which have been classified as women with strong confidence, have an average network composition that consists 77% of women. The segment of women who are classified as women with the lowest classifier confidence, have social networks that consists 40% of women. Similar results were found for men as well. Correspondingly, the usage of lexical same-gender markers increases with the amount of same-gender connections. These results support the theory that language use is related to the audience and social network of the speaker. Bamman, Eisenstein, and Schnoebelen (2014) conclude that language use does not represent gender as a binary category. Whenever linguistic resources points to a certain gender, it expresses and generates the multifaceted positioning inherent to language use.

An interactive approach to analyse gendered word use Another approach in which the multifaceted positioning of language use is show, is in interactive research. Nguyen et al. (2014) implemented an online game, in which players guessed the gender and age of a Twitter user. Participants were shown multiple tweets, without the Twitter username of the author of the tweets nor the usernames of people that they mentioned. Gender was guessed in a binary manner (man or woman), age (in years) could be manually filled in. After guessing, the participants could see the guess of gender and age by a computer, the average guesses of other players and the correct age and gender. For the computer guesses, a logistic regression model was used to predict gender and a linear regression model to predict age in years.

The results suggest that 10.5% to 16% of the Dutch Twitter users do not use language corresponding with language the crowd expected to be used by people of the users' gender. To analyse this further, a gender continuum was created, using the percentage of players who guessed the user to be a man and the percentage of guesses for woman were calculated per Twitter user. This continuum shows that the players guesses were based on the expected linguistic behaviour of women and

men. It also shows that the distribution of percentages of players that guess man and woman cannot be grouped into two distinct groups. These results underline not only that gender should be treated as a social variable, but also that the influence of gender on language use and perception is limited and nuanced.

From the results of Bamman, Eisenstein, and Schnoebelen (2014) and Nguyen et al. (2014) it does not seem that women communicate more 'involved' and man more 'informative' due to socialisation (Argamon et al., 2003), but rather that people communicate and expect other people to communicate in a certain way, based on the social group that they are communicating with, such as Tweets focused on a (gendered) topic (Bamman, Eisenstein, and Schnoebelen, 2014). As Nguyen et al. (2014) did not find distinctive gendered groups guessed a certain gender per Twitter user, it seems that the perception of gendered language use is not related to the gender of the guesser either.

2.2.3 Gender stereotypes & NLP

As seen in the previous section, texts can be accurately classified by author gender, but the differences in writing styles are hard to interpret due to complex influence of society on the expectation and perception of (gendered) language use. Distinctive features, such as determiners and pronouns, are often hard to interpret. Furthermore, the relation between the genre and topic of a text and gender seems to influence the differences in writing style as well (Herring and Paolillo, 2006; Bamman, Eisenstein, and Schnoebelen, 2014). In this section, I will focus on research in which gender stereotypes within texts are researched. I will focus on the variety of methods used to identify the stereotypes present in fiction, tweets and schoolbooks.

Firstly, Fast, Vachovsky, and Bernstein (2016) aimed to identify signals of gender bias in an online community of amateur fiction writers. They found that gender stereotypes are mostly consistent across genres. However, they found mixed results on how gendered stereotypes are received. Conventional stereotypes are not always associated with a high story rating, as they expected. Some stereotypes, such as *sexual* and *violent* men have a positive impact on story rating, whereas *beautiful* women has a negative impact. Gender stereotypes could not be used to accurately predict author gender. The characters written by men and women are extremely similar, as the adjectives and verbs used to describe the characters are not significantly associated with one author gender. This indicates that even though gender stereotypes do not differ between authors of different genders, but gender stereotypes can influence literary ratings.

Another interesting approach to analyse book characters, is by Lucy, Demszky, Bromley, and Jurafsky (2020). They used a combination of NLP methods to uncover the depiction of historically marginalised groups in high school historical books in Texas, such as Named Entity Recognition, coreference resolution, log odds ratio, word embeddings and topic modelling.

Lucy et al. (2020) found that Latinx people are absent in U.S history textbooks used in Texas. Moreover, the named individuals are mostly white men. The word associations found using word embeddings show stereotypes, such as that women are mentioned in relation with marriage, home and work and Black people are associated with actions with low agency and power. The topic modelling shows that the textbooks focus more in political history than social history, and that minority ethnicities are mentioned in relation to white people. Thus, their combination of multiple methods is successful in identifying the description of different groups of people across multiple books.

Lastly, Devinney, Björklund, and Björklund (2020) has a very interesting approach to analyse gender bias in text. They compared differences in topic models for three different corpora of news articles. They used semi-supervised topic modelling, to discover which words are associated with different genders. Semi-supervised topic modelling, is a type of LDA in which the topic assignment by the variable z is assigned by supervised information called z-labels (Andrzejewski and Zhu, 2009). The z-labels assign given words within a subset of topics. In order to semi-supervise the training, Devinney, Björklund, and Björklund (2020) seeded some topics with gendered words. This forces the model to treat these words as belonging to the same, explicitly gendered, topic. For the seed words, they used three gender categories, masculine, feminine and non-binary or neutral. They also implemented unsupervised topic models to identify which topics are implicitly gendered. In their analysis, they compared the top 50 words of the topic models, for both the gendered and the unsupervised topics. They found that women are strongly associated with family, relationships and communication, whereas men are associated with a greater variety of topics. Non-binary people are nearly invisible in not explicitly queer corpora.

Thus, this overview shows that gender stereotypes in text can be detected using NLP techniques, such as topic modelling. For example, topics can be identified that are explicitly related to one gender. Topic modelling can also be used to analyse which groups of people are discussed in texts. Lastly, this overview shows that these stereotypes might influence literary ratings.

2.3 Conclusion

To conclude, the overview given in this chapter shows that it could be possible to use computational stylistics to distinguish nominated and not nominated novels based on word use (Herrmann, Jacobs, and Piper, 2021; Koolen and Cranenburgh, 2017; Cranenburgh and Bod, 2017). However, nominated and not nominated have not been classified using NLP techniques before. Therefore, this thesis will use well researched techniques to classify nominated and not nominated novels based on word use, as this seems to be the most promising approach.

This chapter has also shown the importance to carefully and critically analyse the results when researching author gender. For example, it is important to consider the relationship between author gender and genre of texts (Herring and Paolillo, 2006). Nguyen et al. (2014) and Bamman, Underwood, and Smith (2014) also show that word use is related to social networks, and the societal expectations of certain word use of a social group. Additionally, gender stereotypes within text can also be detected using NLP techniques. This is useful, as it can help analyse the gender stereotypes and groups represented within text, but can also show how gendered stereotype influence ratings of literary quality (Fast, Vachovsky, and Bernstein, 2016). This thesis will follow this nuanced approach to analyse gender inequality within the context of Dutch literary prizes.

Chapter 3

Dataset

As stated in the introduction, the goal of this research is to investigate to what extent the writing styles and topics of authors of different genders associate with winning a literary prize. In order to do so, I will use logistic regression to categorise novels in one of the following three categories: 1) novels that have been nominated for a literary prize, 2) novels that have not been nominated, but have been written by a writer who has written nominated works, or 3) not nominated novels written by a writer who has not been nominated for a literary prize so far. The nominated novels have been selected from the first year the prize was awarded, until 2020. The not nominated novels by nominated authors, are all not nominated novels published since the prize was awarded by a nominated author. The not nominated novels by not nominated authors are novels that have been published in the same time period, and could have been considered for a nomination. The selection procedure will be further discussed in Section 3.1. Then, I will use topic modelling to determine whether the topics in the novels of these three categories differ from each other. Thus, my dataset will consist of Dutch literary novels, in either of these three categories. For readability, I will refer to a set of nominated novels as **NomNov**, a set of not nominated novels by nominated authors as **NomAut** and a set of not nominated novels by not nominated authors as **NotNom**. The titles and authors of the corpus, can be found in Appendix A.

	Literature	Suspense
Literary Prize	Boekenbon Literatuurprijs & Libris Literatuur Prijs	Gouden Strop
Nominated novels	**NomNov**$_{\text{Lit}}$	**NomNov**$_{\text{Susp}}$
Not nominated novels, written by nominated writers	**NomAut**$_{\text{Lit}}$	**NomAut**$_{\text{Susp}}$
Not nominated novels, written by not nominated writers	**NotNom**$_{\text{Lit}}$	**NotNom**$_{\text{Susp}}$

TABLE 3.1: Description of all six dataset that were collected for this thesis. Nominated novels (**NomNov**, not nominated novels by nominated authors **NomAut** and not nominated novels by not nominated authors (**NotNom**) were selected for literature and suspense novels. Due to limitations in the resulting corpora, the suspense novels were not used further in this research.

Suspense dataset Initially, I wanted to perform all experiments on both a literature and suspense dataset. The goal was to identify the difference in results between different genres. Nominated novels from the *Libris Literatuur Prijs*, *Boekenbon Literatuurprijs* and *Gouden Strop* were used. These three prizes were chosen, as they are the main prizes for literary and suspense novels. Since the *Libris Literatuur Prijs* and *Boekenbon Literatuurprijs* are both prizes for literature novels, I have combined the nominated novels and nominated writers *Libris Literatuur Prijs* and the *Boekenbon Literatuurprijs* for the creation of the **NomNov**, **NomAut**, **NotNom** sets. The *Gouden Strop* is a prize for suspense novels only, and can therefore not be combined with the other two prizes. Thus, I collected six different sets of books in total, which is shown in Table 3.1. For both the literature and the suspense novels, I selected **NomNov**, **NomAut** and **NotNom** sets following the same method. This approach is further discussed in Section 3.1. It is important to note that this distinction between these two genre implies that the suspense novels are not literary or are of lesser quality. There is even an overlap between the nominated novels, as novels from René Appel and Charles den Tex have been nominated for both type of prizes.

Unfortunately, as can be seen in Table 3.2, the number of unique authors in the suspense datasets is rather limited. This small number of unique authors complicates researching the relation between author gender and nominations for literary prizes. To illustrate, only 44 novels written by women were collected in the **NomNov**, **NomAut** and **NotNom** suspense sets. To be able to draw clear and general conclusions using these datasets would be hard, as the results could be more related to author specific writing style, than to writing style related to (not) nominated novels or author gender. Therefore, I decided not to use the any of the suspense datasets in this thesis.

	NomNov	**NomAut**	**NotNom**	Total
Books	32	52	35	119
Unique authors	19	22	7	35
Books by women writers	9	24	11	44
Books by men writers	23	28	24	75

TABLE 3.2: Collection of suspense novels, divided by category. The number of unique authors and the numbers of novels written by women is rather limited, and therefore this dataset was not used further in this research.

3.1 General selection procedure corpus

In this section, I will give an overview of the general procedure for collecting the **NomNov**, **NomAut** and **NotNom** sets. Since I've used four different sources to collect novels from, this is the general selection procedure. In Section 3.2, I will explain in more detail how these selection procedures differed per source collection.

Collection NomNov First, I listed all nominated novels for the literature prizes as well as the *Gouden Strop*, including title, author, publisher and year. Then I selected the nominated novels from four different collections of books, using a combination of title and the last name of the author. These collections are books from DBNL, the Dutch Digital Library, the Riddle of Literary Quality, nominated Dutch novels

in 2007-2012 and a set of about 7000 popular epubs. The books in these collections were matched to the title and the last name of the author found in the dataset, using fuzzy matching. These collections will be discussed in depth in Section 3.2.

Collection NomAut Set **NomAut** was collected by selecting all novels written by nominated authors, excluding the nominated novels and novels that were published before the first prizes were awarded. The novels latter were excluded, since the novels in theory could have been nominated for a literary prize, if the novel would have been published in a period where the prize was awarded. The novels were selected using fuzzy matching on the first and last name of the author. The authors from the selected novels were manually compared with the names of the nominated authors, to make sure that no books were excluded due to slightly differing name spellings.

Collection NotNom Lastly, set **NotNom** was collected with the aim to create a set of not nominated novels by not nominated authors, which was most similar to set A. The goal of set **NotNom** is to create a set of books that could have been candidates for nomination. Thus, after the literature and suspense sets **NomNov** were created, the **NotNom** sets were created by selecting books that were published by publishers that have published nominated novels, from the four book collections that were available to me. Since the grosslist of all the books that have been submitted to consider for nomination is not available for every edition of the three prizes, it is more consistent to select books from publishers that publish books that have been nominated in the same publishing year as the books in set **NomNov**. The books have been selected considering author gender. Thus, if set **NomNov** contains three novels that have been nominated in 1995, by two men and one woman, I aimed to select three novels, published in 1995 and written by two men and a woman as well.

3.2 Collection procedure from the different source collections

In this section, I will describe the four collections from which the literature and suspense sets **NomNov**, **NomAut**, **NotNom** were selected. First, I will describe the dataset of about 7600 popular ebooks. Then, I will describe the set of books from DBNL. Lastly, I will describe the collection from the Riddle of Literary Quality and the 50 nominated novels. In all datasets, I focus on works published in 1980 or later, as the three literary prizes were all first awarded in the 1980s.

Popular epubs The dataset of popular epubs contains 7639 Dutch books. The books vary from a wide range of genres, including children's novels, literature and the Dutch dictionary. The oldest books in the collection were first published in the 19th century, but also contains contemporary novels. The collection includes both original Dutch and translated works.

The dataset of popular epubs contains very limited information about the epubs, as it only contains title and author name. This is sufficient for the collection of sets **NomNov** and **NomAut**, as these books can be selected by title and author name (**NomNov**) or author name only (**NomAut**). However for the creation of datasets **NotNom** the original publisher and first publishing date is needed. Therefore, the

linked data-environment of the Dutch Royal Library[1], was used to collect this data using the ISBN of the books.

Collect NotNom dataset from corpus of popular epubs To collect the ISBN numbers of the books in the popular epub dataset, two different methods were used. The majority of the ISBNs were collected from a metadata file containing 113493 urls of the review website Hebban[2], including information such as NUR code, ISBN, author and title. Using the last name of the author and the title, the epubs from the collection were matched with the ISBNs found in the metadata of the reviews. A small part of the ISBNs was collected from the text of the epubs itself, using regex. In total, 3016 ISBNs were collected, which is about 40% of the collection of popular epubs. Thus for the collection of **NotNom**, the other novels in the corpus of popular epubs was not used.

The ISBNs were then used to collect the publishing year and publisher from the linked open-data of the Dutch Royal library using SPARQL. It is important to note that some books have several ISBNs, due to different types of printing, or due to a reprint by a different publisher. From the ISBNs, information on 2362 epubs were collected from `data.bibiliotheken.nl`, which is about 31% of the full dataset of popular epubs. For the majority of the titles multiple editions were found. In that case, the oldest publishing year and corresponding publisher were selected. However, it is still possible that the obtained publishing year does not correspond with the first publishing date. This could for example occur when the first print of a book is not included in the open-data of the Dutch Royal library, but a later version is. Therefore, the first published year and publisher is an estimate.

Since the collection of popular epubs also contains translated works, the authors of the 2362 books were manually checked on nationality. This was done by googling all the authors of which I was not sure that they were Dutch. The authors of which I was sure to be Dutch, were popular authors of which I was sure to mainly publish in Dutch. In doubt, I googled them, even when the authors had Dutch sounding names, or if they published other books in Dutch. It was important to be secure, as some authors published books in several languages. For example, Ayaan Hirsi Ali first published several books in Dutch, but later on published works in English which were translated to Dutch. These originally English books were excluded from the datasets.

DBNL The *Digitale Bibliotheek voor de Nederlandse Letteren* (DBNL) is the Dutch Digital library that includes texts of Dutch literature, linguistics or cultural history from the earliest time until now. The collection represents the Dutch language area. As the DBNL contains a great amount of books, the dataset of books from DBNL [3] was selected from all the available Dutch books of DBNL on two criteria: it had to be categorised as prose as main genre and published in 1980 or later. However, the resulting selection also included magazines and anonymous works. The vast majority of anonymous works are novels stem from the Middle Ages. Since I want to focus on books first published in 1980 or later, the most straightforward solution was to filter out all anonymous works. This resulted in a set of 511 Dutch novels.

However, not all the novels in the dataset were suitable for this research project, as DBNL digitises books of which the copyright has expired. Thus, a great amount of

[1] `http://data.bibliotheken.nl`
[2] `https://www.hebban.nl`
[3] `https://www.dbnl.org`

the books in this set contains works written before 1980, such as a retelling of Floris ende Blancefloer, which is a story from the 13th century. The set also contains work from writers who have passed away before 1980, such as Anna Blaman. As most nominated novels or other works by nominated others are more recently published and still have copyright, this corpus was mainly used to collect books for the sets **NotNom**.

Riddle of Literary Quality Thirdly, the books of the research project the Riddle of Literary Quality were used. This project researched whether the formal characteristics of a text can be identified to influence readers' perception of a novel being literary or not literary. This Riddle of Literary Quality used 401 popular Dutch novels, firstly published in 2007-2012. To measure readers' perception of literary quality, opinions were gathered in the National Reader Survey of 2013. The dataset contains original Dutch as well as translated novels, and has metadata on author gender, year published and publishers. As it is focused on popular novels, the corpus of the Riddle of Literary Quality was mostly used to to create the sets **NomNov** and **NomAut**.

Dutch nominated novels Lastly, the books from the 50 nominated Dutch novels used in Koolen and Cranenburgh (2017). This corpus was necessary to use, as the corpora of popular epubs, the DBNL and the Riddle of Literary Quality only contained a limited amount of nominated novels. In Koolen and Cranenburgh (2017), this corpus of Dutch nominated novels were selected in order to correspond with the corpus of the Riddle of Literary Quality, and were therefore published in the same time period. The metadata of this corpus is similar to the metadata of the Riddle of Literary Quality and also contains author gender, published year and publisher.

Estimation author gender For the DBNL and popular epubs datasets, the author gender had to be estimated. This is done using data from the '*Nederlandse Voornamenbank van het Meertens Instituut KNAW*' [4], which contains first names which were used by more than 500 people registered in the Netherlands, with a Dutch nationality in 2010. This was the most recent available list of Dutch first names. This resulted in a list of 1184 men names and 1493 women names. To ensure the correct author gender was collected, the gender was manually checked when creating dataset **NomNov**, **NomAut** and **NotNom**. This was done by searching for the author online, and see what pronouns or gendered words, such daughter, were used in trustworthy sources, such as the website of the author itself, the website of the publisher of interviews in well known newspapers such as the *Volkskrant*.

3.3 Literary Prizes

In this section, I will analyse the overall nominations of the *Libris Literatuur Prijs*, the *Boekenbon Literatuurprijs* and the *Gouden Strop*. After this analysis, I will compare the distribution of published year, author gender and publisher to the distribution in sets **NomNov**, **NomAut** and **NotNom**.

3.3.1 Analysis Libris Literatuur Prijs

The *Libris Literatuur Prijs* was first awarded in 1987, then named the AKO Literatuur Prijs. It was modelled after the Booker Prize (Stichting Literatuur Prijs, 2021). Every

[4] www.meertens.knaw.nl/nvb

year publishers can submit books, which they think qualify for nomination. This list of titles is called the 'grosslist'. On average 172.2 novels per year are submitted for the grosslist (Dijkgraaf and Appel, 2013). To qualify for the *Libris Literatuur Prijs*, the books have to be published in the previous year, between the first of January and December 31st. From 2010 on, only original Dutch literary adult novels could be submitted. Before 2010, children's novels could be submitted as well. There is no limit on the number of books that publishers can submit, due to the protest from the publishers against such a rule (Stichting Literatuur Prijs, 2021).

The jury, consisting of maximum six people, will have a year to read the novels and select 18 books for the 'longlist'. A few requirements are made for the jury. The jury members have to be a literary author, critic or literary scholar. Also, the jury chairperson has to be a well known public figure. Lastly, at least one jury member has to be Flemish. In the beginning, the longlist was not published, but later on this was made public. In March, the six best novels from this longlist are announced, the so-called 'shortlist'. The authors on the shortlist, receive €2500. In May, the winner of the *Libris Literatuur Prijs* is announced, who wins €50.000.

3.3.2 Analysis Boekenbon

The *Boekenbon Literatuurprijs* is focused on original Dutch fiction in the category 'literary prose for adults' or original Dutch works in the category 'literary non-fiction', such as biographies, essay collections and travel stories which are of similar quality as literary prose. The jury determines whether the non-fiction has the literary quality of prose. The novels have to fulfil three requirements in order to be considered for the prize. Firstly, the books have to have first been published in between the first of July of the previous year until July first of that years prize. For example, the *Boekenbon Literatuurprijs* 2021 will be awarded to a book firstly published in the period 1 July 2020 - 1 July 2021. Secondly, the books have to be written by one author or a collective of authors that can be viewed as one author. Lastly, the author has to have been alive when the book was published. From all the submitted books, a maximum of fifteen books are selected for the longlist. From which a minimum of three and a maximum of five of these books are selected for the shortlist.

Similar to the *Libris Literatuur Prijs*, only publishers submit the novels to be considered for the *Boekenbon Literatuurprijs*. Note that the *Boekenbon Literatuurprijs* asks the publishers of the shortlist novels to pay for certain costs made. The publisher is requested to contribute €2500 for promotion costs, besides providing a sufficient amount of books for the jury and the media. Similar to the *Libris Literatuur Prijs*, the winner receives €50.000. Contrarily, the shortlist nominees of the *Boekenbon Literatuurprijs* do not win any money.

The gender distribution of the overall shortlist nominees of the *Boekenbon Literatuurprijs* is similar to the *Libris Literatuur Prijs*. About 75% of the shortlist nominees of the *Boekenbon Literatuurprijs* until 2020 are men and about 24% of the nominees women. Two nominated novels were written by multiple authors. About 82.3% of the winners are men and about 17.7% women, which is slightly greater portion of women winners. Unfortunately, the longlists of the *Boekenbon Literatuurprijs* are not publicly available and thus that gender distribution cannot be analysed.

3.4 Analysis Literature Datasets

The literature datasets consists of **NomNov**, **NomAut** and **NotNom**. As can be seen in Table 3.3, all three datasets have a similar number of novels. The division between novels written by women and men is similar between in the sets of novels written by nominated writers, **NomNov** and **NomAut**.

