Journal of Data Science 22 (1), 77–96
January 2024

DOI: 10.6339/23-JDS1100
Data Science in Action

Quantifying Gender Disparity in Pre-Modern English Literature using Natural Language Processing

Mayank Kejriwal[1,*] and Akarsh Nagaraj[1]

[1]Information Sciences Institute, Viterbi School of Engineering, University of Southern California, Marina del Rey, CA, United States of America

Abstract

Research has continued to shed light on the extent and significance of gender disparity in social, cultural and economic spheres. More recently, computational tools from the data science and Natural Language Processing (NLP) communities have been proposed for measuring such disparity at scale using empirically rigorous methodologies. In this article, we contribute to this line of research by studying gender disparity in 2,443 copyright-expired literary texts published in the pre-modern period, defined in this work as the period ranging from the beginning of the nineteenth through the early twentieth century. Using a replicable data science methodology relying on publicly available and established NLP components, we extract three different gendered character prevalence measures within these texts. We use an extensive set of statistical tests to robustly demonstrate a significant disparity between the prevalence of female characters and male characters in pre-modern literature. We also show that the proportion of female characters in literary texts significantly increases in female-authored texts compared to the same proportion in male-authored texts. However, regression-based analysis shows that, over the 120 year period covered by the corpus, female character prevalence does not change significantly over time, and remains below the parity level of 50%, regardless of the gender of the author. Qualitative analyses further show that descriptions associated with female characters across the corpus are markedly different (and stereotypical) from the descriptions associated with male characters.

Keywords *digital humanities; gender-specific character prevalence; named entity recognition; project Gutenberg; word embedding*

1 Introduction

Recent innovations in deep neural networks have led to impressive advances in Natural Language Processing (NLP) (Devlin et al., 2018; Wolf et al., 2020). These advances include new state-of-the-art results in tasks as diverse as question answering, information extraction, sentiment analysis, conversational 'chatbot' agents and summarization, to only name a few (Reddy et al., 2019; Han and Wang, 2021; Naseem et al., 2020; Siblini et al., 2019; Liu, 2019). Due to the performance of these models, it has also become possible in recent years to use NLP tools for computational social science and digital

humanities. Jarynowski et al. (2019) provide an introduction to sociologists for modelling communities and populations using computational methods, including both data-driven (which is

*Corresponding author. Email: kejriwal@isi.edu.

Received May 26, 2022; Accepted April 22, 2023

referred to as "black-box" in their work) and declarative rule-based techniques and algorithms. NLP is one set of techniques they cover in their treatment. To quote from their article "As sociology has processed human written or spoken signals...NLP techniques have been widely applied in qualitative analyses" (Jarynowski et al., 2019).

Such methods are especially important for obtaining quantitative results at scale on large datasets that are not possible to examine in a fully manual manner without expending extensive labor and cost. Recognizing this, an illuminating set of use-cases, applications, and fundamental research has emerged in NLP conferences and workshops. An example is the set of peer-reviewed papers published in the *Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science* that was organized in the prestigious 55th Annual Meeting of the Association for Computational Linguistics (Hovy et al., 2017). Despite only being the second edition, the workshop included a range of papers applying advanced NLP methods specifically in support of computational social science, examples including understanding Twitter user behavior across demographic groups (Wood-Doughty et al., 2017), cross-lingual tagging of topics in political documents (Glavaš et al., 2017), and quantifying linguistic features of influence in informal and organic interactions (Prabhumoye et al., 2017). This line of work is not just limited to workshop proceedings. Milli and Bamman (2016) present results in a comprehensive journal article demonstrating gender differences in political discussions on Twitter, using computational methods. Other papers that rely heavily on scalable data analysis and computational methods, but with aims that are similar to more traditional social science research, include work by Rodriguez and Storer (2020); Burley et al. (2020); Mason et al. (2014); Keuschnigg et al. (2018). These works cover a broad range of subjects in sociology, from analysis of social media (Rodriguez and Storer, 2020) and social computing (Mason et al., 2014), to a study of state-led mass killings (Burley et al., 2020), and analytical sociology (Keuschnigg et al., 2018).

Inspired by these findings, this article proposes to use computational methods from the NLP community, implemented in open-source, industrial-grade packages, to quantify and explore the phenomenon of *gender-specific character prevalence* in pre-modern literature. Given a specific gender, gender-specific character prevalence may be defined as the number of *mentions* of characters having that gender within the text of the book. As discussed subsequently, there are at least three plausible ways to define a 'mention.' One plausible way is to count the number of *named* mentions of all characters having the gender. Another is to count the number of pronouns, since each pronoun can also be interpreted as a mention.