	NomNov	**NomAut**	**NotNom**	Total
Books	102	100	98	300
Unique authors	73	35	83	191
Books by women writers	36	42	43	121
Books by men writers	64	60	55	179

TABLE 3.3: Number of novels in literature dataset, divided by category.

The distribution of the publishing years of the novels can be found in Figure 3.1. As stated in Section 3.2, the books in dataset **NotNom** were chosen to resemble the distribution of author gender and publishing year as seen in dataset **NomNov**. Overall, the distribution of publishing years is rather similar in sets **NomNov** and **NotNom**. In both datasets, the publishing years range from 1989-2012 and in both datasets, and most books are from the period 2005-2011. However, for some years the same number of not nominated novels by not nominated authors could not be found in the corpora, as were included in set **NomNov** for that particular year. This is most clearly seen in the period 1992-2001, as none of the books in dataset **NotNom** were published in this period. If a certain number of books could not be found in the four corpora, the aim was to select books from a year close to the publishing year of the books from **NomNov**.

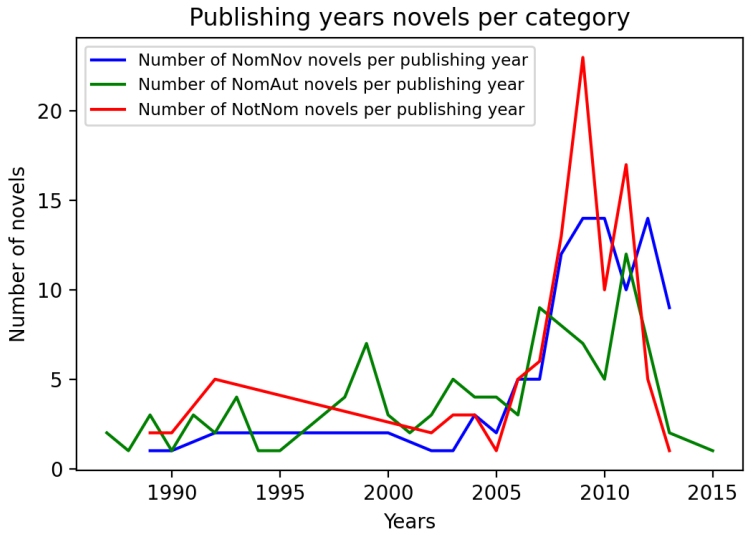

FIGURE 3.1: Distribution of publishing years of the corpus, sorted by **NomNov**, **NomAut** and **NotNom** category. The years are an estimate of the first publication.

Number of novels	Publishing year set **NomNov**	Publishing year set **NotNom**
1	1995	1989
2	1999	1992
2	2000	1992
1	2006	2003
1	2008	2003
9	2012	2009
8	2013	2011

TABLE 3.4: Selection novels of novels for which there were not sufficient **NotNom** novels in that specific publishing year to match the number of **NomNov** novels. In the third column the differing publishing year of the selected not nominated novels for set **NotNom** is shown.

Selection novels from missing publishing years Due to the great number of available books from 2006 on wards, and the limited amount of novels in the years before 2005, I had to find a manner to select not nominated novels from not nominated authors that could still resemble the distribution of the nominated novels. I choose to select novels for the **NotNom** set first published before the publishing dates of the **NomNov** set. For example, **NomNov** includes two books from 1995. Since the four corpora only contain one suitable not nominated novel from 1995, one novel from 1989 was selected for **NotNom** to represent the not nominated novel that could be compared with the second book of 1995. These years were chosen as these were the publishing dates from before 1995, that were closest to 1995. In Table 3.4 is shown per year the number of novels in set **NomNov** that could not be matched with the same number of not nominated novels. In the third column the publishing year of the selected not nominated novels for set **NotNom** is shown.

As can be seen in Table 3.4, due to the high number of nominated novel in my dataset in the years 2012 and 2014, a high number of novels had to be collected from previous years. Eight not nominated novels from not nominated writers from 2009 were used, instead of 2012, and nine not nominated novels from 2013. For the other years, only one or two books had to be used that were published in previous years. Therefore, the distribution of publication years in **NomNov** is approached for **NotNom**, but not exactly matched (see Figure 3.1).

Composition literary nominations set NomNov In set **NomNov**, the majority of the books have been nominated for the *Libris Literatuur Prijs*, as can be seen in Table 3.5. This is due to the longlists from 2005-2020, which have been made public. For the nominations of the *Libris Literatuur Prijs* before 2005 and the *Boekenbon Literatuurprijs*, the longlists have not been made public. Sixteen novels in dataset **NomNov** have been nominated for both the *Boekenbon Literatuurprijs* and the *Libris Literatuur Prijs*.

Most occurring authors In all three datasets, most authors only occur once or twice. However, each dataset contains multiple works of a few authors. In dataset **NomNov**, Arnon Grunberg occurs the most, with four nominated novels. Four authors have three works in dataset **NomNov**, namely Christiaan Weijts, Nelleke Noordervliet, Dimitri Verhulst and Anna Enquist. This leads to 74 unique authors

	Boekenbon Literatuurprijs	Libris Literatuur Prijs
Winners	11	7
Shortlist Nominees	27	22
Longlist Nominees	-	49

TABLE 3.5: Types of nominated novels in the corpus. The longlist of the Boekenbon Literatuurprijs is not publicly available, and therefore these novels are not included in the corpus.

in dataset **NomNov**. In dataset **NomAut**, fourteen not nominated novels of nominated writer Renate Dorrestein are included. Eight not nominated novels of Jan Siebelink are included as well. Two authors have seven not nominated novels in dataset **NomAut**, namely Harry Mulisch and Kristien Hemmerechts. Toon Tellegen, Vonne van der Meer and Herman Brusselmans had three not nominated in the dataset. Five nominated authors have not nominated works in **NomAut**, namely Jeroen Brouwers, Kader Abdolah, J. Bernlef, Remco Campert and Joost Zwagerman. Lastly, Adriaan van Dis, Dimitri Verhulst, Rascha Peper and Connie Palmen have three not nominated novels in dataset **NomAut**. This leads to 33 unique authors in dataset **NomAut**. In dataset **NotNom**, only two authors have multiple novels in the dataset. The author that occurs the most is Martin Bril, with five not nominated novels. The other author is Kluun, who has two novels in the dataset. All other authors have two works or less in the dataset, leading to 83 different authors in dataset **NotNom**.

Chapter 4

Method

This chapter introduces and discusses the different methods and techniques used to answer the main research question: Can quantifiable literary qualities be used to investigate author gender inequality in Dutch literary prizes?

As stated in the introduction (See Chapter 1), this main question will be answered by answering the following sub-questions:

1. RQ1: Can nominated and not nominated novels be identified based on textual features only?
2. RQ2: Is there a relation between classifications on nominated and not nominated novels and author gender, where both classifications are based on textual features?
3. RQ3: Are the differences in topics/writing styles between books that are nominated for literary prizes and those that are not, related to author gender?

To answer these sub-questions, three different NLP techniques are used, namely logistic regression, LDA topic modelling and cosine delta. The first technique, logistic regression, is a supervised algorithm, and the latter two are unsupervised. Therefore, logistic regression was used to derive quantifiable conclusions, whereas LDA topic modelling and cosine delta were used for more qualitative interpretations of the results obtained with logistic regression.

The results of the logistic regression were used to answer RQ1 and RQ2, and the results of the LDA topic modelling and cosine delta for RQ3. Due to the different objectives for the techniques used, the research designs will be discussed per technique.

Furthermore, this chapter introduces the methods used in this research. First, I will explain the research design, by giving an overview of the used techniques, and explaining in what way the techniques are used to answer the research questions. Adding to that, the dataset will be shortly described in Section 4.2, as the creation and content of the dataset was thoroughly discussed in Chapter 3. Lastly, an overview of the procedures used is given. In this section, the details of implementation will be discussed, such as the parameter settings and preprocessing. The different techniques will be discussed per sub-question.

4.1 Research design

In this section, a general overview of research designs will be given. For each technique, a motivation why this technique is used, and a general overview on how the technique is implemented, is given.

Logistic Regression To answer RQ1 and RQ2, logistic regression is used to analyse word features. These word features consist of the most occurring unigrams and bigrams in the complete corpus, created with a Tf-Idf vectorizer. This was chosen as Cranenburgh and Bod (2017) indicated that bigrams can be effectively used to predict literary judgements, and Bamman, Underwood, and Smith (2014) shows that bag of words features can be used to predict and analyse author gender, using logistic regression, in a nuanced manner.

In order to answer these two questions, two different types of logistic regression models were trained. Firstly, models were trained to classify nominated and not nominated novels. Thus, the dependent variable is whether a novel had been nominated or not nominated. The independent variables are 5000 of the most frequent unigram and bigram word features. To further examine the relation between author gender and nominated and not nominated novels, the author gender variables were added to the 5000 most frequent unigram and bigram word features.

The other type of logistic regression model was trained to classify author gender. Thus, the dependent variable is author gender, and the independent variable are the 5000 most frequent unigram and bigram word features.

LDA Latent Dirichlet allocation (LDA) is an unsupervised iterative probabilistic model of a corpus (Blei, Ng, and Jordan, 2003). The idea is that documents in the corpus are represented in a given number of topics. Those topics are defined by the probability that a certain word occurs in a topic. The number of topics in the model is defined beforehand. The algorithm iteratively defines the likelihood of a word occurring in a topic by at first randomly assigning topics to words in the documents. Then, it is assumed that all word assignments to a certain topic are correct, except for one. For the words in this topic, the new topics are assigned by using the probability that the word might occur in that topic. In this way, words are iteratively reassigned to the topics until convergence, based on the probability calculation of the likelihood that a word occurs in a certain topic. These probabilities are then used to calculate which topics occur in the documents in the corpus. It should be noted that multiple topics can be related to one document.

LDA is used in a wide range of different fields, such as Twitter-analysis, biomedical science and literature (Jelodar, Wang, Yuan, Feng, Jiang, Li, and Zhao, 2019). One of the advantages of using LDA to analyse literature, is that it can reveal patterns that are not easily observed. For example, when analysing changes in literature over time, LDA can identify patterns that are not easily recognised because they happened gradually, or simply have slipped under the radar (Goldstone and Underwood, 2014). The same authors also argue that another advantage, is that the unsupervised topics and clusters created, force researchers to analyse literature outside of predefined, traditional concepts. A disadvantage of LDA topic modelling, is that researchers overestimate their ability to explore large corpora quickly, despite the fact that the topics created might not be as coherent and stable as they seem (Schmidt, 2012). The set of words related to one topic, do not per definition have anything in common. Thus, if a topic occurs in two different documents, it does not necessarily mean that this particular set of words related to a topic has the same relation to that topic within these two documents. Therefore, the interpretation of LDA topic modelling is sensitive to the interpretation of the researchers, and conclusions should be carefully drawn from it.

LDA topic modelling is useful to answer RQ3, as it relates the word use across the different novels and groups them into topics. This creates the opportunity to analyse which topics relatively occur most in certain documents. For example, the

most occurring topics in nominated novels written by women can be identified, in comparison to nominated novels written by men, or in comparison to not nominated novels written by women. Thus, a topic model can be used to analyse which topics strongly relate to nominated novels and not nominated novels, but also to analyse whether these topics occur most in novels written by authors of a specific gender. Therefore, topics that relate to nominated and not nominated novels can be identified using LDA, and the relation between author gender across these categories can be identified as well.

Cosine delta To investigate the difference in writing style between books that have been nominated and books that have not been nominated more closely (RQ3), cosine delta is used to identify the difference in writing styles between novels that have been correctly classified in the nominated or not and author gender classifications, and novels that have been misclassified in all these classifications.

Cosine delta is a successful technique to identify authorship and writing style using the most frequent words of novels (Evert et al., 2017). Cosine delta is based on Burrows' delta, which is an algorithm in which the frequencies of 100-5000 most recurrent words of novels are used to calculate the difference in writing style between novels (Burrows, 2002). The frequencies of the most recurrent words are standardised to z-scores, to give each word equal weight. In Burrows' delta, the distance between the most recurrent words are calculated using Manhattan distance. In cosine delta, this distance is measured using cosine similarity, which outperforms authorship identification of Burrows' delta (Smith and Aldridge, 2011). The results show which novels in the corpus have a similar writing style, and how the different writing styles of the novels relate to each other.

Therefore, a comparison between the consistently correctly classified and misclassified novels is chosen, as the novels that have been correctly classified in all models have a word use that is consistently related to the features related to nominated novels (**NomNov**), not nominated novels by nominated authors (**NomAut**) and not nominated novels by not nominated authors (**NotNom**) classes and author gender. The misclassified novels have a word use that is clearly hard to relate to the features related to their target classes. Thus, it can be expected that the most clear distinction in writing style can be found between these sets of novels. This comparison will predominantly be used to obtain an indication on the distinctive writing style that is related to nominated novels and to attempt to relate this distinctive writing style to author gender.

4.2 Dataset

In this section, the datasets used will be shortly described. An in depth description of complete dataset, and the procedure used to collect this dataset is discussed in Chapter 3. The main dataset used in all experiments is the complete dataset. The logistic regression models also used two subset of the main dataset: the **NomNov** or **NotNom** subset and the balanced author gender subset.

Complete dataset The complete dataset consists of 300 books, containing 100 **NomNov**, 102 **NomAut** and 98 **NotNom** novels. 179 novels are written by men, and 121 by women. The nominated novels (**NomNov**) contain 64 works written by men and 36 novels written by women. The not nominated novels written by nominated authors (**NomAut**) contain 60 novels written by men and 42 novels written by women.

Lastly, the not nominated novels by not nominated authors (**NotNom**) contain 55 novels written by men and 43 novels written by women. The complete dataset contains works by 170 different authors, of which 67 women and 103 men.

NomNov-or-NotNom subset The **NomNov**-or-**NotNom** subset consists of the entire complete dataset, except for all the **NomAut** novels. Thus, it contains 198 novels, of which 100 **NomNov** and 98 **NotNom**, with the same author gender distribution as described above.

Balanced author gender subset The balanced author gender subset contains an equal number of novels written by men and women. This subset was created by taking all novels written by women, and randomly selecting an equal number of novels written by men in the corresponding **NomNov**, **NomAut** and **NomNov** classes. Thus, the balanced author gender subset consists of 242 novels, of which 121 written by women and 121 by men. The subset contains 72 **NomNov** novels (36 novels written by men, 36 by women), 84 **NomAut** novels (42 written by men, 42 by women) and 86 **NotNom** novels (43 written by men, 43 by women).

4.3 Procedure

In this section, the implementation of the different techniques per sub-question is discussed. As explained in Section 4.1, RQ1 and RQ2 uses bag-of-words logistic regression classification, and RQ3 uses LDA topic modelling and cosine delta in order to answer the sub-questions.

RQ1: Can nominated and not nominated novels be identified based on textual features only? To answer this question, two different categorisations of nominated and not nominated novels were used. The use of these two categories provides the opportunity to research the difference between **NomAut** and **NotNom**. RQ1 is therefore split up and answered by the following two questions:

1. Can logistic regression classify novels into nominated novels (**NomNov**), not nominated novels by nominated authors (**NomAut**) and not nominated novels by not nominated authors (**NotNom**), based on textual features?

2. Can logistic regression classify novels into nominated novels (**NomNov**) and not nominated novels (**NomAut** and **NotNom**), based on textual features?

As explained in Section 4.2, three different types of novels occur: nominated novels (**NomNov**), not nominated novels written by authors who have been nominated (**NomAut**) and not nominated novels written by authors who have never been nominated (**NotNom**). Therefore, two different logistic regression models were applied, one classifying on three classes: **NomNov**, **NomAut** and **NotNom**, and one nominated-or-not model, classifying whether a novel has been nominated (**NomNov**) or not (**NomAut** and **NotNom**). The predicted classifications were evaluated using precision, recall, F1 score per class, and overall accuracy. The models were all implemented with the complete dataset and the balanced author gender subset. The balanced author gender subset was used to analyse the influence of the author gender imbalance on the results of the complete dataset. Each model was implemented following the same procedure.

The results of the nominated-or-not model were also compared to a model trained on **NomNov** and **NotNom** novels only. The goal of this comparison was to find out to what extend the inclusion of **NomAut** novels influence the results. This comparison was necessary as a large number of authors occurred in both **NomNov** and **NomAut**, which could influence the results due to authorship related word use.

Vectorizer In the models, a TfIdf Vectorizer from sklearn was used (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011). The following parameters were used:

- ngram_range = (1,2)
- max_features= 5000
- min_df= 0.2
- sublinear= True

In the vectorizer, unigrams and bigrams were included. A maximum of 5000 features was chosen, as this led to the best results in comparison to 1000, 5000 and 10.000 features for the cross-validation performed on the **NomNov**, **NomAut** and **NotNom** classification. The min_df is 0.2, meaning that if a unigram or bigram occurred in less than 20% of the documents in the dataset, this feature would not be included in the vocabulary. Sublinear_tf = True was used, to apply sublinear tf scaling, i.e. replace tf with 1 + log(tf). For all other parameters, default settings were used.

To analyse the influence of the variable author gender on the classifications, the author gender variables (man or woman) were added to the 5000 most frequent unigram and bigram word features. Thus, vectors of 5002 features were used to classify whether a novel was nominated or not.

The vectorizers were created using the entire dataset used. Thus, four different vectorizers were generated, one on the complete dataset, one for the balanced author gender subset, one for the complete dataset only including **NomNov** and **NotNom** novels, and one for the balanced author gender subset including only **NomNov** and **NotNom** novels.

Cross-validation The logistic regression classifications were implemented using cross-validation, as the number of novels in each of the categories are rather limited. For the cross-validation, LogisticRegressionCV of sklearn is used (Pedregosa et al., 2011), using the following parameters:

- class_weight= balanced
- max_iter: 4000
- cv = OrderedGroupKFold(n_splits=5)
- groups = author

The class weight parameter is balanced, to ensure that the weights are adjusted to the proportional to the class. This was chosen, as the number of samples are not equal for all target classes. For the iterations, a maximum of 4000 is chosen, as a high number was needed in to converge the algorithm. For the cross-validation

OrderedGroupKFold was used created by Andreas van Cranenburgh[1]. This type of cross-validation ensures that certain groups of data cannot occur in train and test folds. In this case, the groups were the authors, to ensure that the model could not be trained and tested on novels from the same author. Thus, all novels of one author either occur in train or in test folds. 5-fold cross validation was chosen, so that the train folds covered 80% of the data, which is similar to the implementation used by Bamman, Eisenstein, and Schnoebelen (2014).

For the other parameters, default settings were chosen. For the solver, this means that 'lbfgs' solver was used. This was chosen after testing the results of 'newton-cg', 'lbfgs', 'liblinear', 'sag' and 'saga' solvers in combination with 1000, 5000 and 10.000 feature vectorizers. The 'lbfgs' solver led to the best results for all vectorizers.

For all for classifications the standard deviation of the F1-scores and overall accuracy were calculated, in order to determine the stability of the outcome when different folds are created. OrderedGroupKFold sorts documents in folds using the place of the novels in the dataset. Thus, it uses the index of the novel of the pandas dataframe (McKinney et al., 2010) containing all the meta data of the dataset. To ensure that ten different folds were created, the dataframe was shuffled, a new vectorizer was created and then the logistic regression model was trained on the newly created folds. This was done ten times, resulting in ten different classification predictions.

RQ2: Is there a relation between classifications on nominated and not nominated novels and author gender, where both classifications are based on textual features? To answer this question, the results of the logistic regression models used to answer RQ1 were analysed by author gender. Additionally, a model was also trained to classify author gender. The model was implemented using cross-validation with the exact same parameter settings and vectorizers that were used to answer RQ1. The model was implemented on the complete dataset, as well as the balanced author gender set. The goal of this comparison was to identify to influence of the imbalance in author gender on the results of the models.

In order to analyse the different relationship of the **NomNov**, **NomAut** and **NotNom** categories on author gender classification, RQ2 was split into four questions:

1. Can the classifications made for these three classes (**NomNov**, **NomAut** and **NotNom**) be related to author gender?
2. Can the classifications made for nominated novels (**NomNov**) and not nominated novels (**NomAut** and **NotNom**) be related to author gender?
3. Can logistic regression classify author gender, based on textual features?
4. Can the confidence of the nominated-or-not classification be related to the confidence of the author gender classification?

Can the classifications made for these three classes (NomNov, NomAut and NotNom) be related to author gender? In order to analyse the relation between the classification on nominated and not nominated novels, the results of the models were additionally analysed by author gender. The predictions made by the models and target classes were split per author gender. Then, the precision, recall and

[1] https://github.com/andreasvc/literariness

F1-score for the classification of **NomNov**, **NomAut** and **NotNom** novels was calculated for works written by men and works written by women. This approach was also implemented on models trained which included author gender variables (man and woman), to be able to analyse the results of these variables.

Can the classifications made for nominated novels (NomNov) and not nominated novels (NomAut and NotNom) be related to author gender? The same approach as for the classification on **NomNov**, **NomAut** and **NotNom** categories was used on the nominated-or-not model. Thus, the predicted classifications of nominated (**NovNov**) and not nominated (**NomAut** and **NomNov**) were split by author gender. Secondly, the model was also trained including the author gender variables. Lastly, results of the **NomNov** or **NotNom** model were also analysed by author gender, to see what the influence of the **NomAut** novels were on the relation between the classification on nominated and not nominated novels and on author gender.

Can logistic regression classify author gender, based on textual features? To answer this question, the complete dataset and the balanced author gender subset was used to classify author gender (man or woman). The predicted classifications and target classes were split in the classes **NomNov**, **NomAut** and **NotNom**. Then, the precision, recall and F1-score for each of these classes were calculated. These results per **NomNov**, **NomAut** and **NotNom** class were compared to the overall results of the author gender classification, in order to relate the performance on author gender classification to the different types of novels in the dataset.

Can the confidence of the nominated-or-not classification be related to the confidence of the author gender classification? To answer this last sub-question, a different approach was used. The confidence of the author gender classification was compared to the confidence of the nominated-or-not model. Concretely, the predicted probability of whether a novel was written by a man, was compared to the predicted probability of whether a novel had been nominated. The goal of this comparison was to see how the classification on author gender and the nominated-or-not classification related to each other. This approach, using the classification confidence, was chosen as it gives more insight in the relationship between the word features that the author gender model relates to a certain gender and the word features that the nominated-or-not model relates to (not) nominated novels. This usage of classification confidence was inspired by Bamman, Underwood, and Smith (2014), as they use the confidence of their author gender classifications in relation to the gender distribution of the social network of Twitter users. Their research is discussed in detail in Section 2.2.2

RQ3: Are the differences in topics/writing styles between books that are nominated for literary prizes and those that are not, related to author gender? To answer this research question, results from the LDA topic modelling and cosine delta were used, as explained in Section 4.1.

LDA topic modelling For the LDA topic modelling, the complete dataset was used. The LDA was trained to form 50 topics, as this number of topics was also used in Koolen and Cranenburgh (2017) and their corpus was a similar sized corpus of Dutch novels.

The topic weights of the individual novels were used to created average topic weights for **NomNov**, **NomAut** and **NotNom** novels, and also for novels written by women and novels written by men. The probability that a topic occurred in nominated and not nominated novels, and on author gender were analysed. In order to identify which topics occured most in **NomNov**, **NomAut** and **NotNom** novels, and which topics occur most in novels by women and novels by men.

Then, the topics were analysed in six classes, namely: **NomNov** novels by women, **NomNov** novels by men, **NomAut** novels by women, **NomAut** novels by men, **NotNom** novels by women and **NotNom** novels by men. The goal was to identify whether certain topics were more strongly related to author gender, or more strongly related to one of the **NomNov**, **NomAut** and **NotNom** classes.

Preprocessing In order to implement the LDA, the novels in the corpus needed to be divided in chunks of 1000 words. First, a list of stopwords were removed from the documents. The stopwords included the Dutch stopwords from NLTK (Bird, Klein, and Loper, 2009), as well as the 5000 most common Dutch first names[2] and commmon Dutch last names. The common Dutch last names were manually added, including names such as *Jansen*. Names which also corresponded to regular Dutch words, such as *Kok* (cook), and *Boer* (farmer) were not included. Since not all book characters had common Dutch names, other first and last names were manually added to the stopword list by running the LDA several times and removing the names that occurred in the topics. Again, names that also are regular Dutch words were not included. The names were also checked in the *personagebank*[3], a project which collects the characters of popular Dutch novels. The names of the other characters occuring in the personagebank were also added to the stopword list. This resulted in a list of 3174 stopwords.

After the removal of the stopwords, the documents were lemmatised using Spacy (Honnibal and Montani, 2017). Then, the documents were tokenized using NLTK, and all words were lowered. Then NLTK was used to split each document into chunks of 1000 words.

Implementation LDA topic modelling For the implementation of the LDA topic modelling, little-mallet-wrapper[4] by Maria Antoniak was used. This model was trained to create 50 topics, and the topic weights per novels were used to create heatmaps that show the relative occurrence of a certain topic in different types of novels. The heatmaps were used to compare the occurrence of topics between nominated novels (**NomNov**), not nominated novels by nominated authors (**NomAut**) and not nominated novels by not nominated authors (**NotNom**). Furthermore, the heatmaps were used to compare the occurrence of topics in **NomNov**, **NomAut** and **NotNom** novels by author gender. The heat map was created by calculating the average topic weight of each topic from the topic weights of all the type of novels. In this way, for the **NomNov** novels, the average topic weights for all 50 topics was calculated, by calculating the average of the topic weights for each of the topics, from all the **NomNov** novels in the dataset. Then, the average probability of each topic occurring in the novels was calculated. Resulting into taking the average of the topic weights of the **NomNov**, **NomAut** and **NotNom** novels, and this average was thereafter subtracted from the topic weights of **NomNov**, **NomAut** and **NotNom**. In

[2] www.meertens.knaw.nl/nvb

[3] http://personagebank.nl/resultaten/

[4] https://github.com/maria-antoniak/little-mallet-wrapper

this way, the average topic weight per type of novel was normalised, resulting in the relative topic weights of **NomNov**, **NomAut** and **NotNom**.

Cosine delta In this thesis, the 1000 most frequent words of the correctly classified novels were used, to investigate how the writing style in correctly classified novels relate to each other. Correctly classified meaning novels that have been classified in the **NomNov**, **NomAut**, **NotNom** classification, the nominated-or-not-classification and the author gender classification. This was chosen, because these novels seem to have a word use which is consistent with the nomination class that the novels belong to as well as the writing style of other authors of the same gender. Then, the model trained on the correctly classified novels, was applied on the misclassified novels. The misclassified novels are novels that have been misclassified in all three logistic regression models. Thus, these novels have a word use that is consistently differentiating from the writing style of other novels in their nomination class and of authors of the same gender.

Cosine delta was implemented on **NomNov**, **NomAut**, and **NotNom** separately, thus creating three different cosine delta models. In each model, the model was trained on the corresponding correct novels, thus in the **NomNov** model on the correctly classified **NomNov** novels. The model was used to identify the differences and similarities between writing style of the correctly classified **NomNov** novels and the misclassified **NomNov** novels. A cosine delta implementation of Andreas van Cranenburgh[5] was used as basis for the models.

[5] https://gist.github.com/andreasvc/c0742ac5b2f7708971ca32b2aecde90a

Chapter 5

Results

In this chapter, the results of the experiments will be discussed. The results are used to answer the sub-questions of the main question:

1. R1: Can nominated and not nominated novels be identified based on textual features only?
2. R2: Is there a relation between classifications on nominated and not nominated novels and author gender, where both classifications are based on textual features?
3. R3: Are the differences in topics/writing styles between books that are nominated for literary prizes and those that are not, related to author gender?

As explained in Chapter 4, the results of the logistic regression models will be used to answer the first two questions, and the results of the LDA and the cosine delta to answer the last. As three different types of logistic regression classification are used to answer the first two questions, I will answer them by investigating several sub research questions:

1. Q1: Can logistic regression classify novels into nominated novels (**NomNov**), not nominated novels by nominated authors (**NomAut**) and not nominated novels by not nominated authors (**NotNom**), based on textual features? (see Section 5.1)
2. Q2: Can the classifications made for these three classes (**NomNov**, **NomAut** and **NotNom**) be related to author gender? (see Section 5.1)
3. Q3: Can logistic regression classify novels into nominated novels (**NomNov**) and not nominated novels (**NomAut** and **NotNom**), based on textual features?(see Section 5.2)
4. Q4: Can the classifications made for these two classes be related to author gender? (see Section 5.2)
5. Q5: Can logistic regression classify author gender, based on textual features? (see Section 5.3)
6. Q6: Can the confidence of the nominated-or-not classification be related to the confidence of the author gender classification? (see Section 5.3)

After each question is introduced, a short description of the experiments used to answer these question will be given. Then, the results will be discussed and analysed.

As can be seen, the results will be discussed in a different order than they were introduced in Chapter 4. In that chapter, the sub-questions were introduced per research question. In this chapter, the sub-questions are discussed per model used. Thus, the results discussed alter between the sub-questions related to RQ1 and RQ2. This is done so that the results of the **NomNov**, **NomAut** and **NotNom** models (Q1) are grouped together with the discussion the analysis of these results split by author gender (Q2). For the nominated-or-not model (Q3 and Q4), the same order is chosen. Thus, first all the results concerning the **NomNov**, **NomAut** and **NotNom** model are discussed (Q1 and Q2, see Section 5.1). Then, the results of the nominated-or-not and **NomNov** and **NotNom** models are discussed (Q3 and Q4, see Section 5.2). Lastly, the results of the author gender classification (Q5) and the relation between the author gender classification and the nominated-or-not models (Q6) are discussed (see Section 5.3). In Section 5.4 an overall conclusion of the logistic regression analysis will be given.The results answering RQ3 will be discussed in Section 5.5 and Section 5.6.

5.1 Classification: NomNov, NomAut, NotNom

Q1: Based on textual features, can logistic regression classify NomNov, NomAut and NotNom? In order to answer this question, I will analyse the precision, recall and F1 scores of these three classes of a 5-fold cross validation logistic regression model. I will compare these results, and the overall accuracy score, to the scores of a model trained on the same dataset, but with randomly assigned classes.

Lastly, I will try to identify patterns in the misclassifications, which could give an insight in the types of authors and writing styles that are misclassified by the model. I will analyse the misclassified novels of authors which have multiple misclassified works in the dataset, as well as the models confidence of the predictions.

Classification: NomNov, NomAut, NotNom In Table 5.1, the results of the classification of the nominated novels (**NomNov**), the not nominated books written by nominated authors (**NomAut**) and the not nominated novels written by not nominated authors (**NotNom)** are shown. As can be seen, the overall accuracy score is 58.7%. As comparison, a model was trained to predict the three classes, **NomNov**, **NomAut**, **NotNom**, but the labels of those three classes were randomly assigned, with similar distribution, to the novels of the entire dataset. Training a model on randomly assigned classes led to an accuracy of 30.6%. This overall accuracy was used as a baseline to check whether the results are better than a random classification, and thus to check whether the model actually makes generalisations over the features. The standard deviation on the F1 scores and overall accuracy can be found in Table 5.1.

The results shown in Table 5.1 show that the model clearly surpasses the classification scores of a model trained on randomly assigned classes, and thus can be concluded that the model has detected patterns in the textual features of the novels, which distinguish **NomNov**, **NomAut** and **NotNom** novels. However, the scores differ greatly between the three different classes. Particularly, the not nominated novels by nominated authors (**NomAut**) have a remarkably lower recall and F1-score. This suggests that the model is less able to detect not nominated novels by nominated authors, than the other two classes.

For both the complete dataset and the balanced author gender subset can be seen that the **NomNov** class have a low precision, but a high recall. This shows that the

COMPLETE CORPUS	Precision	Recall	F1-score	Standard deviation	Number of novels
NomNov	0.569	0.700	0.628	0.0134	100
NomAut	0.567	0.333	0.420	0.0285	102
NotNom	0.615	0.735	0.735	0.0284	98
Accuracy			0.587	0.0155	300

BALANCED AUTHOR GENDER SUBSET	Precision	Recall	F1-score	Standard deviation	Number of novels
NomNov	0.500	0.681	0.576	0.0254	72
NomAut	0.562	0.321	0.409	0.0421	84
NotNom	0.635	0.709	0.670	0.0201	86
Accuracy			0.566	0.0145	242

TABLE 5.1: Results Classification **NomNov**, **NomAut** and **NotNom** classification. The lowest scores per column are underlined. **NomAut** has the lowest precision, recall and F1 score. This suggests that the model is less able to detect not nominated novels by nominated authors, than the other two classes. The same pattern can be seen in the results of the model trained on a balanced author gender subset, thus these results do not seem to be strongly influenced by author gender imbalance. Overall, all three categories surpass the scores of the model trained on randomly assigned classes. Thus can be concluded that the model has detected patterns in the textual features of the novels, which distinguish nominated novels (**NomNov**), not nominated novels by nominated writers (**NomAut**) and not nominated novels by not nominated writers (**NotNom**).

models succeed to correctly classify the majority of the nominated novels, but also classify a high number of not nominated novels as nominated. For example, 39 **NomAut** novels were classified as **NomNov** in the complete dataset. In comparison, only 34 **NomAut** novels were correctly classified, and 29 **NomAut** novels were classified as **NotNom**. Thus, the model seems biased to classify **NomAut** novels as **NomNov** novels. A similar pattern is also seen in the balanced author gender subset, so these misclassifications do not seem to be strongly related to the author gender imbalance in the complete dataset.

To further examine these results, the misclassifications have been studied and the results of classification on the three classes have been split and analysed per author gender.

Q1: Conclusion To conclude, it is possible to classify nominated novels (**NomNov**), not nominated novels by nominated authors (**NomAut**) and not nominated novels by not nominated authors (**NotNom**) using a 5-fold cross-validation logistic regression model trained on textual features. The results greatly surpass the results obtained with a similar model trained on the same dataset with randomly assigned classes. The model has the lowest precision, recall and F1-score on not nominated novels by nominated authors. This suggests that based on textual features, it is hardest to precisely distinguish **NomAut** novels from **NomNov** and **NotNom** novels. As

the majority of the **NomAut** novels were classified as **NomNov**, the misclassifications of **NomAut** novels suggest that the word use of nominated authors in nominated novels are close related to the word use in not nominated novels.

Q2: Can the classifications made for these three classes (NomNov, NomAut and NotNom) be related to author gender? To answer this question, I will split the results shown in Table 5.1 by author gender (see Section 4.3). The goal is to identify author gender specific patterns in the results of the classification of **NomNov**, **NomAut** and **NotNom**.

I will also compare these patterns to the results of a similar model trained on an author gender balanced subset of the complete dataset, to identify the influence of the imbalanced number of novels per gender on the performance of the model. The balanced author gender subset set consists of an equal number of the books by man and woman writers (see Section 4.2 for details, and Appendix A for the corpus).

In the Appendix (see Appendix B), the results can be found of the classification on **NomNov**, **NomAut** and **NotNom** novels using a vectorizer of word features and the gender variables (man and woman). The precision, recall, F1-score and overall accuracy are overall slightly lower than with the vectorizer based on word features only. The difference is small, thus the inclusion of gender variables does not seem to have a strong effect on the performance of the classification task. Similar results were observed when the gender variables were included in the classification on the balanced author gender subset.

In Table 5.2 can be seen that when only looking at the subset of novels by women, the precision is also lowest in **NomNov**, and the recall and F1 scores are lowest in **NomAut**. This pattern is seen in the balanced author gender subset as well, but the difference between the scores is smaller. This is interesting, because the exact same novels written by women were included in the complete dataset and the balanced author gender subset. Thus, balancing author gender seems to slightly diminish the difference in performance for **NomNov** and **NomAut**.

For the men, **NotNom** have the lowest precision on the complete dataset, but this seems to be due to the author gender imbalance of the dataset, as for the balanced author gender subset, **NomNov** written by men writers have the lowest precision. For both sets, **NomAut** has the lowest recall and F1 score. **NotNom** has the highest precision, recall and F1 score for novels of both genders.

Q2: Conclusion To conclude, analysing the results of the classification task by splitting the results on author gender, shows that the best performing category for works written by women and men is **NotNom**. However, for **NomNov** and **NotNom** novels the precision, recall and F1 scores of novels written by women, are lower than for the novels written by men. Thus it seems that for these two classes, it is harder to classify novels written by women than novels written by men. This is not the case for **NomAut**, due to the low recall of **NomAut** novels written by men.

5.2 Classification: nominated novels or not nominated novels

In this section, question 3 and question 4 will be answered. These questions are focused on the same datasets as in Section 5.1, but now categorised in two categories: nominated and not nominated. **NomNov** remains the subset of nominated novels, but the categories **NomAut** and **NotNom** are joined to form the new category of

COMPLETE CORPUS

Women	Precision	Recall	F1-score	Number of novels
NomNov	0.500	0.583	0.538	36
NomAut	0.517	0.357	0.423	42
NotNom	0.680	0.791	0.731	43
Accuracy			0.579	121
Men	Precision	Recall	F1-score	Number of novels
NomNov	0.605	0.766	0.676	64
NomAut	0.613	0.317	0.418	60
NotNom	0.567	0.691	0.623	55
Accuracy			0.592	179

BALANCED
AUTHOR GENDER SUBSET

Women	Precision	Recall	F1-score	Number of novels
NomNov	0.455	0.556	0.500	36
NomAut	0.583	0.333	0.424	42
NotNom	0.604	0.744	0.667	43
Accuracy			0.545	121
Men	Precision	Recall	F1-score	Number of novels
NomNov	0.537	0.806	0.644	36
NomAut	0.542	0.310	0.394	42
NotNom	0.674	0.674	0.674	43
Accuracy			0.587	121

TABLE 5.2: Results classification complete dataset and balanced author gender dataset split by author gender. For the complete dataset, for women, the precision is lowest for **NomNov** and the recall and f1-score for **NomAut**. For men precision is lowest for **NotNom** novels. For the balanced author gender subset, the lowest scores are all for **NomAut**. Thus, **NomAut** is the most difficult class to classify, regardless of author gender.

not nominated novels. Consequently, the complete dataset becomes a unbalanced dataset regarding these categories, as it contains 100 nominated novels, and 200 not nominated novels. The standard deviation and variance of the F1-score per class and overall accuracy can be found in Table 5.3

Q3: Based on textual features, can logistic regression classify nominated novels and not nominated novels? As can be seen in Table 5.3 below, the precision, recall, F1-score and overall accuracy scores are higher for the model trained on two classes nominated or not, than on the one trained on **NomNov**, **NomAut** and **NotNom**. The overall accuracy is 0.713, which is higher than the model trained to classify on three classes. A logistic regression model trained prediction on randomly labelled binary classes with a similar distribution has an accuracy of 0.658.

COMPLETE CORPUS	Precision	Recall	F1-score	Standard deviation	Number of novels
Nominated Novels	0.561	0.640	0.598	0.019	100
Not Nominated Novels	0.806	0.750	0.777	0.0158	200
Accuracy			0.713	0.0143	300
BALANCED AUTHOR GENDER SUBSET	Precision	Recall	F1-score	Standard deviation	Number of novels
Nominated Novels	0.545	0.583	0.564	0.0355	72
Not Nominated Novels	0.818	0.794	0.806	0.0158	170
Accuracy			0.731	0.0214	246

TABLE 5.3: Results Logistic Regression Literature nominated-or-not, classification performed on two datasets: the complete dataset and the balanced author gender subset. The nominated novels perform the least, in both dataset, but the difference in performance is the biggest in the complete dataset.

All precision, recall and F1-scores are lower for the nominated novels than for the not nominated novels. This is expected, due to the distribution of the two categories. Despite the use of balanced class weight, this cannot compensate for the imbalance between nominated and not nominated novels.

Classification: NomNov and NotNom Since the results in Table 5.2 seem to indicate that **NomAut** is the hardest to classify, I have trained two models on **NomNov** and **NotNom** only. The goal is to compare the results of the nominated-or-not model with the results shown in Table 5.4, to examine to what extent the patterns seen in the nominated-or-not models are influenced by the **NomAut** novels.

As can be seen in Table 5.4, the overall results are higher than for the nominated-or-not models. Interestingly, the **NotNom** novels have the lowest performance in the model trained on the complete dataset and **NomNov** on the model trained on the balanced author gender subset. This is different from the results in the nominated-or-not models, in which the nominated novels perform the least for both datasets. This underlines the conclusion that the results of Table 5.3 are influenced by the uneven distribution of nominated and not nominated novels. Another interesting difference is that the F1 score for the complete dataset only differs 0.016 between the two classes, and for the balanced author gender subset 0.036. This is smaller than the difference in F1-scores in the nominated-or-not models. Thus, it seems that only including **NomNov** and **NotNom** novels leads to a smaller difference in classification performance.

Another remarkable results if that all the precision, recall and F1-scores for **NomNov** and **NotNom** are higher in Table 5.4 are higher than the results of the **NomNov**, **NomAut** and **NotNom** classification (see Table 5.1). Thus, identifying **NomNov** and **NotNom** is easier when **NomAut** is not included.

Q3: Conclusion To conclude, the nominated-or-not model is able to classify whether a novel has been nominated or not. A logistic regression model trained prediction on randomly labelled binary classes with a similar distribution has an accuracy of

COMPLETE DATASET	Precision	Recall	F1-score	Standard deviation	Number of novels
NomNov	0.796	0.820	0.808	0.013	100
NotNom	0.811	0.786	0.798	0.0136	98
Accuracy			0.803	0.0129	198
BALANCED AUTHOR GENDER SUBSET	Precision	Recall	F1-score	Standard deviation	Number of novels
NomNov	0.753	0.764	0.759	0.0198	76
NotNom	0.800	0.791	0.795	0.0187	86
Accuracy			0.778	0.0188	158

TABLE 5.4: Results Logistic Regression Literature nominated or not, only on **NomNov** and **NotNom**. In the complete dataset, the nominated had the lowest recall, but the not nominated novels the lowest recall and F1-score. For the balanced author gender subset, the nominated novels score the lowest for all metrics.

0.658. The overall accuracy is 0.731 on the complete dataset, which clearly shows that using textual features an accurate model can be trained to predict which novels have been nominated. As expected, nominated novels have the lowest scores, as the dataset contains twice as many not nominated novels than nominated novels.

For the **NomNov** and **NotNom** classification, such an influence is not seen as the number of novels in the classes are evenly distributed. It is noteworthy that the results of the two classes are close to each other, and higher than the results for these two classes discussed in Section 5.1. Thus, the inclusion of **NomAut** seems to make the it more difficult to correctly classify **NomNov** and **NotNom**.

Q4: Can the classifications made for these two classes be related to author gender? As stated above, in both the complete dataset as the balanced author gender subset, the nominated novels have the lowest precision, recall and F1-score. The accuracy of both models is similar, thus the author gender imbalance does not seem to have influenced the overall accuracy.

In the Appendix (see B), the results can be found of the nominated-or-not model using a vectorizer that includes the gender variables (man and woman). The precision, recall, F1-score and overall accuracy are slightly lower for the not nominated novels on the balanced author gender subset than with the vectorizer based on word features only. For the complete dataset and the not nominated novels of the balanced author gender subset the results are higher. The difference is small, thus the inclusion of gender variables does not seem to have a strong effect on the performance of the classification task.

When looking at the novels written by men and the novels written by women separately, it is clear that in both cases, the classification is most successful for not nominated novels (see Table 5.5). The high performance of not nominated novels written by men is interesting, as in a model classifying **NomNov**, **NomAut** and **NotNom**, nominated novels written by men (**NomNov**) perform best. The influence the number of novels per class could also explain this pattern, as the majority of the novels in the complete dataset are not nominated novels written by men.