For our empirical study, we use a subset of the publicly available English-language texts in the Project Gutenberg corpus (with links and selection methodology provided subsequently in Section 3). These texts were published in the period ranging from the early nineteenth to the early twentieth century. Many texts published within this period, which coincides significantly with the Victorian era, are now considered historically and culturally significant, including (among others) novels by Charles Dickens and Elizabeth Barrett Browning (Adams, 2012).

In the extreme case, a text could contain characters that, with one or two exceptions are all male. An example is *Treasure Island* by Robert Louis Stevenson, a classic novel first published as a book in 1883 (Stevenson, 1883). The book has several major characters, all of whom as male, which is also the case for almost all of the minor characters. Although it was much more uncommon for a book during the pre-modern period to contain all (or even a majority of) female characters, some books do have many more female characters than male characters. A paradigmatic example is the coming-of-age novel *Little Women* by May Alcott (1868), published originally in two volumes in 1868 and 1869, but available today as a single volume. This book is a rare example where the majority of the characters are female, as the story follows the lives of four sisters from childhood to womanhood.

There are socio-cultural reasons why studying such character-based gender disparity, in aggregate, in these texts is important. First, studies continue to show that there is a gap between male and female representation in various cultural, political, scientific, and economic spheres (Yang et al., 2020; Nixon, 1994; Miller, 2016). We discuss the connection between disparities in gender representation in culture and literature to gender bias in Section 2; herein, we note that cultural movements and trends generally play an important role in shaping a society's values (Rochon, 2000; Katz, 1999). More recently, this has been witnessed firsthand both in the global diffusion (and homogenization, often along Western lines) of culture (Belkhyr, 2013; Rosenmann, 2016), and in rapid assimilation of Internet 'culture' and memes into mainstream culture (Nath and Murthy, 2004). Literary texts were an important component of the culture in the period that we are studying in this article (John, 2016), as there was no radio, television or Internet in that era. Hence, there is an argument to be made that gender disparity in the prevalence and qualitative description of female characters played a non-trivial cultural role in perpetuating gender bias in the broader society.

Even if such a causal link is not possible to scientifically confirm, the possibility suggested by the evidence raises the question of whether, and to what extent, such disparity was present in the thousands of texts (read widely even today) published in that period. Fortunately, the maturity of computational, NLP-based techniques and the availability of data from Project Gutenberg, allows us to measure such disparity with strong statistical guarantees.

The broad aim of this article is to compute both male- and female-specific character prevalence using several robust definitions of gendered character mentions. This aim is similar, in structure, to that of the *Gender Novels Project*, which was a similar study (and over a similar corpus) led by Digital Humanities Lab, MIT (2022). According to the website, their goal was to "build computational tools that helped us understand how conceptions of gender were expressed and changed through distant reading of thousands of books." However, there are some important differences between that project and ours both in the specific hypotheses that we investigate and in the statistical analysis underlying each hypothesis. We contextualize these differences further when describing our research hypotheses below.

1.1 Research Hypotheses

Three specific hypotheses are investigated in this work in the context of the Project Gutenberg corpus mentioned earlier. We also draw comparisons to other relevant work (most notably, the *Gender Novels Project*) and briefly state the main similarities and differences between that work and our investigation.

- **Hypothesis 1: Female-specific character prevalence is less than male-specific character prevalence.** This hypothesis is best related to the larger goal of this article, which is to investigate if female-specific character prevalence is indeed significantly lower than male-specific character prevalence. We note that there is some qualitative support for this hypothesis in the literature, and in more recent years, results from the Gender Novels Project have also yielded quantitative support. Specifically, when gender-specific character prevalence is measured using pronouns (which is one of the three methods we adopt), the Gender Novels Project has found that female-specific pronouns are less than half that of male-specific pronouns in the overall corpus. However, unlike our investigation, their result was not tested for statistical significance (Digital Humanities Lab, MIT, 2022). Another key difference between our investigation and theirs is that, methodologically, we investigate the hypothesis using two other prevalence measures, as defined subsequently in Section 3. Although these measures are correlated, they consider the measurement of prevalence from different viewpoints and enable a more robust inference. We expand on the underlying theoretical basis behind each of these measurements in Section 3.3, and also provide an appropriate statistical correction for addressing the correlation.

- **Hypothesis 2: The difference between male- and female-specific character prevalence significantly declines when stratified by the gender of books' authors.** This hypothesis aims to understand the impact (if any) on Hypothesis 1, if we stratify by the gender of the book's author. Although a number of variables could be considered as a stratifying variable (e.g., the genre of the book, or other variables related to the textual content itself), we choose to investigate the specific variable of the gender of the book's author because of work that has demonstrated the impact that greater diversity (in a given field or sphere of study) can have on increasing representation (Yang et al., 2020). Similar to Hypothesis 1, the Gender Novels Project comes closest to investigating this question. They find, for example, that female pronoun counts significantly increase, as a proportion of pronoun counts, in female-authored texts versus male-authored texts. The result is also statistically significant at the 95% confidence level. However, our methodology considers two other measures of gender-specific character prevalence (similar to our treatment of Hypothesis 1), and we tabulate a much more detailed set of statistics than is available for their finding.
- **Hypothesis 3: Female-specific character prevalence changes significantly, relative to male-specific character prevalence, in the 120-year period (1800 to 1920) spanning the majority of books in our corpus.** This hypothesis aims to understand if female-specific character prevalence has been increasing relative to male-specific character prevalence (at least approximately) over time. Certainly, in the period that is covered by the corpus, broader societal forces were in play that ultimately led to more rights for women in parts of the Western world. Although the suffrage movement is cited as an important historical example by O'Connor (1996), historians have recognized other such activities occurring well in the early nineteenth century. Both gender and class are also well-recognized today to be key organizing principles of Victorian society (Tusan, 2004). We note that, because this hypothesis is seeking to measure changes over time, as opposed to prevalence statistics, the underlying methodology is more complex. While the Gender Novels Project also includes some data on whether there are changes in pronoun counts by publication date (and also publication location, which we do not consider), significance results are missing from the analysis. Also, we measure the trend over time by using a regression-based model, and we also report results separately by the gender of the author, which is not the case for their analysis.

Finally, we note that, while the three hypotheses are quantitative in nature, we also complement the analysis by conducting a *qualitative* assessment of the kinds of words associated with male and female character occurrences, using computational techniques from NLP. Full details on the methodology and results of this analysis are provided in the supplementary material.

2 Related Work

A number of recent social science studies, some of which are computational, have shown that gender bias continues to exist in many aspects of economic, social and cultural life, including movies (Montasseri et al., 2020), executive positions in top corporations (Jordan et al., 2007), board membership (Burke and Mattis, 2013), and political leadership (Setzler, 2019). Legally, gender bias is defined by Legal Information Institute, Cornell Law School (2020) as "a person receiving different treatment based on the person's real or perceived gender identity". Under United States legal doctrine, it is considered synonymous with "discrimination on the basis of sex." A lack of gender representation on a corporate board may be (although is not necessarily) a consequence of gender bias. However, a lack of adequate representation of characters of a certain gender in culture (including art, poetry and fictional text) does not fall under the definition of gender bias. On the contrary, artistic expression is often protected in Western nations by free speech laws.

For these reasons, we argue that it is technically more appropriate to refer to a lack of representation of a certain gender within cultural and written accounts as *gender disparity*. Gender disparity may be associated with gender bias (but again, this is not a necessary relation), but at times, such disparity may occur by volition e.g., in the specific context of a women's studies course, it may well be that female authors and figures are over-represented in the reading list of the course's syllabus. However, a *systematic* difference can be argued to be problematic, especially if it persists over time, occurs in a cultural body of work that is sufficiently general, and is significantly and consistently different from the underlying population. In Section 5, we further discuss and cite the significance of gender disparity in culture (and other forms of cultural diffusion) in influencing social values for the worse.

While researchers have tackled the challenge of accurately measuring such disparity in large and fact-based corpora like Wikipedia (or even the news) using computational methodologies (Reagle and Rhue, 2011), such studies have been generally lacking in literature and cultural corpora, especially those that have not been published first on the Web (and thereby lack additional context, such as hyperlinks). The closest study that we are aware of is the aforementioned *Gender Novels Project* (Digital Humanities Lab, MIT, 2022). This project closely matches the goals of this work and was conducted independently. Similar to the research herein, their goal is to "study the ever-adapting and changing view on gender by writers all around the globe in the nineteenth and twentieth centuries." Our study serves as a complement to the Gender Novels Project, but we consider a broader set of statistical analyses than they do, and there are also key differences in the scope of our hypotheses and corresponding methodologies.