COMPLETE DATASET				
Women	Precision	Recall	F1-score	Number of novels
Nominated novels	0.528	0.528	0.528	36
Not nominated novels	0.800	0.800	0.800	85
Accuracy			0.719	121
Men	Precision	Recall	F1-score	Number of novels
Nominated novels	0.577	0.703	0.634	64
Not nominated novels	0.812	0.713	0.759	115
Accuracy			0.709	179

BALANCED AUTHOR GENDER SUBSET				
Women	Precision	Recall	F1-score	Number of novels
Nominated novels	0.486	0.472	0.479	36
Not nominated novels	0.779	0.788	0.784	85
Accuracy			0.694	121
Men	Precision	Recall	F1-score	Number of novels
Nominated novels	0.595	0.694	0.641	36
Not nominated novels	0.861	0.800	0.829	85
Accuracy			0.769	121

TABLE 5.5: Results of two logistic regression models, trained to classify nominated and not nominated novels. One model is trained on the complete dataset and one on the balanced author gender subset. The results are split per author gender. For both datasets, the nominated novels are the least performing for both works written by women and works by men. The precision, recall and F1-score are lowest for nominated novels written by women are the lowest in both datasets, suggesting that this difference in the complete dataset is not completely due to author gender imbalance.

As **NotNom** is the best performing classification for work written by women, it is less surprising that in a model trained on two classes, the not nominated novels by women reach the highest accuracy.

In the balanced author gender subset, not nominated novels are also the best performing class. For the novels written by men, the difference in performance between the two classes has diminished, in comparison to the results on the complete dataset. For the works by women, the performance of nominated novels has decreased, and the performance of the not nominated novels has increased, compared to the results of the complete dataset. This pattern is also seen in Table 5.5. As less novels written by men are included in the balanced author gender subset, the models on **NomNov**, **NomAut** and **NotNom** and nominated-or-not seem to have more difficulty to distinguish the nominated and not nominated novels written by women.

Classification: NomNov and NotNom by author gender When analysing the models trained on **NomNov** and **NotNom** using author gender, it is surprising that for the novels written by women, the **NotNom** novels have the lowest precision, and

the **NomNov** novels the lowest recall and F1-score, see Table 5.6. This is different from the results in Tables 5.2 and 5.5, as **NomNov** has consistently the lowest scores there. For the novels written by men, the **NotNom** novels have the lowest recall and F1-score on the complete dataset and the **NomNov** novels the lowest precision and F1-score on the balanced author gender set.

Overall, the **NomNov** novels F1-scores are the lowest across genders, except for the novels written by men in the model trained on the complete dataset. These results oppose the results of the nominated-or-not models, in which the nominated novels perform the worst across genders. Thus, it seems that the inclusion of **NomAut** influences the relatively low performance of nominated novels.

Q4: Conclusion To conclude, the nominated novels are the least performing for both works written by women and works by men. When comparing the results of between genders, nominated novels written by women, obtain the lowest F1-scores with **NomAut** in the datasets and without.

The results of the **NomNov** or **NotNom** classification task suggest that the inclusion of **NomAut** decreases the scores of the nominated novels, as the difference between the precision, recall and F1-score of novels written by men and novels written by women diminish when **NomAut** is excluded. Also, the nominated novels do not consistently have the lowest scores across all genders and for all datasets (see Table 5.6. Despite the decrease in difference between scores of nominated and not nominated novels in these results, the F1-scores of novels written by women are still consistently lower than the performance of novels written by men. Thus can be concluded that whether **NomAut** is included in the dataset or not, novels written by women are harder to classify than novels written by men, in particular nominated novels.

5.3 Author gender prediction

In this section, the results of a logistic regression model trained to classify whether a book is written by a man or women is discussed. This model uses almost the same datasets as used in Section 5.1 and 5.2. Also, similar methods are used to analyse the results. First, the overall results will be analysed and compared to the results of the balanced author gender subset. Then, these results will be split by **NomNov**, **NomAut** and **NotNom**, to see if author gender can be related to the type of books which are used. Lastly, the confidence of the author gender prediction will be mapped to the confidence of the two class nominated-or-not model.

Q5: Based on textual features, can logistic regression classify author gender? In Table 5.7 can be seen that the precision, recall and F1-score of the model trained on the complete dataset are the consistently lowest for novels written by women. The difference in results is rather big, particularly for recall. The overall accuracy is 0.740, which clearly surpasses the chance of 0.5. The overall accuracy also surpasses the majority prediction of 0.609, namely the accuracy score that would be obtained if all novels were predicted to be written by men.

The results of the complete dataset show that it is harder to classify novels written by women, than novels written by men. This pattern is not as clearly seen in the balanced author gender subset, as the lowest precision is for novels written by men. Additionally, the difference between the scores of novels written by men and by women is lower than on the complete dataset. Thus, author gender imbalance

seems to be related to a lower precision on novels written by women. Additionally, the recall of novels written by men remains high in the balanced author gender subset. Therefore, the F1-score of novels written by men is still higher than the score of novels written by women, despite a low precision. These results therefore suggest that the despite a balance in author gender, the model seems biased to classify novels as 'written by men'.

Q5: Conclusion To conclude, logistic regression can classify author gender, based on textual features, as the accuracy score clearly surpass chance. However, it is important to note that the model is biased towards novels written by men, leading to a higher overall accuracy score on the complete dataset than in the balanced author gender subset.

Q6: If this is possible, can the classification on author gender be related to whether a novel has been nominated or not? To answer this question, the results of the classification on author gender are related to nominated novels in two different manners. Firstly, I will split the results of the author gender prediction into **NomNov**, **NomAut** and **NotNom**, and compare these to the results of the balanced author gender subset. Secondly, I will plot the confidence of the author gender predictions against the confidence of the nominated-or-not predictions.

Author gender prediction: NomNov, NomAut, NotNom In Table 5.8 can be seen that for the complete dataset, for **NomNov**, **NomAut** and **NotNom** the novels written by women have the lowest, precision, recall and F1-score. The biggest difference with the novels written by men is seen in **NomNov**. This is not unexpected, novels written by women also have the lowest scores on the classifications on **NomNov**, see Table 5.1. The overall accuracy is highest for **NomNov**, and the lowest for **NomAut**, but the accuracy scores are close to each other, as they range from 0.735-0.745.

In the results for the balanced author gender subset, this pattern is not seen. For **NomNov**, **NomAut** and NotNom works by men have the lowest precision and novels by women the lowest recall and F1-score. The accuracy is lowest for **NomNov**, and highest for **NotNom**. Thus, the author gender prediction and the relation with **NomNov**, **NomAut** and **NotNom** seems to be influenced by the author gender balance in the dataset.

Confidence of classification In Figure 5.1, the confidence of the classification on author gender is plotted against the confidence of nominated-or-not classification. The confidence of these two classification is shown per author gender, and per dataset (complete dataset and balanced author gender subset). As can be seen, the confidence for novels written by men ranges from 0.0-1.0, and as expected, most novels have a confidence higher than 0.5 to be written by a man. This pattern is seen in the balanced author gender subset as well. For the novels written by women writers, there is not such a clear skew towards the left side of the x-axis. This is interesting, because it would be expected that novels written by women would have a high probability to be written by a man.

In the graphs of the novels written by men, few novels are predicted to be written by a woman and nominated for a literary prize. For the complete dataset, there are 4 outliers, which are all not nominated novels, with a probability higher than 0.6 to be nominated for a literary prize, and a high confidence to be written by a woman (probability to be written by a man is 0.2 or lower). In the balanced author gender

dataset, six of such novels can be identified, of which three nominated novels and three not nominated novels. Also, the nominated novels all have a high probability to be written by men, only 10 nominated novels have a probability lower than 0.5 in both the complete dataset and the balanced author gender subset. Thus, it seems that nominated novels by men authors have a high probability to be written by a men according to the nominated-or-not model.

For the graphs written by women writers, it is noticeable tat none of the novels have a probability to be nominated that approaches 1 in the complete dataset. The highest probability is around 0.9. Also, a limited number of novels has a probability higher than 0.6. This clearly shows that in general, novels written by women overall have a lower probability to be nominated according to the nominated-or-not model. This is more balanced in the balanced author gender subset, with novels approaching a probability score of 1 for nomination. Thus, this pattern could be partially due to the author gender imbalance in the complete dataset. However, the graph of the balanced author gender subset also shows that about half of the nominated novels has a probability lower than 0.5 to be nominated. For the novels written by men authors, only a few nominated novels have a probability score lower than 0.5. Thus, in a balanced author gender subset, there is still a clear lower probability score for nominated novels written by women. This difference indicates that in general, the probability score on nomination of novels written by women is lower than the probability score of novels written by men, according to the nominated-or-not model.

Thus, Figure 5.1 shows that for the nominated-or-not model, there seems to be a relation between a high probability to be nominated for a literary prize, and novels written by men. This is shown by the few books written by men authors that have a low probability to be written by men and a high probability to be written by a woman. For the novels written by women, relatively more novels have a high probability to be written by a men and have won a literary prize. Also, the probability to be nominated for a literary prize is in general lower for novels written by women, than for novels written by men.

Q6: Conclusion To conclude, a relation can be found on whether a novel has been nominated or not, as Figure 5.1 shows that there are few books which have a high confidence to be nominated and a high confidence to be written by a woman. This is in line with the results discussed in Sections 5.1 and 5.2. Additionally, the difference in F1-scores for **NomNov** written by women is lowest, when classifying on author gender (see Table: 5.8, suggesting that nominated novels written by women are the hardest to distinguish from novels written by men.

5.4 RQ1 & RQ2: Conclusion

In conclusion, the results show that based on textual features, logistic regression can be used to classify on **NomNov**, **NomAut** and **NotNom** successfully. Also, it shows that logistic regression can also be used to classify whether a book has been nominated or not. The **NomAut** novels seem to be the hardest to classify across all the models. For the logistic regression on three classes, the scores were lowest across author gender and datasets for **NomAut**. The comparison of the nominated-or-not models and the models trained on **NomNov** and **NotNom** show that the differences in performance decreased can be related to the exclusion of **NomAut**.

The results also suggest a relation between not nominated novels by not nominated authors, and women writers. Figure 5.1 shows that few books have a high

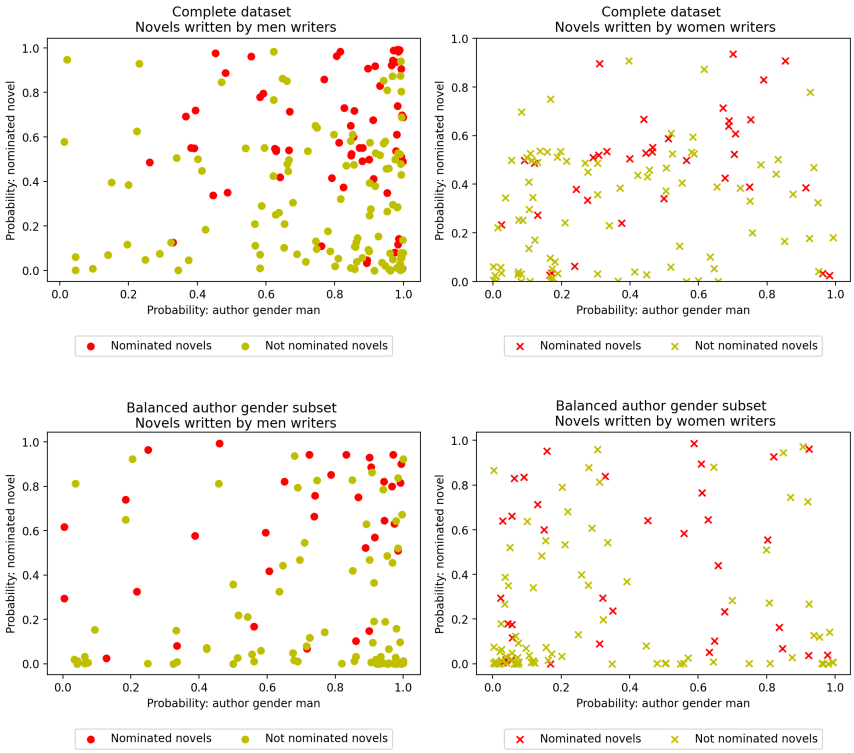

FIGURE 5.1: Probability of nominated-or-not and author gender logistic regression models that a novel is written by a man, and that a novels has been nominated.

confidence to be written by a women and a high confidence to be nominated. For the three class logistic regression models, the nominated-or-not models and the models trained on **NomNov** and **NotNom**, the not nominated novels written by women have a higher F1-score than the nominated novels written by women. Thus, textual features in not nominated novels written by women are more accurately identified than nominated novels by women.

For novels written by men, such a pattern is not seen across models. The F1-scores of the nominated novels by are always higher than the F1-scores of the nominated novels by women, for all classes.

5.5 Q3: LDA Topic Model

For the LDA Topic model, 50 topics were created based on all novels in the corpus, to answer RQ3: Are the differences in topics between books that are nominated for literary prizes and those that do not related to author gender? The LDA results in 36 interpretable topics, of which the ten words with the highest topic weights can be found in the appendix (see Appendix C). An example of an interpretable topic is war, including words such as: *majoor* (major), *soldaat* (soldier) and *oorlog* (war). An example of an uninterpretable topic is topic 32, which consists of the words: *oom* (uncle), *ieder* (each), *twee* (two) and Alkmaar. Some topics, such topics 19 and 42 are clearly related to certain books in the corpus. Topic 19 is about *Congo* by David van Reybrouck, and topic 42 is about *De papegaai, de stier, en de klimmende bougainvillea* by Anil Ramdas.

The topics have been analysed by type of novel, and can be seen in Figures 5.2 and 5.3. In Figure 5.2, the topics per **NomNov**, **NomAut** and **NotNom** novels are shown. The heatmap shows the occurrence of topics in these three classes, relative to each other. Some topics clearly are more strongly related to certain classes, such as topic 0, war. This topic occurs most in **NomNov** novels. Topics 22 to 25 are more strongly related to **NotNom** novels, including topics on the Second World War and international politics. Topics 35-38 are more strongly related to **NomAut** novels, including a topic about love.

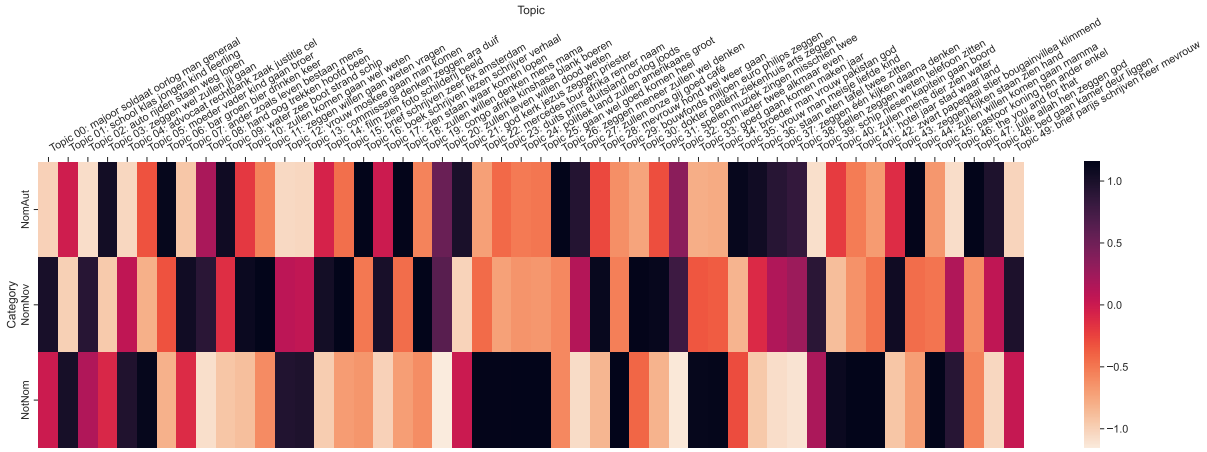

FIGURE 5.2: Heatmap showing relative occurrence of topics in **NomNov**, **NomAut** and **NotNom** novels

When the topics are split by author gender, the division between the **NomNov**, **NomAut** and **NotNom** novels becomes less clear. Certain topics remain related to one of the three classes. I will only discuss topics which are clearly interpretable. Topics that are strongly related to **NomNov** novels, for both women and men writers are: topic 14, art, topic 16, on writing and topic 17, going home and. Topic 23 on the Second World war occurs relatively more in **NotNom** novels, by men and women writers. For the **NomAut** novels, such topics cannot be defined.

Some topics appear to be more related to authors of a certain gender, such as the topic 0 on war, which most strongly is related to **NomNov** novels written by men, and to **NotNom** novels written by men. Some topics also seem more related to gender than to whether or not a has been nominated. For example, topics on religion (Christian and Islamic), appear more in **NomNov** and **NomAut** novels written by men. This can be explained by certain nominated authors that write about these topics, such as Kader Abdolah. Other gendered topics cannot be related to particular authors, such as topic 38, the office. This topic appears more in novels written by men authors, for all **NomNov**, **NomAut** and **NotNom** novels. Lastly, there are also topics that are related to authors of a certain gender in particular classes. For example, topic 30, on hospitals, occurs most in **NomNov** novels written by men and **NomAut** and **NotNom** novels by women. This could suggest that some topics are perceived to be of literary quality when a novel is written by a man, but not when it is written by a woman. This difference in judgement on literary quality related to author gender on certain topics is also shown by Koolen (2018).

Conclusion: LDA To conclude, Figure 5.2 show that there are certain topics that relate to **NomNov**, **NomAut** and **NotNom** novels specifically. Thus, there seem to be differences between the topics in nominated novels and topics in not nominated

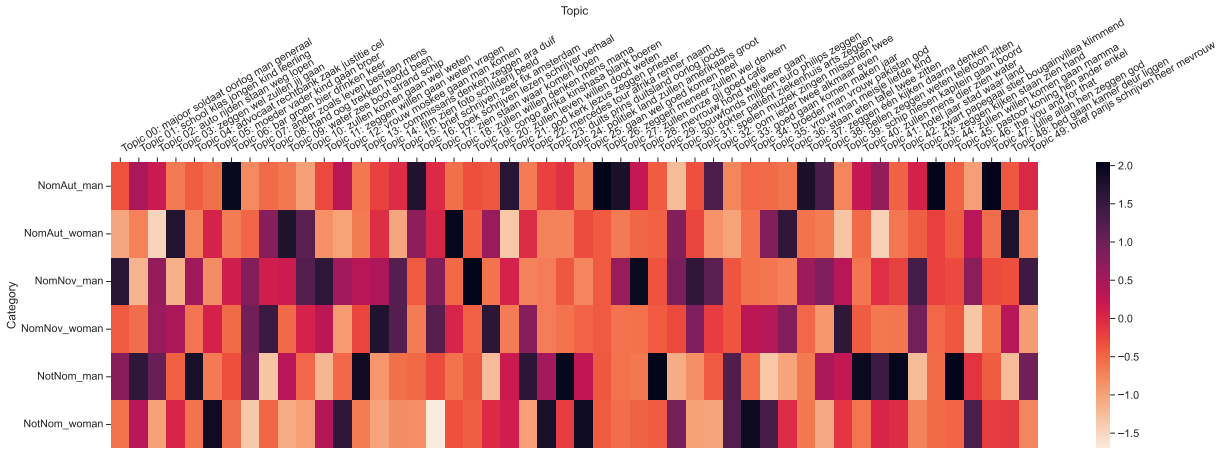

FIGURE 5.3: Heatmap showing relative occurence of topics in **NomNov**, **NomAut** and **NotNom** novels, by author gender

novels. These difference in topics cannot be relate to author gender, as Figure 5.3 shows no clear pattern in relation to author gender. Figure 5.3 does show that the relation of certain topics to **NomNov**, **NomAut** and **NotNom** novels is complex. Some topics are related to a class due to author gender, such as war, which is related to **NomNov** novels, but actually predominantly occurs in novels written by men. Other topics, such as hospitals, seem to be related to nominated novels when a novel is written by a man, and to not nominated novels when a novel is written by a woman.

5.6 RQ3: Cosine Delta

The 1000 most frequent words of the novels are used for a cosine delta analysis. This analysis will show how the writing styles of the novels relate and differ from each other. It will be used to answer the question: Are the differences between writing styles between books that are nominated for literary prizes and those that do not related, to author gender?

The cosine delta analysis is implemented on correctly classified novels, namely novels that have been correctly classified in all three logistic regression models: **NomNov**, **NomAut**, **NotNom**, nominated-or-not and author gender. The correctly classified **NomNov**, **NomAut** and **NotNom** are analysed separately and compared to the misclassified novels in the corresponding classes. The misclassified novels are novels that have been misclassified in all three classification tasks. I will analyse the results of the cosine delta comparison with a dendogram and a heatmap. The dendogram shows how the correctly classified novels relate to each other, and the heatmap shows how the misclassified novels relate to the writing style of the correctly classified novels.

Dendogram In Figure D.1 (see Appendix D), the relation between the correctly classified **NomNov**, **NomAut** and **NotNom** novels are shown. The branches of the dendogram show which novels relate directly to each other in writing style. The further up the dendogram, the less directly writing styles are related. In the **NomNov** and **NomAut** graphs, multiple novels of the same author are shown. For most authors, such as Kristien Hemmerechts (**NomNov**) and Jeroen Brouwers (**NomAut**),

these novels are grouped together directly on the same branch. In the **NomNov** graph, the correctly classified novels of Arnon Grunberg and Kristien Hemmerechts are closely related to each other. This is interesting, as these are both authors which have been nominated multiple times for the *Libris Literatuur Prijs* and the Boekenbon Literatuur Prijs (see: Chapter 1). However, in each of the graphs, not all novels of the same author are grouped together. For the **NomNov** novels, the novels of Anna Enquist are not placed on the same branche. For the **NomAut** novels, the novels of Mensje van Keulen are not grouped together and in **NotNom** graph, the novels of Kluun are not directly related.