One reason for the dearth of work in this area, despite the availability of many out-of-copyright texts, is that (until quite recently) using NLP techniques for such studies was non-trivial due to concerns over quality, which is paramount in studies of this nature. This may be the reason why there was much more focus on pronouns (e.g., by Digital Humanities Lab, MIT (2022)) than on named mentions, since the former is easier to extract computationally from text, than the latter, which requires tuning named entity recognition models (Nadeau and Sekine, 2007). However, continuing improvements in NLP suggest that the time is ripe for conducting such studies, with appropriate quality control measures in place (Hovy et al., 2017).

A number of papers have also proposed using new metrics for quantifying important sociological phenomena, which is related to the goals of this article. For example, Hu and Kejriwal (2022) use Twitter to analyze "spatio-textual affinity" between different cities by adapting nearest-neighbors and information-theoretic metrics. Other examples, where such metrics (typically in the social media setting e.g., see Peters et al. (2013)) have been adapted to study sociological phenomena, include work by Montjoye et al. (2013) for predicting personality through mobile phone usage, and Cabrera et al. (2018) for studying alternative metrics for academic promotion and tenure. Although our primary contribution in this article is not the proposal of new metrics for studying gender disparity in literature, we do propose a broader set of metrics and statistical procedures for measuring this phenomenon than is the case in similar analytical projects, such as the Gender Novels Project (Digital Humanities Lab, MIT, 2022).

While our study is primarily concerned with quantifying disparities in character prevalence between male and female genders, our qualitative analysis suggests bias in the words used to describe, or in the textual vicinity of mentions of, female characters. There is good reason to assume, therefore, that gender bias (at least of a descriptive nature) and gender disparity (in terms of gendered character mentions) both exist in an important cultural corpus. Books in this corpus are read widely even today, with many continuing to be in print world-wide, and some continue to be part of school and college curricula. We are not disputing the cultural merits of these works; rather, we cite these examples to argue that these books continue to enjoy cultural diffusion. If it is true that cultural diffusion, even implicitly, can affect social norms, practices, and beliefs (Rosenmann, 2016), our findings suggest that, for a class of influential fictional texts, it is worth discussing if gender disparity can be lessened through more judicious selection of texts (e.g., in classroom reading lists).

More recently, textual analysis of novels and literary texts, especially using structural and statistical analyses, has been a major theme in Digital Humanities research. We cite the book by Jockers (2013) on digital methods and literary history as a notable example in this regard. There is also a long line of gender studies literature in the humanities that is complementary to the research herein (Asghari, 2016; Budzise-Weaver, 2016); however, many of those articles use qualitative methods and tend to deeply analyze relatively small corpora, such as the books of a single author (Homans, 1993), or within a narrow period or genre (Fine, 1998), rather than a broad-based computational analysis such as in this work or other similar projects such as the Gender Novels Project. Examples include work by Homans (1993) and Pilcher and Whelehan (2016). These works are necessary for understanding and contextualizing gender in literature and the humanities, and this context should be borne in mind for a causal interpretation of the statistical and computational results that we present in this paper. In contrast to these deep, but qualitative analyses, there have been some computational studies on gender, but they tend to involve either corpora like Wikipedia (Reagle and Rhue, 2011), or multi-modal datasets, such as work by Oh et al. (2020); Hu and Kearney (2021), to only cite a few examples.

3 Materials and Methods

3.1 Data

The raw data for the study was obtained from Project Gutenberg (1971). Named after Johannes Gutenberg, who introduced moving-type book printing in Europe in the medieval era, Project Gutenberg is a volunteer-based project with the stated mission (see Lebert (2009)) to “encourage the creation and distribution of eBooks”. Items in the Project Gutenberg corpus are in the public domain by definition, and all files can be accessed freely using an open format (such as plain text). Although the corpus is technically multi-lingual, books are primarily in the English language. We acquired the original corpus for our study from the website by issuing a web-based query. We obtained a total of 3,036 texts, all in the English language penned by 142 authors (of whom 14 are female), and published between 1800 and 1922. The books published during this period are out of copyright in the United States, and this period also largely coincides with the Victorian era, when gender and class were known to play organizing (and exclusionary) roles (Rose, 2009; Tusan, 2004). We determined the gender of the authors manually by consulting encyclopedic sources such as Wikipedia.

We used a subset of the 3,036 texts for this study by excluding non-fiction, biographical works, public addresses and letters (e.g., the *Lincoln Letters* by Abraham Lincoln). Following this filtering, the final dataset comprised 2,443 texts that cover genres ranging from adventure and science fiction, to mystery and romance, and span the same publication period as the original corpus of 3,036 texts. Nagaraj and Kejriwal (2022) provide a summary of author metadata acquired by us through this manual process. Therein, we also provide details on the query that was used to acquire the original 3,036 texts from Project Gutenberg. Not all works are necessarily in the format of a novel, since the corpus also includes short stories, plays and poems. Technical details on data preprocessing, as well as NLP-based steps such as character extraction and disambiguation, and gender classification are provided in the supplementary material. Additionally, the filtering script, as well as other data and methods used for generating the results, are also included as supplementary material to facilitate reproducibility.

3.2 Descriptive Statistics and Availability of Processed Data for Replication

Additional descriptive statistics of the final data used for evaluating the hypotheses are provided in Table 1. The average number of texts per female author is less than that for male authors, but there are still more than ten texts, on average, per female author. Furthermore, as discussed subsequently, the statistical tests that we conduct when comparing different prevalence measures on male- vs. female-authored texts rely on the unpaired Student's t-test, and does not assume equal variance. Hence, the difference between the total numbers of male- and female-authored texts is not expected to pose a problem for the investigation.

Although only publicly available packages have been used in the NLP-based pipeline described earlier, we recognize that the steps noted above are non-trivial to execute, especially by social scientists and digital humanities scholars who may want to use the data to run their own studies, or replicate the results described subsequently. Hence, we have published the processed version of the data as a peer-reviewed data-in-brief article, with links to a Mendeley repository containing the processed data (Nagaraj and Kejriwal, 2022). Also, code and data to replicate the specific experiments in this article are included as supplementary material.

3.3 Experimental Methodology

All three hypotheses enumerated earlier in the introduction require us to define measures of gender-specific character prevalence. As noted in Section 1, however, measuring prevalence itself depends on defining a methodological procedure for measuring character *mentions*. We begin this section by presenting three such definitions, each of which then becomes associated with its own prevalence measure. The actual prevalence measure depends on the hypothesis and is discussed subsequently. Each of these definitions of mentions can be male-specific or female-specific, and

Table 1: Key descriptive statistics of the filtered and processed Project Gutenberg corpus used in the empirical study in this article.

Total male-authored / female-authored / all texts	2,278 / 165 / 2,443
Mean texts per male / female / each author	17.8 / 11.8 / 17.2
Mean number of ASCII-characters (excluding spaces) / words per text	317,846 / 72,928
Mean characters (of any gender) per text	44.17

is computed per text:

1. **Character Count**: The number of *unique* (male or female) characters extracted from the text. Recall that this was achieved by applying NER on the segmented sentences of the text.
2. **Character Occurrence Count**: The number of times (or 'occurrences') male or female characters are mentioned by name in the text. In other words, we simply remove the uniqueness constraint from the definition of Character Count to compute the Character Occurrence Count. Importantly, however, only named mentions are considered in the definition.
3. **Pronoun Count**: The number of male pronouns (he, him, his) or female pronouns (she, her, hers) present in each text. We do not address gender-neutral pronouns in this article as we could not identify, despite our best efforts, a publicly available or established computational package for distinguishing the uses of pronouns, such as 'they', as referring to multiple individuals or a single individual. However, gender-neutral pronouns referring to individuals that choose their pronouns as 'they', 'their' or 'them' could be considered in future research that draws on more modern texts to replicate the following analyses.

Although not theoretically necessary, in practice, the absolute counts of each of these mentions (per text) increases in the order shown above. For example, the number of pronouns in a text is generally much higher than the number of named character occurrences in the text.

With these notions of gendered mentions in place, we discuss the specific methodology and statistical tests for evaluating each of the three hypotheses below, in turn.

3.3.1 Hypothesis 1

To investigate Hypothesis 1, we define gender-specific character prevalence, for a given choice of mention (e.g., Character Count), as the mean of the mention across all texts. Recall that Hypothesis 1 does not distinguish between the gender of the text's author. We separately plot male- and female-specific character prevalence for all three definitions of mentions. We use bar graphs to plot each such within-group mean, and show standard errors computed at a confidence level of 95% ($\alpha = 0.05$). To assess statistical significance, our working research hypothesis (per prevalence measure) is that male-specific character prevalence is greater than female-specific character prevalence across the corpus of texts under study. The null hypothesis is the converse.