Due to the limited number of novels written by women in the correctly classified novels, it is not possible to draw strong conclusions on author gender from Figure D.1 (see Appendix D). For the limited number of novels written by women writers, it seems that they are evenly distributed across the branches, creating an even distribution of author gender in the dendograms.

Comparison Misclassifications Figure 5.4 shows how the writing style of the misclassified novels relate to the correctly classified novels. For the **NomNov** and **NomAut** graphs, a pattern is seen where a specific author relates more to one author, and less to another. For the **NotNom** novels, this pattern is less strong. The rows, representing the correctly classified novels, are either more blue or more red, which shows that the writing style in that novel either is positively related to all misclassified novels, or negatively related to all misclassified novels. This could indicate that for the **NomAut** correctly classified novels, a less distinctive writing style is identified for each author. This could also indicate that the writing styles in the misclassified novels are closely related to each other, and that the correctly classified **NotNom** novels either relate to them or not.

The writing styles of the nominated authors seem to be more distinctive in relation to each other. Each author relates differently to the misclassified novels. For example, in the **NomNov** graph, the misclassified novel of Arnon Grunberg is close in writing style to the correctly classified novels of Arnon Grunberg as well as of Kristien Hemmerechts. This is similar to the results of Figure D.1.

Still, there are a few nominated authors that are positively related to all misclassified novels. For the correctly classified **NomNov** novels, these are novels of Christiaan Weijts, A.F.Th. van der Heijden, Herman Koch and Harry Mulisch. For the **NomAut** novels, these are novels of Jeroen Brouwers, Joost Zwagerman, Vonne van der Meer and Harry Mulisch. It is remarkable that Harry Mulisch is positively correlated with all missclassified novels. Harry Mulisch is recognised as one of the main three Dutch authors (*de grote drie*). Due to the big influence of Harry Mulisch on the Dutch literary scene, this could indicate that the writing style of nominated authors is positively related to the writing style of Harry Mulisch, even if these novels are consistenly misclassified by the logistic regression models.

The correctly classified **NomAut** novels of Mensje van Keulen and Kader Abdolah are consistently negatively related to the misclassified **NomAut** novels. For the correctly classified **NomNov** novels, such pattern was not seen.

Conclusion: Cosine Delta In conclusion, the difference in the writing style between **NomNov**, **NomAut** and **NotNom** novels cannot be clearly identified. For the nominated novels (**NomNov** and **NomAut**), the relation between the correctly classified novels and the misclassified novels differ per author. In the not nominated

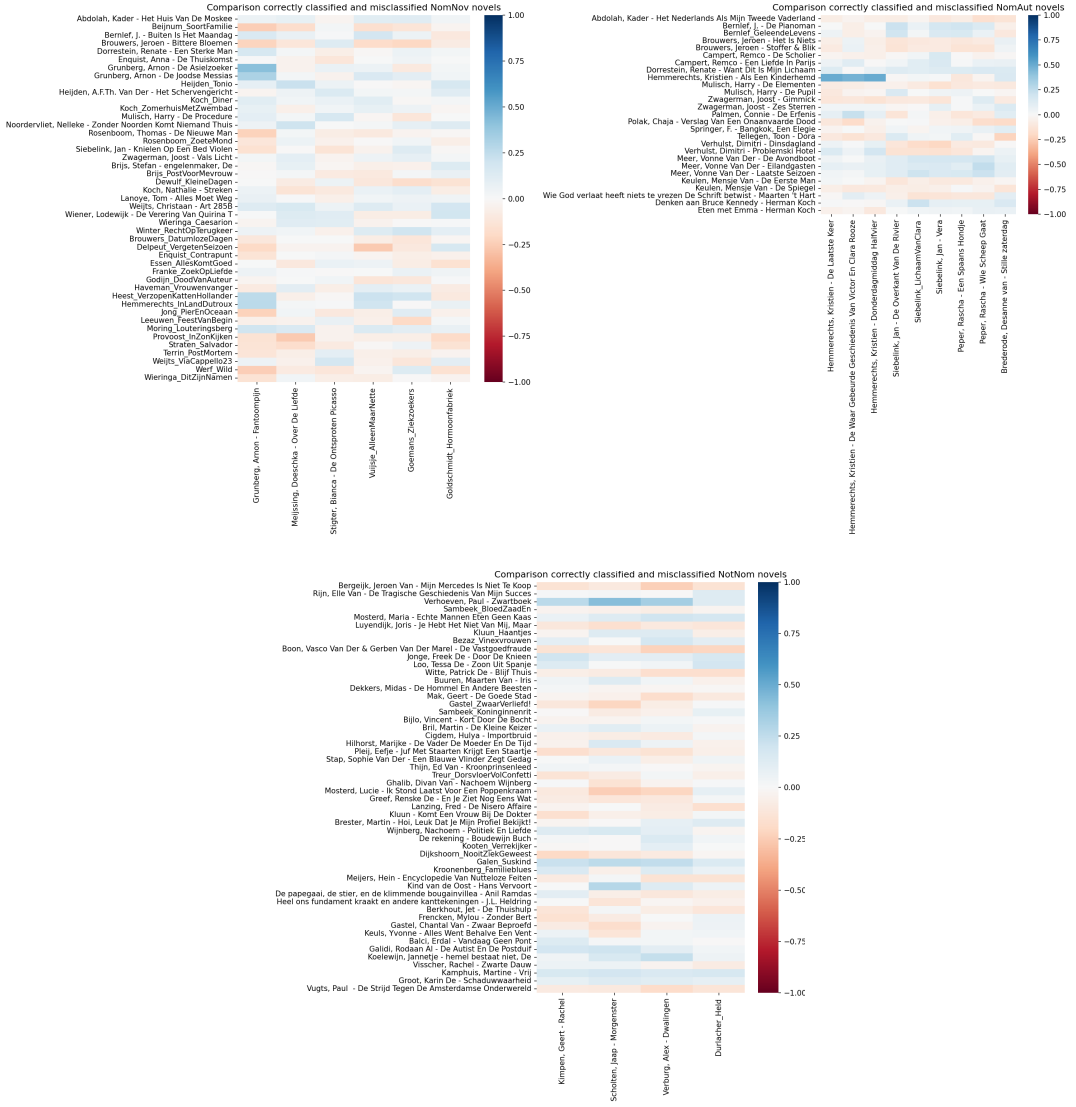

FIGURE 5.4: Heatmap showing how the writing styles of the correctly classified novels relate to the misclassified novels: (a) Nominated novels (**NomNov**) (b) Not nominated novels by nominated writers (**NomAut**) (c) Not nominated novels by not nominated writers (**NotNom**)

novels, a correctly classified novel either positively relates to all misclassified novels or not. No clear relation with author gender could be identified.

The analysis using cosine delta shows that the writing style of Kristien Hemmerechts and Arnon Grunberg, both multiple nominated authors, are positively related to each other. Also, the writing style of Harry Mulisch is positively related to all misclassified **NomNov** and **NomAut** novels, which could show the great influence of the writing style of Harry Mulish on other nominated authors.

COMPLETE DATASET				
Women	Precision	Recall	F1-score	Number of novels
NomNov	0.824	0.778	0.800	36
NotNom	0.822	0.860	0.841	43
Accuracy			0.823	79
Men	Precision	Recall	F1-score	Number of novels
NomNov	0.783	0.844	0.812	64
NotNom	0.800	0.727	0.762	55
Accuracy			0.790	119

BALANCED AUTHOR GENDER SUBSET				
Women	Precision	Recall	F1-score	Number of novels
NomNov	0.743	0.722	0.732	36
NotNom	0.773	0.791	0.782	43
Accuracy			0.759	79
Men	Precision	Recall	F1-score	Number of novels
NomNov	0.763	0.806	0.784	36
NotNom	0.829	0.791	0.810	43
Accuracy			0.797	79

TABLE 5.6: Results of a logistic regression model trained on the complete dataset and a model trained on the balanced author gender subset split by author gender. Unexpectedly, for the novels written by women, the **NotNom** novels have the lowest precision, and the **NomNov** novels the lowest recall and F1-score on the complete dataset. For the balanced author gender subset, the NomNov do score lowest, so the high precision for **NomNov** novels written by women in the complete dataset could be influence by author gender imbalance in the dataset. For the novels written by men, the **NotNom** novels have the lowest recall and F1-score on the complete dataset and the **NomNov** novels the lowest precision and F1-score on the balanced author gender set. Overall, the **NomNov** novels score the lowest across genders, except for the novels written by men in the model trained on the complete dataset.

COMPLETE DATASET	Precision	Recall	F1-score	Standard deviation	Number of novels
Man	0.759	0.827	0.791	0.015	179
Woman	0.705	0.612	0.655	0.0255	121
Accuracy			0.740	0.0176	300
BALANCED AUTHOR GENDER SUBSET	Precision	Recall	F1-score	Standard deviation	Number of novels
Man	0.691	0.777	0.732	0.0183	121
Woman	0.745	0.653	0.696	0.0172	121
Accuracy			0.715	0.0174	242

TABLE 5.7: Results Logistic Regression Author Gender prediction. For the complete dataset, the results of the novels written by women are the lowest. On the balanced author gender set, the precision on novels written by men is the lowest, and the recall and F1 score of novels written by women are the lowest.

COMPLETE DATASET	Precision	Recall	F1-score	Number of novels
NomNov				
Man	0.771	0.844	0.806	64
Woman	0.667	0.556	0.606	36
Accuracy			0.740	100
NomAut				
Man	0.746	0.833	0.787	60
Woman	0.714	0.595	0.649	42
Accuracy			0.735	102
NotNom				
Man	0.759	0.800	0.779	55
Woman	0.725	0.674	0.699	43
Accuracy			0.745	98
BALANCED AUTHOR GENDER SUBSET	Precision	Recall	F1-score	Number of novels
NomNov				
Man	0.628	0.750	0.684	36
Woman	0.690	0.556	0.615	36
Accuracy			0.653	72
NomAut				
Man	0.708	0.810	0.756	42
Woman	0.778	0.667	0.718	42
Accuracy			0.738	84
NotNom				
Man	0.733	0.767	0.750	43
Woman	0.756	0.721	0.738	43
Accuracy			0.744	86

TABLE 5.8: Results Logistic Regression Author Gender prediction split by **NomNov**, **NomAut** and **NotNom**. Novels written by women have the lowest performance over all classes, except for**NotNom** in the balanced author gender subset.

Chapter 6

Conclusion

The goal of this thesis was to answer the research question: Can quantifiable literary qualities be used to investigate author gender inequality in Dutch literary prizes? Before answering the main question, I will answer the three questions formulated to answer the main research question:

1. RQ1: Can nominated and not nominated novels be identified based on textual features only?
2. RQ2: Is there a relation between classifications on nominated and not nominated novels and author gender, where both classifications are based on textual features?
3. RQ3: Are the differences in topics/writing styles between books that are nominated for literary prizes and those that are not, related to author gender?

RQ1: Can nominated and not nominated novels be identified based on textual features only? The results of the classification models clearly show that it is possible to identify nominated and not nominated novels based on textual features only (see Section 5.4). The three models on classification of nomination (nominated novels (**NomNov**), not nominated novels from nominated authors (**NomAut**) and not nominated novels from not nominated authors (**NotNom**), nominated-or-not and **NomNov** or **NotNom**) all obtained an accuracy higher than chance. This means that the model predicts classes based on generalisations made on textual features. The results also show that the not nominated novels by nominated authors (**NomAut**) are the hardest to classify. This could be due to the limited number of unique authors in this category, resulting in the models being trained on a limited number of writing styles, making it harder to generalise the distinguishing features for not nominated novels by nominated authors. Another reason for the low performance of **NomAut** novels, is that this category is the least well defined. To illustrate, the **NomAut** novels also include *Boekenweekgeschenken*, which are shorter novels that are not considered for literary prizes. The reason why one novel of a nominated author is nominated and the other is not, could be related to other aspects than word use. For example, if an author has been nominated many times, a jury could decide to nominate other novels to create some balance.

RQ2: Is there a relation between classifications on nominated and not nominated novels and author gender, where both classifications are based on textual features? The results of the classification tasks can be related to author gender. The results show a relation between the word use in novels written by women and not nominated novels by not nominated authors (**NotNom**). This relation is most clearly shown by the scores of the classification tasks. Novels written by women writers

consistently have the highest score for the **NotNom** class, in comparison to **NomNov** and **NomAut**. This shows that for a classification task on nominated and not nominated novels, it is easiest to classify **NotNom** for novels written by women. For the classification task on author gender, **NotNom** has the highest scores on novels written by women. Thus, there seems to be a relation between novels written by women and **NotNom**.

For the novels written by men, such a relation between the **NomNov**, **NomAut** and **NotNom** classes was not found, but the novels written by men did consistently have higher results than the novels written by women, for all classes. This was probably not due to the higher number of books by men authors in the dataset, as this pattern was also seen, but not as strong, in the subset with an equal author gender balance.

One could argue that the results do not show a strong relation between nominated and not nominated novels and author gender, as adding author gender variables to the textual features does not remarkably change the results of the classification task. One could even argue that this shows that the classification tasks do not show a generalisation based on the textual features, and that the relation found is coincidental. However, the addition of these two variables is very small on the 5000 textual features that are used in the classification task. Therefore, it is not surprising that the author gender variables do not remarkably alter the results of the classification class. Additionally, it is more important that four different classification tasks show a relation between not nominated novels and novels written by women, based on the same textual features. Since different task suggest the same relation, it is plausible that the models made generalisations over the textual features and that this relation shown is not based on coincidence.

RQ3: Are the differences in topics/writing styles between books that are nominated for literary prizes and those that are not, related to author gender? There can be topics identified which are related to books that have been nominated and book that have not. For the writing styles, such difference cannot be defined.

For the topics, a few topics could be identified that occur relatively more in nominated or not nominated novels, which could sometimes be related to author gender. For example, the topic Second World War occurs most in not nominated novels of not nominated authors, and the topic writing in nominated novels. Both topics have a high probability to occur in novels written by men and novels by women. Other topics, such as war, seem to relate a specific nomination class, but are actually more gender specific. The topic war occurs relatively most in **NomNov** novels, in comparison to **NomAut** and **NotNom** novels, but this is due to the high accuracy of this topic in **NomNov** novels written by men. Another interesting result, is that some topics are judged to be of higher literary quality when written by a man author. For example, the topic hospital occurs most in **NomNov** novels written by men and not nominated novels (both **NomAut** and **NotNom**) written by women. This supports the theory that for particular topics and genres, the judgement of literary quality of a certain topics or genres is higher when a novel is written by a man writer (Koolen et al., 2020).

For the difference in writing styles between nominated and not nominated, it was expected that the relation between novels that are consistently correctly classified in the logistic regression models could be related to author gender. Such a pattern could not be identified, nor falsified. Another expectation was that a pattern could be found in the relation between correctly classified novels and misclassified novels. For the novels by nominated authors (**NomNov** and **NotNom**) such a pattern cannot

be seen. The relation between the writing styles seem to highly depend on how close the writing style of particular authors are related to each other. The results do show that highly frequently nominated authors Arnon Grunberg and Kristien Hemmerechts have writing style closely related to each other, and that all nominated authors have a writing style that is close to the writing style of Harry Mulisch. This indicates that there is a particular writing style in Dutch literature, which sets the norm of writing styles that are judged to be of high literary quality.

For the **NotNom** novels, such pattern could not be identified. All the misclassified novels either were closely related to all correctly classified novels, or not. Due to the limited number of misclassified novels, no further general conclusion can be drawn from this observation.

6.1 Main conclusion

The answers to these research questions show that it is possible to investigate author gender inequality in Dutch literary prizes using quantifiable literary qualities. The different experiments performed on the dataset of nominated and not nominated authors show that it is possible to identify nominated and not nominated novels based on textual features only, and also to identify a relation between these classifications and nominated and not nominated novels. The results indicate that the word use of women authors is related to not nominated novels.

The analysis of the topics in nominated and not nominated novels indicate that the relation between nominated and not nominated novels and author gender is rather complex, and highly depends on the topic which is investigated. The relation is complex, since some topics relate differently to nominated or nominated novels, depending on author gender. Some topics are related to nominated or not nominated novels specifically, but for other topics, the judgement of literary quality seems to depend more on the gender of the author. For example, the topic hospital relates most strongly to nominated novels written by men and not nominated novels written by women. Lastly, also topics have been identified that relate more strongly to author gender, than to whether a novel has been nominated or not.

The difference in writing style of nominated and not nominated novels cannot be clearly defined, but the results do indicate that the writing style of Harry Mulisch, who is considered to be one of the three greatest Dutch authors, has a strong similarity with writing styles that are perceived to be of literary quality.

To conclude, this thesis not only shows that it is possible to investigate author gender inequality in Dutch literary prizes with quantifiable literary qualities, but it also indicates that the inequality in Dutch literary prizes is rooted in a homogeneous writing style that is related to the writing style of men. The results clearly show that nominated and not nominated novels are distinguishable both for men and women writers, as the predictions of all the classifications tasks surpass chance. Thus, the results suggest that a particular word use exists that identifies literary quality. However, this word use seems to be further removed from women writers, even from their word use in nominated novels.

These conclusions support the theory that some works are viewed as 'literary' and 'universal', due to the identity of the authors, whereas authors outside this framework are perceived as the other (Koren and Delhaye, 2019). In this case, the women are the 'other', for which it is harder to enter the framework of literary authors. The influential literary authors, such as Harry Mulisch, are mostly men, and determine the 'literary' word use that is needed to be accepted in the frame work.

The conclusions also supports the theory that the homogeneous idea of literary quality is reproduced due to the Dutch education system, in which students are taught on literary quality by predominantly reading works written by white men (Dera, 2020). Lastly, the results support that the critique of Dutch authors, that the homogeneity in the literary environment is due to a notion of literariness which is related to white men (Ramdas, 1997; Amatmoekrim, 2015; Rouw, 2015; Weijers, 2014). It seems harder for women authors to be recognised as literary author, and even if they are recognised, it is harder for them to get nominated again.

Thus, the inequality in Dutch literary prizes cannot be solved by simply having more women in juries (Ahmed, 2012), but is rooted in inequality in the idea of literary quality. This is reproduced by the Dutch school system (Koren and Delhaye, 2019), as well as reviews in news papers (Berkers, 2009). The literary prizes itself uphold this inequality as well, as white men are still more often nominated and thus continue to set the norm to be centred around the word use of white men.

6.2 Limitations

The conclusions of this research are limited by several factors. Firstly, the dataset contains a limited number of authors, especially the not nominated novels by nominated authors. This limited number of different authors make it harder to generalise the results, as it is unclear to what extend the high occurrence of certain authors, such as Renate Dorrestein influence the experiments. Secondly, this thesis only focuses on author gender inequality, in particular between men and women without taking in account other factors influencing language use, such as ethnicity, age and social class (Eckert, 2012). In order to analyse inequality in Dutch literary prizes, this should be researched as well, preferably in an intersectional manner. Also, influences within the literary environment, such as prestige of a publisher or the reviews of novels, could be considered as well.

The interpretation of my results are also limited due to the computational focus of my research. For example, the analysis of the topics is limited due to the uninterpretable topics. It would be worthwhile to combine the techniques I have used with qualitative methods, such as close reading, so that the results that cannot be interpreted can be analysed using other methods and approaches. Also, the results show that the difference in topics and writing styles are rather nuanced. This nuance could be further explored with close reading research.

Lastly, the dataset used is rather limited. A greater number of novels in all three classes, would lead to more general results. It would in particular be worthwhile to include more different authors in the corpus. As authors have a very distinguishable personal writing style Tuzzi and Cortelazzo, 2018; Herrmann, Jacobs, and Piper, 2021, having more authors in the dataset would lead to more different writing styles in the corpus, and therefore more general results. Another improvement of the dataset would be to select not nominated novels which were send in by their publishers, but were not selected for the longlist. The idea behind of the selection of not nominated novels in this corpus, was to select novels that in theory could have been opponents of the nominated novels. In this manner, the actual opponents of the nominated novels could be used.

6.3 Future work

As this thesis mainly focuses on identifying inequality, it would be interesting to further research the textual factors that relate to the author gender inequality that has been identified. Firstly, it would be interesting to research to what extend the writing styles of certain great writers, such as Harry Mulisch, relate to the general writing style of nominated novels.

Another suggestion for further research, is the difference in topics of nominated and not nominated novels, in particular in relation to author gender. One could argue that the relations found in this thesis are coincidental, as the premise of unsupervised learning is that they will find a relation. However, the topics found and the classes and author gender they identify with, seem to relate with the topics in the corpus. To investigate to what extend the topics found by LDA topic modelling are representative of the relations between the topics in the novels, it would be worthwhile to conduct further research combining computational techniques with close reading.

The combination of techniques used in this thesis, could also be applied to research other (potential) forms of inequality in the Dutch literary scene, such as ethnic and cultural background, socio-economic class and queerness. These techniques could also be used outside of the literary scene, for example to research inequality in job applications. Resumes and motivational letters of candidates that were rejected, selected for interviews or hired could be used to investigate if certain word use and writing styles relate to the chance of being hired. Another interesting use of the techniques would be to investigate whether certain writing styles and word use in essays lead to higher grades at the universities, and how this relates to socio-economic background, gender and ethnic and cultural background. Lastly, further research could also investigate to what extend machine learning using textual features can judge literary quality, for example by predicting which novels will win a literary prize in the future. In this way, the possibility of an algorithm picking the winners of literary prizes could be explored.

Appendix A

Corpus

Author	Author gender	Title	Year	Libris Literatuur Prijs	Boekenbon Literatuurprijs	Target	Balanced author gender subset
Thomas van Aalten	Man	De Schuldigen	2012	Longlist		NomNov	Yes
Kader Abdolah	Man	De Boodschapper	2008			NomAut	No
Kader Abdolah	Man	De Koran	2008			NomAut	Yes
Kader Abdolah	Man	De Kraai	2011			NomAut	Yes
Kader Abdolah	Man	Het Nederlands Als Mijn Tweede Vaderland	1996			NomAut	Yes
Kader Abdolah	Man	Het Huis van de Moskee	2006	Longlist		NomNov	Yes
Ayaan Hirsi Ali	Woman	De Zoontjesfabriek	2002			NotNom	Yes
Erdal Balci	Man	Vandaag Geen Pont	2009			NotNom	No
Kees van Beijnum	Man	Een Soort Familie	2010		Shortlist	NomNov	No
Kees van Beijnum	Man	De Oesters van Nam Kee	2000			NomAut	Yes
Abdelkader Benali	Man	De Stem van mijn Moeder	2010	Longlist		NomNov	No
Marinus van den Berg	Man	Nooit te Oud	2007			NotNom	No
Jeroen Bergeijk	Man	Mijn Mercedes is niet te koop	2006			NotNom	Yes
Jet Berkhout	Woman	De Thuishulp	2009			NotNom	Yes
J. Bernlef	Man	Geleendelevens	2010			NomAut	No
J. Bernlef	Man	De Pianoman	2008			NomAut	Yes
J. Bernlef	Man	Buiten is het Maandag	2004	Shortlist	Shortlist	NomNov	No
J. Bernlef	Man	Zijn Dood	2011			NomAut	Yes
Hanna Bervoets	Woman	Lieve Celine	2011			NomAut	Yes
Naima el Bezaz	Woman	Vinex Vrouwen	2010			NotNom	Yes
Vincent Bijlo	Man	Kort door de Bocht	2008			NotNom	Yes
Aliefka Bijlsma	Woman	Mede Namens Mijn Vrouw	2010			NotNom	Yes
Oscar van den Boogaard	Man	Majesteit	2010			NomAut	Yes
Oscar van den Boogaard	Man	Meer dan een Minnaar	2010		Shortlist	NomNov	Yes
Vasco van der Boon	Man	De Vastgoedfraude	2009			NotNom	Yes
Johan de Boose	Man	De Poppenspeler en de Duivelin	2009			NomAut	Yes
Martin Bossenbroek	Man	De Boerenoorlog	2013		Shortlist	NomNov	Yes
Désanne van Brederode	Woman	Stille Zaterdag	2011			NomAut	Yes
Désanne van Brederode	Woman	Door Mijn Schuld	2010	Longlist		NomNov	Yes
Claudia de Breij	Woman	Dingen die fijn zijn	2009			NotNom	Yes
Martin Brester	Man	Hoi, leuk dat je mijn profiel bekijkt!	2009			NotNom	Yes
Stefan Brijs	Man	Post voor mevrouw Bromley	2012	Longlist		NomNov	No
Stefan Brijs	Man	De Engelenmaker	2006			NomNov	Yes
Martin Bril	Man	Overal Wonen Mensen	2011			NotNom	Yes
Martin Bril	Man	Vaarwel Evelien	2011			NotNom	Yes
Martin Bril	Man	De Kleine Keizer	2008			NotNom	Yes
Martin Bril	Man	Evelien 2 Gelukkig Niet	2003			NotNom	Yes
Martin Bril	Man	Plat du Jour	2011			NotNom	Yes
Jan Brokken	Man	De Wil en de Weg	2006			NomAut	Yes
Jan Brokken	Man	Zeedrift	2009			NomAut	Yes
Jeroen Brouwers	Man	Het is Niets	1993			NomAut	Yes
Jeroen Brouwers	Man	Bittere Bloemen	2011	Shortlist	Shortlist	NomNov	Yes
Jeroen Brouwers	Man	Datumloze Dagen	2008	Shortlist		NomNov	No
Jeroen Brouwers	Man	Stoffer & Blik	2004			NomAut	Yes
Herman Brusselmans	Man	[Guggenheimer 01] De terugkeer van Bonanza	1995			NomAut	Yes
Herman Brusselmans	Man	[Guggenheimer 02] Guggenheimer wast witter	1996			NomAut	Yes
Herman Brusselmans	Man	[Guggenheimer 03] Uitgeverij Guggenheimer	1999			NomAut	No
Herman Brusselmans	Man	Trager dan Snelheid	2010			NomAut	Yes
Herman Brusselmans	Man	Het Einde van Mensen in 1967	1999			NomAut	Yes
Miquel Bulnes	Man	Attaque	2007			NomAut	Yes
Maarten van Buuren	Man	Iris	2011			NotNom	No
Boudewijn Büch	Man	De Rekening	1990			NotNom	Yes
Remco Campert	Man	Dagboek van een Poes	2007			NomAut	Yes
Remco Campert	Man	De Scholier	2009			NomAut	Yes
Remco Campert	Man	Een Liefde in Parijs	2004			NomAut	Yes
Hülya Cigdem	Woman	Import Bruid	2008			NotNom	Yes
Eveline Crone	Woman	Het Puberende Brein	2008			NotNom	Yes
Midas Dekkers	Man	De Hommel en Andere Beesten	2005			NotNom	Yes
Peter Delpeut	Man	Het Vergeten Seizoen	2008	Longlist		NomNov	Yes
Bernard Dewulf	Man	Kleine Dagen	2010	Winner		NomNov	Yes
Nico Dijkshoorn	Man	Nooit Ziek Geweest	2012			NotNom	Yes
Adriaan van Dis	Man	Tikkop	2011	Shortlist		NomNov	Yes
Adriaan van Dis	Man	Leeftocht	2007			NomAut	Yes
Adriaan van Dis	Man	Een Barbaar in China	1987			NomAut	Yes
Adriaan van Dis	Man	De Wandelaar	2007			NomAut	No
Renate Dorrestein	Woman	Een Sterke Man	1995	Shortlist		NomNov	Yes
Renate Dorrestein	Woman	De Leesclub	2010			NomAut	Yes
Renate Dorrestein	Woman	De Stiefmoeder	2011			NomAut	Yes
Renate Dorrestein	Woman	Echt Sexy	2007			NomAut	Yes
Renate Dorrestein	Woman	Een Hart van Steen	1998			NomAut	Yes
Renate Dorrestein	Woman	Heden Ik	1993			NomAut	Yes
Renate Dorrestein	Woman	Het Duister Dat Ons Scheidt	2003			NomAut	Yes
Renate Dorrestein	Woman	Het Hemelse Gerecht	1991			NomAut	Yes
Renate Dorrestein	Woman	Is Er Hoop	2013			NomAut	Yes
Renate Dorrestein	Woman	Mijn zoon heeft een sexleven en ik lees mijn moeder Roodkapje voor	2006			NomAut	Yes
Renate Dorrestein	Woman	Zonder Genade	2002		Shortlist	NomNov	Yes
Renate Dorrestein	Woman	Ontaarde Moeders	1992			NomAut	Yes
Renate Dorrestein	Woman	Noorderzon	2009			NomAut	Yes
Renate Dorrestein	Woman	Want dit is mijn lichaam	1997			NomAut	Yes
Renate Dorrestein	Woman	Zolang er leven is	2015			NomAut	Yes
Renate Dorrestein	Woman	Voor liefde druk op f	1999			NomAut	Yes
Dirk Draulans	Man	Beagledagboek	2010			NotNom	Yes
Jessica Durlacher	Woman	Held	2010			NotNom	Yes
G.L. Durlacher	Man	Godvergeten Tijd	2009			NomAut	No

Author	Author gender	Title	Year	Libris Literatuur Prijs	Boekenbon Literatuurprijs	Target	Balanced author gender subset
Anna Enquist	Woman	Contrapunt	2009	Shortlist		NomNov	Yes
Anna Enquist	Woman	Het Geheim	1997			NomAut	Yes
Anna Enquist	Woman	Het Meesterstuk	1994			NomAut	Yes
Anna Enquist	Woman	Mei	2007			NomAut	Yes
Anna Enquist	Woman	De Verdovers	2012	Longlist		NomNov	Yes
Anna Enquist	Woman	De Thuiskomst	2006	Longlist		NomNov	Yes
Rob van Essen	Man	Alles komt goed	2013	Longlist		NomNov	No
Louis Ferron	Man	Karelische Nachten	1990		Winner	NomNov	Yes
Herman Franke	Man	Zoek op Liefde	2009	Longlist		NomNov	Yes
Mylou Frencken	Woman	Zonder Bert	2009			NotNom	Yes
Louise O. Fresco	Woman	De Utopisten	2008	Shortlist		NomNov	Yes
Alex van Galen	Man	Süskind	2012			NotNom	No
Rodaan Galidi	Man	De Autist en de Postduif	2009			NotNom	Yes
Chantal van Gastel	Woman	Zwaar Verliefd!	2008			NotNom	Yes
Chantal van Gastel	Woman	Zwaar Beproefd	2009			NotNom	Yes
Esther Gerritsen	Woman	Superduif	2011	Shortlist		NomNov	Yes
Esther Gerritsen	Woman	Dorst	2013	Shortlist		NomNov	Yes
Wim Gijsen	Man	[Merisse 01] Kring van Stenen	1989			NotNom	Yes
Wim Gijsen	Man	[Merisse 02] Groene Eiland	1990			NotNom	Yes
Wouter Godijn	Man	De dood van een auteur die een beetje op Wouter Godijn lijkt	2008	Longlist		NomNov	Yes
Anne-Gine Goemans	Woman	Glijvlucht	2012	Longlist		NomNov	Yes
Anne-Gine Goemans	Woman	Ziekzoekers	2008	Longlist		NomNov	Yes
Saskia Goldschmidt	Woman	De Hormoonfabriek	2013	Longlist		NomNov	Yes
Renske Greef	Woman	En je ziet nog eens wat	2009			NotNom	Yes
Karin Groot	Woman	Schaduwwaarheid	2011			NotNom	Yes
Arnon Grunberg	Man	De Joodse Messias	2005	Longlist	Shortlist	NomNov	No
Arnon Grunberg	Man	De Asielzoeker	2004			NomNov	No
Arnon Grunberg	Man	Huid en Haar	2011	Shortlist	Shortlist	NomNov	No
Arnon Grunberg	Man	Fantoompijn	2000		Winner	NomNov	Yes
Arnon Grunberg	Man	Onze Oom	2009	Shortlist		NomNov	No
Kees 't Hart	Man	Hotel Vertigo	2013	Longlist		NomNov	Yes
Kees 't Hart	Man	Ter Navolging	2004	Longlist	Shortlist	NomNov	No
Maarten 't Hart	Man	Wie God verlaat heeft niets te vrezen: de Schrift betwist	2011			NomAut	Yes
Mariëtte Haveman	Woman	De Vrouwenvanger	2011	Longlist		NomNov	Yes
Detlev van Heest	Man	De verzopen katten en de Hollander	2011	Longlist		NomNov	No
A.F.Th. van der Heijden	Man	Het Schervengericht	2007	Longlist	Winner	NomNov	No
A.F.Th. van der Heijden	Man	Weerborstels	1992			NomAut	No
A.F.Th. van der Heijden	Man	Tonio	2012	Winner		NomNov	No
Ellen Heijmerikx	Woman	Blinde Wereld	2009			NotNom	Yes
Ellen Heijmerikx	Woman	Wij Dansen Niet	2011			NotNom	Yes
J.L. Heldring	Man	Heel ons fundament kraakt en andere kanttekeningen	2003			NotNom	No
Kristien Hemmerechts	Woman	In het land van Dutroux	2008	Longlist		NomNov	Yes
Kristien Hemmerechts	Woman	Wit Zand	1993			NomAut	Yes
Kristien Hemmerechts	Woman	Een jaar als (g)een ander	2003			NomAut	Yes
Kristien Hemmerechts	Woman	De waar gebeurde geschiedenis van Victor en Clara Rooze	2005			NomAut	Yes
Kristien Hemmerechts	Woman	Donderdagmiddag Halfvier	2002			NomAut	Yes
Kristien Hemmerechts	Woman	Ann	2008			NomAut	Yes
Kristien Hemmerechts	Woman	Als een kinderhemd	2006			NomAut	Yes
Kristien Hemmerechts	Woman	De laatste keer	2004			NomAut	Yes
Joke Hermsen	Woman	De liefde dus	2009	Longlist		NomNov	Yes
Marijke Hilhorst	Woman	De vader, de moeder en de tijd	2008			NotNom	Yes
Oek de Jong	Man	Pier en Oceaan	2013	Shortlist		NomNov	No
Freek de Jonge	Man	Door de knieën	2004			NotNom	Yes
Atte Jongstra	Man	De avonturen van Henry II Fix	2008	Longlist		NomNov	No
Lieve Joris	Woman	Zangeres op Zanzibar en andere reisverhalen	2008			NotNom	Yes
Lieve Joris	Woman	De Golf	2007			NotNom	Yes
Martine Kamphuis	Woman	Vrij	2011			NotNom	Yes
Martine Kamphuis	Woman	Ex	2011			NotNom	Yes
Marie Kessels	Woman	Ruw	2010	Shortlist		NomNov	Yes
Frank Ketelaar	Man	Avond aan avond	2006			NotNom	Yes
Mensje van Keulen	Woman	Liefde heeft geen hersens	2012	Longlist	Shortlist	NomNov	Yes
Mensje van Keulen	Woman	De Spiegel	2008			NomAut	Yes
Mensje van Keulen	Woman	De eerste man	2011			NomAut	Yes
Mensje van Keulen	Woman	Een goed verhaal	2010	Shortlist		NomNov	Yes
Yvonne Keuls	Woman	Alles went behalve een vent	2009			NotNom	Yes
Geert Kimpen	Man	Rachel	2011			NotNom	No
Kluun	Man	Komt een vrouw bij de dokter	2009			NotNom	Yes
Kluun	Man	Haantjes	2010			NotNom	Yes
Nathalie Koch	Woman	Streken	2007	Longlist		NomNov	Yes
Herman Koch	Man	Denken aan Bruce Kennedy	2005			NomAut	Yes
Herman Koch	Man	Eten met Emma	2000			NomAut	Yes
Herman Koch	Man	Odessa Star	2003			NomAut	No
Herman Koch	Man	Het Diner	2010	Longlist		NomNov	Yes
Herman Koch	Man	Zomerhuis met Zwembad	2012	Longlist		NomNov	Yes
Herman Koch	Man	Red ons Maria Montanelli	1989			NomAut	Yes
Jannetje Koelewijn	Woman	De hemel bestaat niet	2011			NotNom	Yes
Kees van Kooten	Man	De Verrekijker	2013			NotNom	Yes
Yvonne Kroonenberg	Woman	Familieblues	2012			NotNom	Yes
Ernest van der Kwast	Man	Mama Tandoori	2010			NotNom	Yes
Tom Lanoye	Man	Sprakeloos	2010	Shortlist	Shortlist	NomNov	No
Fred Lanzing	Man	De Nisero-affaire	2009			NotNom	Yes
Rik Launspach	Man	1953	2009			NotNom	Yes
Stan Lauryssens	Man	Rode Rozen	2004			NotNom	Yes
Joke van Leeuwen	Woman	Alles Nieuw	2009	Longlist	Shortlist	NomNov	Yes
Joke van Leeuwen	Woman	Feest van het begin	2013		Winner	NomNov	Yes
Tomas Lieske	Man	Dünya	2009			NomNov	Yes
Celine Linssen	Woman	Duet	2007			NotNom	Yes
Tessa de Loo	Woman	Zoon uit Spanje	2004			NotNom	Yes
Karel Glastra van Loon	Man	De Onzichtbaren	2013			NomAut	Yes
Karel Glastra van Loon	Man	Lisa's adem	2000			NomAut	No
Karel Glastra van Loon	Man	De passievrucht	1999		Winner	NomNov	No
Karel Glastra van Loon	Man	De romans	2008			NomAut	Yes
Joris Luyendijk	Man	Je hebt het niet van mij, maar	2010			NotNom	Yes
Geert Mak	Man	De goede stad	2007			NotNom	Yes
Geert Mak	Man	Reizen zonder John	2012			NotNom	No
Vonne van der Meer	Woman	De vrouw met de sleutel	2012	Longlist		NomNov	Yes
Vonne van der Meer	Woman	Eilandgasten	1999			NomAut	Yes
Vonne van der Meer	Woman	Laatste seizoen	2002			NomAut	Yes
Vonne van der Meer	Woman	De Avondboot	2001			NomAut	Yes
Vonne van der Meer	Woman	De reis naar het kind	1989			NomAut	Yes
Vonne van der Meer	Woman	Take 7	2007			NomAut	Yes
Hein Meijers	Man	Encyclopedie van nutteloze feiten	2012			NotNom	Yes
Doeschka Meijsing	Woman	Over de liefde	2008	Longlist	Winner	NomNov	Yes
Jan van Mersbergen	Man	Naar de overkant van de nacht	2012	Longlist		NomNov	Yes
Marente de Moor	Woman	De nederlandse maagd	2011			NomNov	Yes
Margriet de Moor	Woman	Op de rug gezien	1989		Shortlist	NomNov	Yes
Margriet de Moor	Woman	De schilder en het meisje	2011	Longlist		NomNov	Yes
Maria Mosterd	Woman	Echte mannen eten geen kaas	2008			NotNom	Yes
Lucie Mosterd	Woman	Ik stond laatst voor een poppenkraam	2009			NotNom	Yes

Author	Author gender	Title	Year	Libris Literatuur Prijs	Boekenbon Literatuurprijs	Target	Balanced author gender subset
Harry Mulisch	Man	De Pupil	1987			NomAut	Yes
Harry Mulisch	Man	De ontdekking van de hemel	1992		Shortlist	NomNov	No
Harry Mulisch	Man	De Procedure	1999	Winner		NomNov	No
Harry Mulisch	Man	De Elementen	1988			NomAut	Yes
Charlotte Mutsaers	Woman	Koetsier Herfst	2009	Shortlist		NomNov	Yes
Marcel Möring	Man	Louteringsberg	2012	Longlist		NomNov	Yes
Willem Nijholt	Man	Met bonzend hart	2011			NotNom	Yes
Nelleke Noordervliet	Woman	Zonder noorden komt niemand thuis	2010	Longlist		NomNov	Yes
Nelleke Noordervliet	Woman	Vrij Man	2013	Longlist		NomNov	Yes
Nelleke Noordervliet	Woman	Snijpunt	2009	Longlist		NomNov	Yes
Michiel Klein Nulent	Man	Het Koekoeksei	2011			NotNom	No
Ellen Ombre	Woman	Maalstroom	1992			NotNom	Yes
Connie Palmen	Woman	Logboek van een onbarmhartig jaar	2011			NomAut	Yes
Connie Palmen	Woman	De Wetten	1991			NomAut	Yes
Connie Palmen	Woman	De Erfenis	1999			NomAut	Yes
Connie Palmen	Woman	De Vriendschap	1995		Winner	NomNov	Yes
Koen Peeters	Man	Grote Europese roman	2008	Shortlist		NomNov	Yes
Rascha Peper	Woman	Vossenblond	2011			NomAut	Yes
Rascha Peper	Woman	Dooi	1999			NomAut	Yes
Rascha Peper	Woman	Een Spaans hondje	1998			NomAut	Yes
Rascha Peper	Woman	Wie scheep gaat	2003			NomAut	Yes
Yves Petry	Man	De maagd Marino	2011	Winner		NomNov	Yes
Eefje Pleij	Woman	Juf met staarten krijgt een staartje	2008			NotNom	Yes
Chaja Polak	Woman	Verslag van een onaanvaarde dood	2007			NomAut	Yes
Chaja Polak	Woman	Wachten op de schemering	2007			NomAut	Yes
Anne Provoost	Woman	In de zon kijken	2008	Longlist		NomNov	Yes
Anil Ramdas	Man	De papegaai, de stier, en de klimmende bougainvillea	1992			NotNom	Yes
David van Reybrouck	Man	Congo	2010	Winner		NomNov	Yes
Elle van Rijn	Woman	De tragische geschiedenis van mijn succes	2006			NotNom	Yes
Thomas Rosenboom	Man	Zoete Mond	2010	Longlist		NomNov	Yes
Thomas rosenboom	Man	De nieuwe man	2003		Shortlist	NomNov	Yes
Helga Ruebsamen	Woman	Beer is terug	2000	Shortlist		NomNov	Yes
Ciel van Sambeek	Woman	Bloedzaaden	2011			NotNom	Yes
Ciel van Sambeek	Woman	Koninginnenrit	2008			NotNom	Yes
Peter Schaap	Man	De bruiden van Tyobar	1992			NotNom	Yes
Jaap Scholten	Man	De wet van Spengler	2009			NotNom	Yes
Jaap Scholten	Man	Morgenster	2009			NotNom	Yes
Jan Siebelink	Man	Verdwaald Gezin	1993			NomAut	No
Jan Siebelink	Man	Vera	1997			NomAut	No
Jan Siebelink	Man	Suezkade	2008			NomAut	Yes
Jan Siebelink	Man	Knielen op een bed violen	2005	Shortlist	Winner	NomNov	No
Jan Siebelink	Man	De overkant van de rivier	1990			NomAut	Yes
Jan Siebelink	Man	Engelen van het duister	2001			NomAut	Yes
Jan Siebelink	Man	Hartje zomer	1991			NomAut	Yes
Jan Siebelink	Man	Het lichaam van Clara	2010			NomAut	Yes
Mart Smeets	Man	De Afrekening	2010			NotNom	Yes
Susan Smit	Woman	Wat er niet meer is	2007			NotNom	Yes
Susan Smit	Woman	Wijze Mannen	2010			NotNom	Yes
F. Springer	Man	Kandy	1998			NomAut	No
F. Springer	Man	Bangkok, een elegie	2005			NomAut	No
Rosalie Sprooten	Woman	De pest voor een schip	1989			NotNom	Yes
Sophie van der Stap	Woman	Een blauwe vlinder zegt gedag	2008			NotNom	Yes
Bianca Stigter	Woman	De ontsproten Picasso	2008		Shortlist	NomNov	Yes
Henk van Straten	Man	Salvador	2012	Longlist		NomNov	No
Henk van Straten	Man	Superlul	2011			NomAut	Yes
Henk van Straten	Man	Kleine Stinkerd	2008			NomAut	No
Toon Tellegen	Man	Dora	1998			NomAut	Yes
Peter Terrin	Man	Post mortem	2012	Longlist	Winner	NomNov	Yes
Charles den Tex	Man	Cel	2009	Longlist		NomNov	Yes
Charles den Tex	Man	Spijt	2009			NomAut	Yes
Charles den Tex	Man	De macht van meneer Miller	2005			NomAut	Yes
Christiaan Thijm	Man	Het proces van de eeuw	2011			NotNom	Yes
Ed Thijn	Man	Kroonprinsenleed	2008			NotNom	Yes
Theo Thijssen	Man	De gelukkige klas	2007			NotNom	No
P.F. Thomése	Man	De weldoener	2011		Shortlist	NomNov	Yes
Anneloes Timmerije	Woman	Aus liebe	2009			NotNom	Yes
Willem Toorn	Man	Rooie en andere verhalen over mijzelf en mijn klas	1992			NotNom	Yes
Franca Treur	Woman	Dorsvloer vol confetti	2009			NotNom	Yes
Carolina Trujillo	Woman	De terugkeer van Lupe García	2009			NomNov	Yes
Betsy Udink	Woman	Allah & Eva	2006			NotNom	Yes
Monica Vanleke	Woman	Pelgrimstocht op hoge hakken	2011			NotNom	Yes
Annelies Verbeke	Woman	Vissen redden	2010	Longlist		NomNov	Yes
Alex Verburg	Man	Dwalingen	2009			NotNom	Yes
Paul Verhoeven	Man	Zwartboek	2006			NotNom	Yes
Dimitri Verhulst	Man	Godverdomse dagen op een godverdomse bol	2009			NomNov	Yes
Dimitri Verhulst	Man	De helaasheid der dingen	2006	Longlist	Shortlist	NomNov	Yes
Dimitri Verhulst	Man	Mevrouw Verona daalt de heuvel af	2007	Longlist	Shortlist	NomNov	Yes
Dimitri Verhulst	Man	Problemski Hotel	2003			NomAut	No
Dimitri Verhulst	Man	De kamer hiernaast	1999			NomAut	Yes
Dimitri Verhulst	Man	Dinsdagland	2004			NomAut	No
Hans Vervoort	Man	Kind van de Oost	1992			NotNom	No
Hans Vervoort	Man	Geluk is voor de dommen	2003			NotNom	No
Rachel Visscher	Woman	Zwarte Dauw	2011			NotNom	Yes
Arjan Visser	Man	Paganinipark	2011			NomAut	Yes
Carolijn Visser	Woman	Vrouwen in den vreemde	2008			NotNom	Yes
Erik Vlaminck	Man	Brandlucht	2012	Longlist		NomNov	Yes
Paul Vugts	Man	De strijd tegen de Amsterdamse onderwereld	2011			NotNom	Yes
Robert Vuijsje	Man	Alleen maar nette mensen	2009	Shortlist		NomNov	Yes
Christiaan Weijts	Man	Via Cappello 23	2009		Shortlist	NomNov	No
Christiaan Weijts	Man	Art 285b	2006			NomNov	No
Christiaan Weijts	Man	De etaleur	2010	Longlist		NomNov	No
Gerwin van der Werf	Man	Wild	2012	Longlist		NomNov	No
Lodewijk Wiener	Man	De verering van Quirina T.	2007			NomNov	Yes
Tommy Wieringa	Man	Caesarion	2009		Shortlist	NomNov	Yes
Tommy Wieringa	Man	Dit zijn de namen	2013	Winner		NomNov	No
Nachoem Wijnberg	Man	Politiek en liefde	2002			NotNom	No
Nachoem Wijnberg	Man	Divan van Ghalib	2009			NotNom	Yes
Leon de Winter	Man	Het recht op terugkeer	2008	Longlist	Shortlist	NomNov	Yes
Patrick Witte	Man	Blijf Thuis	2009			NotNom	Yes
Ivan Wolffers	Man	Onweer in de verte	2009			NotNom	Yes
Annejet Zijl	Woman	Bernhard	2010			NotNom	Yes
Joost Zwagerman	Man	Transito	2007		Shortlist	NomNov	No
Joost Zwagerman	Man	Duel	2011			NomAut	No
Joost Zwagerman	Man	Gimmick	1989			NomAut	No
Joost Zwagerman	Man	Vals licht	1992		Shortlist	NomNov	Yes
Joost Zwagerman	Man	Zes sterren	2002			NomAut	No
Joost Zwagerman	Man	De buitenvrouw	2009			NomAut	Yes

Appendix B

Classification on nominated and not nominated novels, including author gender variables

COMPLETE DATASET				
Class	Precision	Recall	F1-score	Number of novels
NomNov	0.565	0.700	0.625	100
NomAut	0.508	0.304	0.380	102
NotNom	0.600	0.704	0.648	98
Accuracy			0.567	300
BALANCED AUTHOR GENDER SUBSET				
Class	Precision	Recall	F1-score	Number of novels
NomNov	0.462	0.597	0.521	72
NomAut	0.471	0.286	0.356	84
NotNom	0.622	0.709	0.663	86
Accuracy			0.529	242

TABLE B.1: In this Table, the results of a logistic regression classification on **NomNov**, **NomAut** and **NotNom** trained on the 5000 most frequent unigrams and bigrams and the gender variables (man and woman). The results are lower than the score for the 5000 most frequent word features, without the gender variables (see Table 5.1). The same patterns are shown in the results, namely: **NomAut** has the lowest scores for the complete dataset, and **NomNov** has the lowest precision on the balanced author gender subset, and **NomAut** the lowest recall and F1-score.

COMPLETE DATASET				
Women	Precision	Recall	F1-score	Number of novels
NomNov	0.489	0.611	0.543	36
NomAut	0.480	0.286	0.358	42
NotNom	0.647	0.767	0.702	43
Accuracy			0.554	121
Men	Precision	Recall	F1-score	Number of novels
NomNov	0.608	0.750	0.671	64
NomAut	0.528	0.317	0.396	60
NotNom	0.562	0.655	0.605	55
Accuracy			0.575	179
BALANCED AUTHOR GENDER SUBSET				
Women	Precision	Recall	F1-score	Number of novels
NomNov	0.426	0.556	0.482	36
NomAut	0.500	0.262	0.344	42
NotNom	0.615	0.744	0.674	43
Accuracy			0.521	121
Men	Precision	Recall	F1-score	Number of novels
NomNov	0.500	0.639	0.561	36
NomAut	0.448	0.310	0.366	42
NotNom	0.630	0.674	0.652	43
Accuracy			0.537	121

TABLE B.2: In this Table, the results of a logistic regression classification on **NomNov**, **NomAut** and **NotNom** trained on the 5000 most frequent unigrams and bigrams and the gender variables (man and woman), split by author gender. The results are lower than the score for the 5000 most frequent word features, without the gender variables (see Table 5.5), except for the precision of **NomAut** novels written by women in the complete dataset, and the precision, recall and F1-score of **NomNov** novels written by men in the balanced author gender subset. For these four scores, the results were higher when the gender variables were included. Slightly different patterns are shown in these results, as **NomAut** have lowest scores overall for the complete dataset, and for the men in the balanced author gender subset. This differs from the results without the gender variables, as the precision is the lowest score in these results, alternating between the precision of the **NomNov** set and the **NotNom** set.

COMPLETE DATASET				
Class	Precision	Recall	F1-score	Number of novels
Nominated	0.579	0.660	0.617	100
Not nominated	0.817	0.760	0.788	200
Accuracy			0.727	300

BALANCED AUTHOR GENDER SUBSET				
Class	Precision	Recall	F1-score	Number of novels
Nominated	0.542	0.542	0.542	72
Not nominated	0.806	0.806	0.806	170
Accuracy			0.727	242

TABLE B.3: In this Table, the results of a logistic regression classification on nominated (**NomNov**) and not nominated (**NomAut** and **NotNom**) trained on the 5000 most frequent unigrams and bigrams and the gender variables (man and woman). The overall accuracies are lower than the score for the 5000 most frequent word features, without the gender variables (see Table 5.3). This due to lower the precision, recall and F1-score of the not nominated novels when the gender variables are added. The precision, recall and F1-score for the nominated novels are higher when the gender variables are added. This results in a lower overall accuracy, as the datasets contain more not nominated novels than nominated novels. The same patterns are shown in the results, namely: nominated novels have the lowest results for the complete dataset and the balanced author gender subset.

COMPLETE DATASET				
Women	Precision	Recall	F1-score	Number of novels
Nominated	0.541	0.556	0.548	36
Not nominated	0.810	0.800	0.805	85
Accuracy			0.727	121
Men	Precision	Recall	F1-score	Number of novels
Nominated	0.597	0.719	0.652	64
Not nominated	0.824	0.730	0.774	115
Accuracy			0.726	179
BALANCED AUTHOR GENDER SUBSET				
Women	Precision	Recall	F1-score	Number of novels
Nominated	0.486	0.500	0.493	36
Not nominated	0.786	0.776	0.781	85
Accuracy			0.694	121
Men	Precision	Recall	F1-score	Number of novels
Nominated	0.600	0.583	0.592	36
Not nominated	0.826	0.835	0.830	85
Accuracy			0.760	121

TABLE B.4: In this Table, the results of a logistic regression classification on nominated (**NomNov**) and not nominated (**NomAut** and **NotNom**) trained on the 5000 most frequent unigrams and bigrams and the gender variables (man and woman). Most overall accuracies are higher than the score for the 5000 most frequent word features, without the gender variables (see Table 5.3). Only the overall accuracy of the novels written by men on the complete dataset are lower.

Appendix C

Topics LDA Topic model

The names of the topics were manually chosen. For several topics, general themes could not be identified. Therefore, these topics are not labelled.

Topic 0: War 0.0174 majoor 0.0145 soldaat 0.0142 oorlog 0.0141 man 0.0134 generaal 0.0096 leger 0.0089 twee 0.0085 militair 0.008 luitenant 0.0066 komen 0.006 werden

Topic 1: School 0.0438 school 0.0299 klas 0.0195 jongen 0.0149 kind 0.0146 leerling 0.0121 les 0.0121 één 0.0095 meester 0.0094 leraar 0.0091 boek 0.009 student

Topic 2: Car, travelling 0.0299 auto 0.0231 rijden 0.0157 staan 0.0136 weg 0.0123 lopen 0.0118 komen 0.0104 zien 0.0096 straat 0.0086 twee 0.0084 gaan 0.0078 man

Topic 3 0.0543 zeggen 0.0276 wel 0.018 zullen 0.0179 jij 0.0164 gaan 0.0163 nou 0.0161 weten 0.0145 goed 0.0139 denken 0.0124 heel 0.0122 weer

Topic 4: Law enforcement, court of law 0.0107 advocaat 0.0098 rechtbank 0.0097 zaak 0.008 justitie 0.0074 cel 0.0073 rechter 0.0061 moord 0.0055 choreo 0.0054 twee 0.0048 zeggen 0.0046 jaar

Topic 5: Family 0.0663 moeder 0.0612 vader 0.0223 kind 0.0152 gaan 0.0132 broer 0.0128 komen 0.0122 ouder 0.0111 zeggen 0.0104 jaar 0.0099 huis 0.0093 zitten

Topic 6: Bar, pub, cafe 0.0112 bar 0.0111 groen 0.0102 bier 0.0095 drinken 0.0085 keer 0.0083 jongen 0.0077 twee 0.0075 lul 0.0074 echt 0.0071 meisje 0.0069 ieder

Topic 7: Life 0.0126 ander 0.0099 zoals 0.0091 leven 0.0088 bestaan 0.0082 mens 0.0079 waar 0.0076 wereld 0.0059 jaar 0.0055 tijd 0.0052 eigen 0.0051 waarin

Topic 8: Human body 0.0169 hand 0.0101 oog 0.0093 trekken 0.0073 hoofd 0.007 been 0.0063 laten 0.0062 houden 0.0062 arm 0.006 gezicht 0.006 mond 0.0058 alsof

Topic 9: Sea, sailing 0.0328 water 0.0284 zee 0.019 boot 0.0128 strand 0.0123 schip 0.0115 eiland 0.01 wind 0.0084 weer 0.0083 liggen 0.0082 zien 0.0078 dijk

Topic 10 0.0348 zullen 0.0123 komen 0.0111 gaan 0.011 wel 0.0106 weten 0.0102 ander 0.0096 laat 0.0096 één 0.0095 weer 0.0093 maken 0.0089 dag

Topic 11: Verbs 0.0711 zeggen 0.0242 willen 0.0221 gaan 0.0204 weten 0.0192 vragen 0.016 zullen 0.0129 goed 0.0128 denken 0.0125 kijken 0.0124 komen 0.0116 zien

Topic 12: Islam 0.0123 vrouw 0.0122 moskee 0.012 gaan 0.0111 man 0.0101 komen 0.0089 huis 0.0083 imam 0.0079 stad 0.0069 waar 0.0066 chinees 0.006 pont

Topic 13 0.0218 commissaris 0.0197 denken 0.0176 zeggen 0.0165 ara 0.0132 duif 0.012 vragen 0.0099 viool 0.0086 twee 0.0054 antwoorden 0.0052 dood 0.0048 wij

Topic 14: Art 0.0322 film 0.0206 zien 0.0188 foto 0.0161 schilderij 0.0129 beeld 0.012 maken 0.0116 camera 0.0114 kunst 0.0092 schilder 0.0088 kunstenaar 0.0082 werk

Topic 15: Letters, mail 0.0322 brief 0.0147 schrijven 0.0096 zeer 0.0074 fix 0.0058 heer 0.0058 amsterdam 0.0056 zwolle 0.0055 parijs 0.0055 wij 0.0054 slechts 0.0046 stad

Topic 16: Writing 0.0553 boek 0.045 schrijven 0.0245 lezen 0.0228 schrijver 0.0185 verhaal 0.0086 jaar 0.0085 zin 0.0078 één 0.0069 literatuur 0.0068 woord 0.0066 gedicht

Topic 17: Going home 0.0129 zien 0.0124 staan 0.0105 waar 0.0102 komen 0.0096 lopen 0.0086 huis 0.0086 weg 0.0081 liggen 0.0072 groot 0.007 boom 0.0066 weer

Topic 18 0.0316 zullen 0.0139 willen 0.0095 mens 0.0095 denken 0.009 mama 0.0087 nooit 0.0078 zeggen 0.0074 jij 0.0071 zelfs 0.0065 opnieuw 0.0065 misschien

Topic 19: Congo 0.0187 congo 0.0096 afrika 0.0066 kinshasa 0.0063 blank 0.0063 boeren 0.0059 zuid 0.0056 belgisch 0.0054 brits 0.0054 koloniaal 0.0048 congolees 0.0046 afrikaans

Topic 20: Life and death 0.0201 zullen 0.0188 leven 0.0156 willen 0.0135 dood 0.0131 weten 0.0099 laten 0.0089 denken 0.0078 voelen 0.0071 maken 0.0069 houden 0.0069 gaan

Topic 21: Christianity 0.033 god 0.0202 kerk 0.0117 jezus 0.0097 zeggen 0.0087 priester 0.0086 woord 0.0079 heilig 0.0079 dominee 0.0079 bijbel 0.0078 zullen 0.0073 zoon

Topic 22: Africa 0.0089 mercedes 0.0066 tour 0.0066 afrika 0.0054 renner 0.005 naam 0.005 wel 0.0046 kilometer 0.0042 nokia 0.0042 per 0.0036 één 0.0036 sahara

Topic 23: Second World War, Germany 0.0223 duits 0.012 prins 0.0103 duitsland 0.0091 oorlog 0.0082 joods 0.0078 duitsers 0.0072 wij 0.0069 koningin 0.0064 berlijn 0.0063 aus 0.0061 laat

Topic 24: Politics, international relations 0.0131 politiek 0.0112 land 0.0102 zullen 0.0091 amerikaans 0.0075 groot 0.0073 jaar 0.0068 amerika 0.0062 minister 0.0062 europa 0.0059 schrijven 0.0058 oorlog

Topic 25: Verbs 0.0303 gaan 0.0151 wel 0.0149 goed 0.0132 komen 0.0124 heel 0.0111 willen 0.0111 zien 0.0106 weer 0.0103 vinden 0.0102 zitten 0.0088 kijken

Topic 26: Verbs 0.0379 zeggen 0.0203 meneer 0.0173 zullen 0.0116 wel 0.0111 denken 0.0083 gaan 0.0082 jij 0.0071 goed 0.007 zitten 0.0066 vragen 0.0057 twee

Topic 27 0.0211 zullen 0.0113 onze 0.0091 gij 0.009 goed 0.0077 café 0.0077 wij 0.0073 één 0.0058 mogen 0.0056 moeten 0.0054 twee 0.0047 wel

Topic 28 0.0428 mevrouw 0.0373 hond 0.0177 wel 0.015 weer 0.0142 gaan 0.0138 komen 0.0129 meneer 0.0117 nee 0.0072 goed 0.0068 heel 0.0065 vragen

Topic 29: Corporate 0.0172 bouwfonds 0.0119 miljoen 0.0108 euro 0.0106 philips 0.0101 zeggen 0.0073 geld 0.007 betalen 0.0066 bedrijf 0.0065 project 0.0065 directeur 0.0065 pensioenfonds

Topic 30: Health care 0.0312 dokter 0.0195 patiënt 0.0172 ziekenhuis 0.0146 arts 0.0105 zeggen 0.0091 kind 0.0083 vrouw 0.0076 twee 0.0074 ziekte 0.0069 week 0.0062 zuster

Topic 31: Music 0.0232 spelen 0.0225 muziek 0.0094 zingen 0.0082 misschien 0.0079 twee 0.0068 piano 0.0067 elkaar 0.0062 gaan 0.0061 natuurlijk 0.0059 denken 0.0058 alleen

Topic 32 0.0106 oom 0.0103 ieder 0.0089 twee 0.0052 alkmaar 0.0044 even 0.0041 amsterdam 0.0036 ooit 0.0036 inmiddels 0.0036 natuurlijk 0.0034 goedemorgen 0.003 blijken

Topic 33 0.0126 goed 0.0104 gaan 0.0092 komen 0.008 maken 0.0072 jaar 0.0072 krijgen 0.0069 groot 0.0067 wel 0.0065 moeten 0.0059 ander 0.0054 geven

Topic 34: Pakistan 0.0134 broeder 0.0116 man 0.0107 vrouw 0.0102 pakistan 0.0088 god 0.0081 non 0.0065 pap 0.0056 hen 0.005 pakistaans 0.0049 mam 0.0047 krijgen

Topic 35: Love 0.0761 vrouw 0.0562 man 0.0179 meisje 0.0127 liefde 0.0122 kind 0.0121 ander 0.0113 jong 0.01 jaar 0.0097 mooi 0.009 leven 0.0081 trouwen

Topic 36: Dinner 0.0122 staan 0.0102 eten 0.0097 twee 0.0097 tafel 0.0093 zitten 0.0087 huis 0.0082 jaar 0.0075 dag 0.007 oud 0.0068 nemen 0.0067 maken

Topic 37 0.0306 zeggen 0.0097 één 0.0092 kijken 0.0092 daarna 0.0082 denken 0.0077 oog 0.0068 wanneer 0.0067 keer 0.0066 glas 0.0064 heel 0.0064 twee

Topic 38: Office 0.0158 bellen 0.0157 zeggen 0.015 weten 0.0125 telefoon 0.0091 zitten 0.0088 naam 0.0087 waar 0.0073 twee 0.0073 kantoor 0.0073 bureau 0.0064 computer

Topic 39: Ships, sailing 0.0301 schip 0.0153 niesen 0.0125 kapitein 0.0111 gaan 0.0104 boord 0.0084 man 0.0074 weer 0.0065 matroos 0.0063 varen 0.0062 zee 0.0056 hut

Topic 40: Darwin 0.0128 zullen 0.0111 mens 0.0092 dier 0.009 zien 0.0077 water 0.0074 vis 0.0071 groot 0.0067 één 0.0062 leven 0.0062 onze 0.0062 maken

Topic 41: Hotel, travelling 0.0137 hotel 0.0134 jaar 0.0123 stad 0.0087 waar 0.0073 land 0.007 dag 0.0061 nederland 0.006 uur 0.0059 oud 0.0056 twee 0.0055 reis

Topic 42: Surinam 0.0128 zwart 0.0095 papegaai 0.009 stier 0.008 klimmend 0.008 bougainvillea 0.008 suriname 0.0069 ashirwad 0.0068 blank 0.0064 maalstroom 0.0063 ombre 0.0056 india

Topic 43 0.0303 zeggen 0.0268 kijken 0.0202 staan 0.0172 zien 0.0158 hand 0.0128 man 0.0124 lopen 0.0115 komen 0.0113 gaan 0.0091 gezicht 0.009 oog

Topic 44 0.0153 zullen 0.0141 willen 0.0138 komen 0.0119 gaan 0.008 mamma 0.0072 heel 0.0056 maken 0.0055 denken 0.0055 pappa 0.005 moment 0.005 hand

Topic 45 0.0142 pastoor 0.0123 koning 0.0102 hen 0.0077 ander 0.0074 enkel 0.0064 maarschalk 0.0064 eiland 0.006 baron 0.006 werden 0.0053 schip 0.0051 wendag

Topic 46: English words 0.0645 the 0.0257 you 0.0234 and 0.0091 for 0.0088 that 0.0067 new 0.0062 your 0.0057 with 0.0057 what 0.0054 not 0.0048 are

Topic 47: Islam 0.0344 jullie 0.02 allah 0.0197 hen 0.0188 zeggen 0.0169 god 0.0169 wij 0.0145 zullen 0.0118 geven 0.009 komen 0.0087 mens 0.0081 vrouw

Topic 48: Sleeping at home 0.02 bed 0.0183 gaan 0.018 kamer 0.0167 deur 0.0164 liggen 0.0155 staan 0.0111 huis 0.0106 zitten 0.0104 komen 0.0096 kijken 0.0096 slapen

Topic 49: France 0.009 brief 0.0086 parijs 0.0076 schrijven 0.0071 heer 0.0065 mevrouw 0.0061 naam 0.0054 wel 0.0053 jaar 0.005 onderzoek 0.0049 paar 0.0049 monsieur

Appendix D

Dendogram Cosine Delta

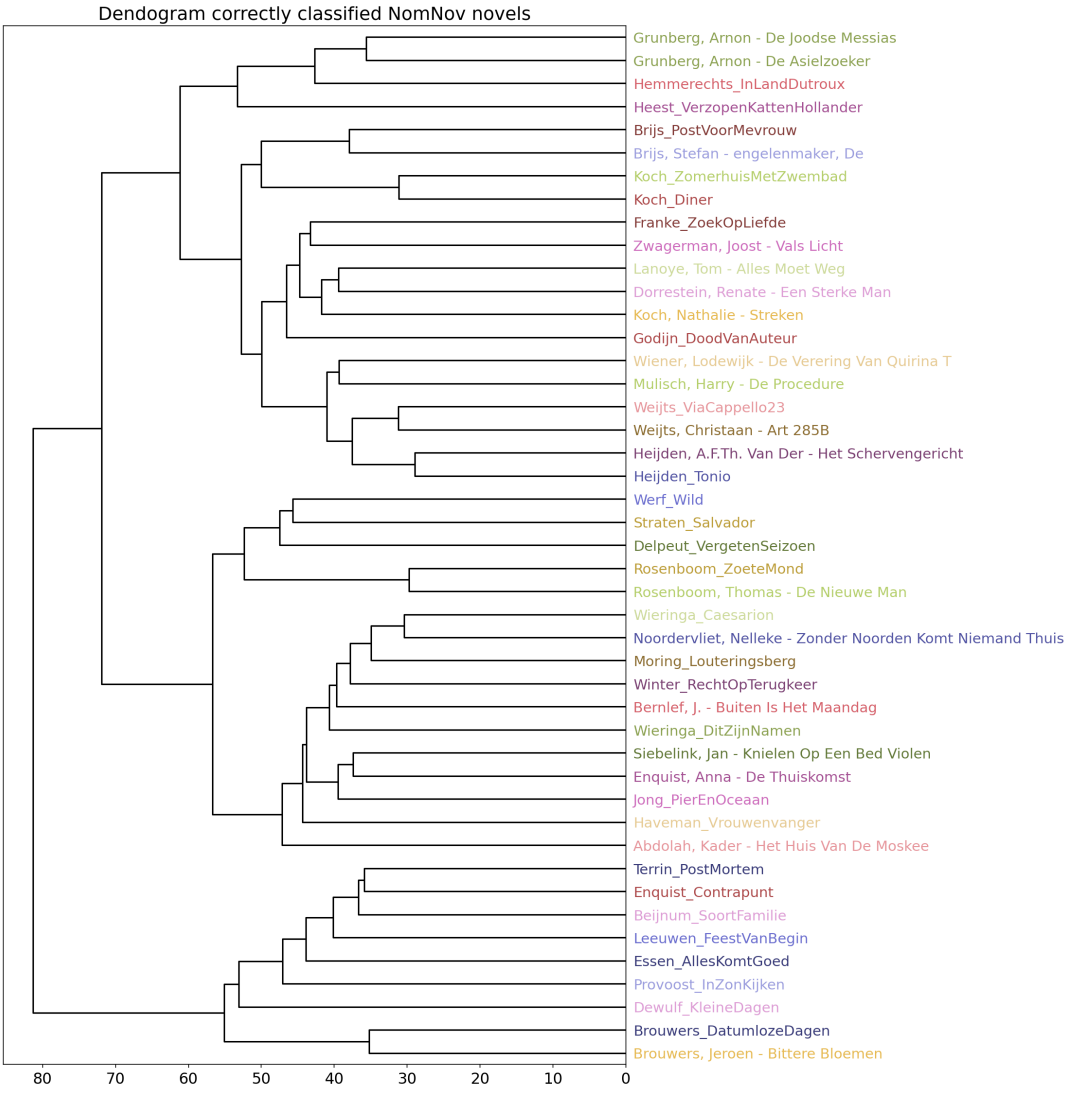

FIGURE D.1: Dendogram correctly classified novels, showing how the writing styles of the novels relate to each other for not nominated novels.

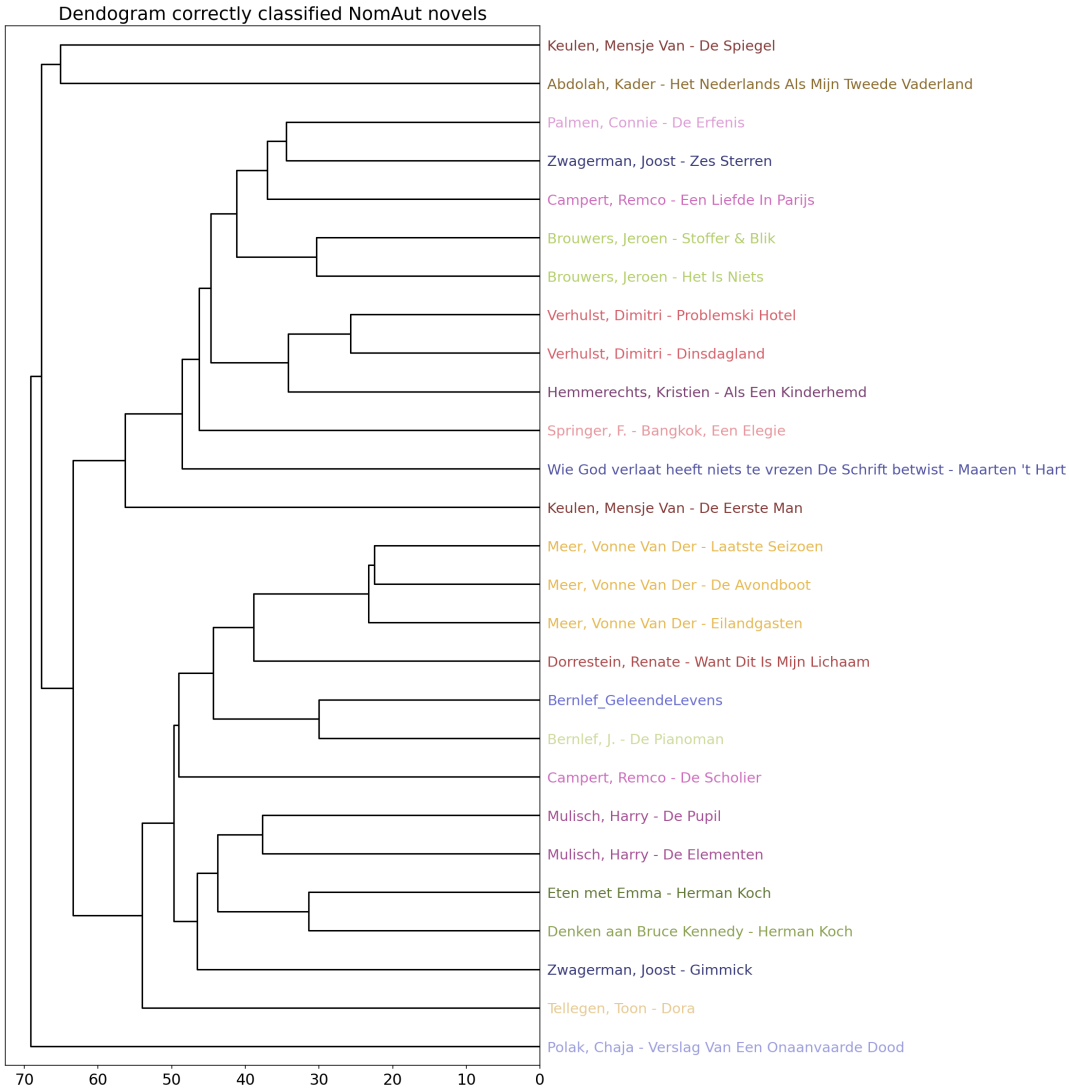

FIGURE D.2: Dendogram correctly classified novels, showing how the writing styles of the novels relate to each other for not nominated novels written by nominated authors.

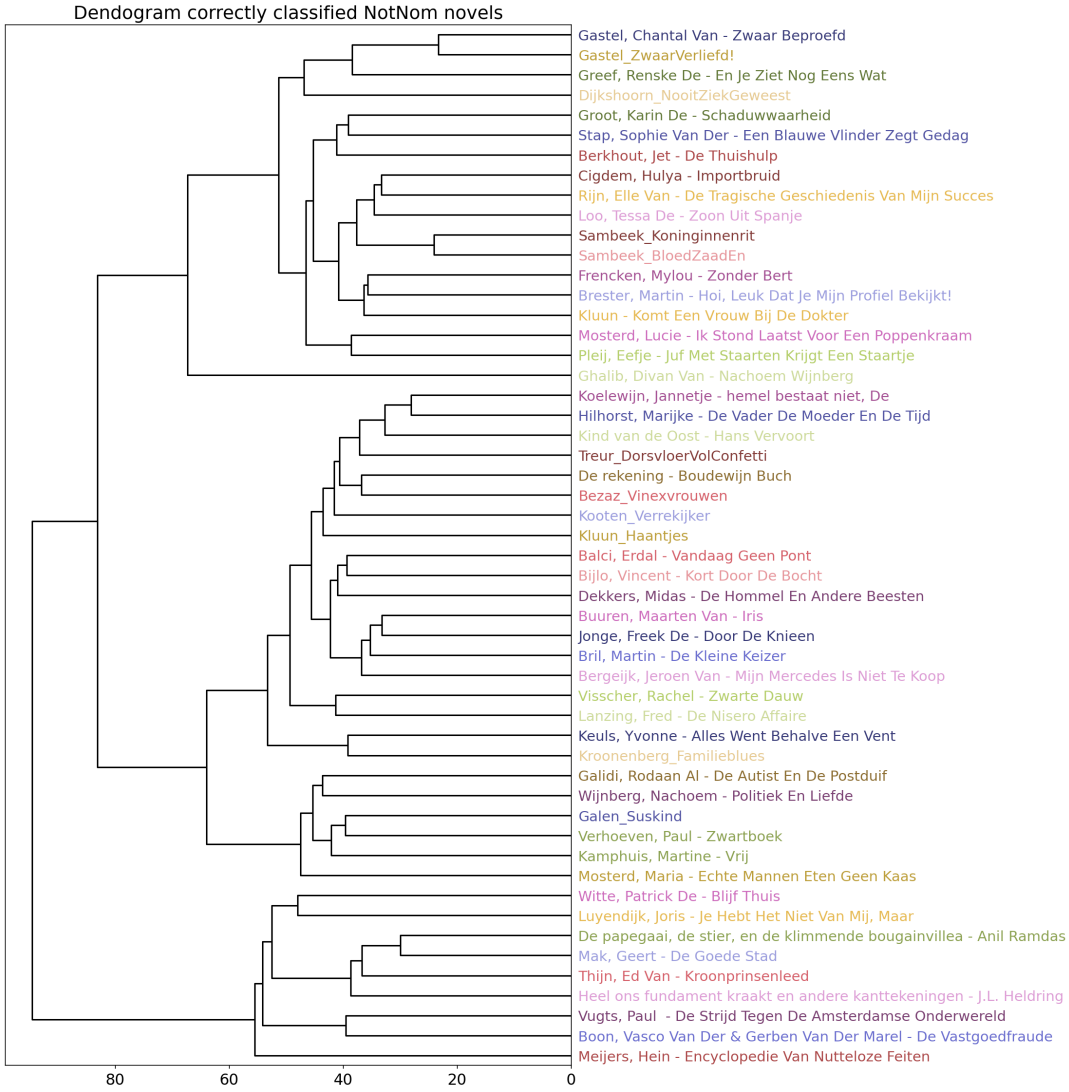

FIGURE D.3: Dendogram correctly classified novels, showing how the writing styles of the novels relate to each other for not nominated novels written by not nominated authors.

Bibliography

Ahmed, Sara (2012). *On being included*. Duke University Press, pp. 1–17.

Amatmoekrim, K (2015). "Een monoculturele uitwas. De ondraaglijke witheid van de Nederlandse letteren." In: *De Groene Amsterdammer*.

Andrzejewski, David and Xiaojin Zhu (2009). "Latent Dirichlet Allocation with topic-in-set knowledge". In: June, pp. 43–48. DOI: 10.3115/1621829.1621835.

Argamon, Shlomo et al. (2003). "Gender, genre, and writing style in formal written texts". In: *Text* 23.3, pp. 321–346. ISSN: 18607349. DOI: 10.1515/text.2003.014.

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen (2014). "Gender identity and lexical variation in social media". In: *Journal of Sociolinguistics* 18.2, pp. 135–160. ISSN: 14679841. DOI: 10.1111/josl.12080. arXiv: 1210.4567.

Bamman, David, Ted Underwood, and Noah A. Smith (2014). "A bayesian mixed effects model of literary character". In: *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference* 1, pp. 370–379. DOI: 10.3115/v1/p14-1035.

Beauvoir, Simone de (1953). *The second sex*. Alfred Knopf New York.

Berkers, Pauwke (2009). *Classification into the literary mainstream? Ethnic boundaries in the literary fields of the United States, the Netherlands and Germany, 1955-2005*.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3, pp. 993–1022.

Boudewijn, Petra (2020). "'And the award goes to…'Women on the Dutch literary award scene". In: *Journal of Dutch Literature* 11.1.

Burrows, John (2002). "'Delta': a measure of stylistic difference and a guide to likely authorship". In: *Literary and linguistic computing* 17.3, pp. 267–287.

Butler, Judith (1998). "Imitation and Gender Insubordination". In: *Literary Theory: An Anthology*. Blackwell Publishers, pp. 722–730.

Cranenburgh, Andreas van and Rens Bod (2017). "A Data-Oriented Model of Literary Language". In: *arXiv preprint arXiv:1701.03329*.

Cranenburgh, Andreas van and Corina Koolen (2020). "Results of a Single Blind Literary Taste Test with Short Anonymized Novel Fragments". In: arXiv: 2011.01624. URL: http://arxiv.org/abs/2011.01624.

Deijl, LA van der, RJH Smeets, and APJ van den Bosch (2019). "The Canon of Dutch Literature According to Google". In: *Journal of Cultural Analytics*. DOI: https://doi.org/10.22148/16.046.

Dera, J (2020). "The Cultural Diversity of Text Selection in Dutch Literary Education: An Analysis of Reading Tips, Teaching Packs, and Student Choices." In: preprint on webpage at https://doi.org/10.31234/osf.io/a9tuq.

— (2021). "De helaasheid der leeslijsten. Over diversiteit in het literatuuronderwijs". In: *De Lage Landen* 64 (1), pp. 115–121.

Devinney, Hannah, Jenny Björklund, and Henrik Björklund (2020). "Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish". In: *GeBNLP2020,*

COLING'2020–The 28th International Conference on Computational Linguistics, December 8-13, 2020, Online. Association for Computational Linguistics, pp. 79–92.

Dijkgraaf, Margot and René Appel (2013). *Vrouwen, mannen en de Libris Literatuur Prijs*. Stichting Libris Literatuur Prijs.

Eckert, Penelope (2012). "Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation". In: *Annual Review of Anthropology* 41, pp. 87–100. ISSN: 00846570. DOI: 10.1146/annurev-anthro-092611-145828.

Evert, Stefan et al. (2017). "Understanding and explaining Delta measures for authorship attribution". In: *Digital Scholarship in the Humanities* 32.suppl_2, pp. ii4–ii16.

Fanon, Frantz (2008). *Black skin, white masks*. Grove press, pp. 1–24.

Fast, Ethan, Tina Vachovsky, and Michael S. Bernstein (2016). "Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community". In: *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pp. 112–120. arXiv: 1603.08832.

Goldstone, Andrew and Ted Underwood (2014). "The quiet transformations of literary studies: What thirteen thousand scholars could tell us". In: *New Literary History* 45.3, pp. 359–384.

Herring, Susan C. and John C. Paolillo (2006). "Gender and genre variation in weblogs". In: *Journal of Sociolinguistics* 10.4, pp. 439–459. ISSN: 13606441. DOI: 10.1111/j.1467-9841.2006.00287.x.

Herrmann, J Berenike, Arthur M Jacobs, and Andrew Piper (2021). "Computational Stylistics". In: *Handbook of Empirical Literary Studies*, p. 451.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear.

Jelodar, Hamed et al. (2019). "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey". In: *Multimedia Tools and Applications* 78.11, pp. 15169–15211.

Keyes, Os, Chandler May, and Annabelle Carrell (2021). "You Keep Using That Word: Ways of Thinking about Gender in Computing Research". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1, pp. 1–23.

Koolen, Corina (2018). *Reading beyond the female: The relationship between perception of author gender and literary quality*. Universiteit van Amsterdam.

— (Sept. 2020). *Dit is geen vrouwenboek*. 1st ed. Amsterdam: HarperCollins. Chap. 3, pp. 71–87.

Koolen, Corina and Andreas van Cranenburgh (2017). "These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution". In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 12–22.

Koolen, Corina et al. (2020). "Literary quality in the eye of the Dutch reader: The National Reader Survey". In: *Poetics* 79.February, p. 101439. ISSN: 0304422X. DOI: 10.1016/j.poetic.2020.101439. URL: https://doi.org/10.1016/j.poetic.2020.101439.

Koren, Timo and Christine Delhaye (2019). "Depoliticising literature, politicising diversity: ethno-racial boundaries in Dutch literary professionals' aesthetic repertoires". In: *Identities* 26.2, pp. 184–202. ISSN: 15473384. DOI: 10.1080/1070289X.2017.1391561. URL: https://doi.org/10.1080/1070289X.2017.1391561.

Kraicer, Eve and Andrew Piper (2019). "Social characters: the hierarchy of gender in contemporary English-language fiction". In: *Journal of Cultural Analytics* 1.1, p. 11055.

Lejun, Gong, Tang Xiangyu, and Li Huakang (2021). "Analysis of Literary based on Deep Emotional Network". In: *2021 7th International Conference on Big Data Computing and Communications (BigCom)*. IEEE, pp. 227–233.

Lucy, Li et al. (2020). "Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks". In: *AERA Open* 6.3, p. 233285842094031. ISSN: 2332-8584. DOI: 10.1177/2332858420940312.

Lupei, Maksym et al. (2020). "Identification of authorship of Ukrainian-language texts of journalistic style using neural networks". In: pp. 30–36. DOI: 10.15587/1729-4061.2020.195041.

Marsden, John et al. (2013). "Language individuation and marker words: Shakespeare and his Maxwell's demon". In: *PloS one* 8.6, e66813.

McKinney, Wes et al. (2010). "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX, pp. 51–56.

Nguyen, Dong et al. (2014). "Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment". In: *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, pp. 1950–1961.

Nguyen, Dong et al. (2016). "Computational sociolinguistics: A survey". In: *Computational Linguistics*. ISSN: 15309312. DOI: 10.1162/COLI_a_00258. arXiv: 1508.07544.

Oyewumi, Oyeronke (2002). "Conceptualizing gender: the eurocentric foundations of feminist concepts and the challenge of African epistemologies". In: *Jenda: A Journal of Culture and African Women Studies* 2.1, pp. 1–9.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Ramdas, A (1997). "Moedwil en kwade trouw bij blanke schrijvers. Niemand heeft oog voor het vreemde". In: *NRC Handelsblad*.

Rouw, E (2015). "Literatuur blijft te wit". In: *NRC Handelsblad*.

Rybicki, Jan (2016). " Vive la différence : Tracing the (authorial) gender signal by multivariate analysis of word frequencies ". In: *Digital Scholarship in the Humanities* 31.4, pp. 746–761. ISSN: 2055-7671. DOI: 10.1093/llc/fqv023.

Schmidt, Benjamin M (2012). "Words alone: Dismantling topic models in the humanities". In: *Journal of Digital Humanities* 2.1, pp. 49–65.

Smeets, RJH, EP Sanders, and APJ van den Bosch (2019). "Character Centrality in Present-Day Dutch Literary Fiction". In: *Digital Humanities Benelux Journal*, pp. 71–90.

Smith, Peter WH and William Aldridge (2011). "Improving Authorship Attribution: Optimizing Burrows' Delta Method". In: *Journal of Quantitative Linguistics* 18.1, pp. 63–88.

Staszak, Jean-François (2008). "Other/otherness". In: *International Encyclopedia of Human Geography*.

Stichting Literatuur Prijs (2021). *Over de Libris Literatuur prijs*. URL: https://www.librisprijs.nl/index.php/over-de-libris-literatuur-prijs.

Tuzzi, Arjuna and Michele A Cortelazzo (2018). "What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer". In: *Digital Scholarship in the Humanities* 33.3, pp. 685–702.

Underwood, Ted, David Bamman, and Sabrina Lee (2018). "The Transformation of Gender in English-Language Fiction". In: *Journal of Cultural Analytics*, pp. 1–25. DOI: 10.22148/16.019.

Underwood, Ted and Jordan Sellers (2012). "The emergence of literary diction". In: *Journal of Digital Humanities* 1.2, pp. 1–2.

Van Der Deijl, Lucas et al. (2016). "Mapping the Demographic Landscape of Characters in Recent Dutch Prose Journal of Dutch Literature". In: *Journal of Dutch Literature* 7.1, pp. 20–42. URL: http://www.revisor.nl/entry/2095/slachtoffers-positiebepaling..

Varela, Paulo et al. (2016). "A computational approach for authorship attribution of literary texts using sintatic features". In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 4835–4842.

Varela, Paulo et al. (2018). "A Computational Approach for Authorship Attribution on Multiple Languages". In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2021). "Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law". In: *SSRN Electronic Journal*, pp. 1–51. DOI: 10.2139/ssrn.3792772.

Weijers, N (2014). "Vrouwen *schrijven* niet met hun *tieten*". In: *NRC Handelsblad*.

Wekker, Gloria et al. (2016). *Let's do Diversity: report of the University of Amsterdam Diversity Commission*. University of Amsterdam.