To evaluate the null hypothesis, we conduct the one-sided *paired* Student's t-test, because the gender-specific mentions for both genders can be computed on a per-text level. By virtue of being one-sided and paired, as is appropriate given these assumptions, the power of the test becomes higher than if we were conducting either a two-sided or unpaired test (or both). Neither one of these is appropriate given the experimental design and the stated research hypothesis.

Note that, because there are three different definitions of mentions, there are three 'different' null (and correspondingly, alternative) hypotheses being tested. Although we are separately testing each of the three null hypotheses above, independence should not be assumed between them, since the prevalence measures are likely to be correlated per text. This can lead to problems when combining the results and making claims about the *overall* hypothesis. For example, consider the hypothetical example, where each of the three null hypotheses is individually rejected with a P value of (say) 0.03. In this situation, while each of the null hypotheses can be rejected with maximum Type-I error $\alpha = 0.05$, the hypothesis overall cannot be rejected at the desired α level of 0.05 due to potential correlation between the hypotheses (and their statistics). The reason is that there is greater likelihood of rejecting a null hypothesis (and committing a Type-I error) by chance if highly correlated statistical hypotheses are tested *multiple* times.

This issue of *multiple comparisons* is well understood in the statistical literature and typically addressed by using the *Bonferroni correction* (Bonferroni, 1936; Napierala, 2012), which tests each individual hypothesis at a more stringent confidence level. For the sake of clarity, we always report the 'independent' P values (i.e., before applying the correction), but in discussing the result, we refer to the Bonferroni correction and determine a result to be significant only if the effective α (i.e., after applying the correction) is at most 0.05. Finally, we also tabulate additional statistical details, including sample size, degrees of freedom, the appropriate means (with 95% confidence intervals) and the one-sided P values using the Student's t-test, as discussed above. Key results are reported in the main text, with full details on the significance testing reproduced in the supplementary material, along with code and series data to facilitate further analysis.

3.3.2 Hypothesis 2

Investigating Hypothesis 2 requires data on authors' genders, since the text author's gender is used as the stratifying variable for the analysis. Because there are 142 unique authors in our corpus, we manually tagged their gender using both their names, and public resources like Wikipedia. Of these 142 authors, only 14 were found to be female. Some other descriptive statistics on the texts were tabulated earlier in Table 1.

The statistical procedure used for this hypothesis also relies on the one-sided Student's t-test. However, unlike Hypothesis 1, for each of the three definitions of mentions, the definition of female-specific character prevalence that we adopt is the *proportion* of female-specific character mentions in a text to the total number of character mentions in the same text. This yields a real value ranging between 0 and 1, regardless of the mention definition used. With this definition of prevalence in place, the research hypothesis underlying Hypothesis 2 is that female-specific character prevalence in female-authored texts is greater than that in male-authored texts. The null hypothesis is the converse. Although the proportion is computed per text, the nature of the hypothesis above indicates that an unpaired test is suitable (since female-authored texts are necessarily disjoint from male-authored texts). This is another important difference (besides the manner in which prevalence is defined) between the testing for Hypothesis 1 and 2. Specifically, we use the one-sided *unpaired* Student's t-test for assessing significance. Furthermore, we do not assume equal variance when conducting the test.

A second possibility for statistical significance testing that could be considered for Hypothesis 2 is to assess whether female-specific character prevalence equals 0.5 (indicating parity with male-specific character prevalence). The test for proportion could be applied to investigate this claim, but for the claim above, the proportion could only be properly tested on a per-text level. Comparing male-authored texts as a whole, and female-authored-texts as a whole, then becomes more complex and indirect. We leave such a comparison for future work, but the workbooks used for investigating each hypothesis are provided as supplementary material to facilitate such a comparison without having to extract the data and statistics from scratch.

A commonality between the statistical procedures for both hypotheses is that we again use the Bonferroni correction to assess whether potential correlation of the three tests (because the definitions of the mentions are highly correlated) leads to a significance result that provides overall support for Hypothesis 2. Similar to Hypothesis 1, we tabulate precise statistical information, such as the P values, means with 95% confidence intervals, and other such supporting information. Key results are reported in the main text, with complete details in the supplementary information.

3.3.3 Hypothesis 3

To investigate Hypothesis 3, we need to extract the publication year of each text. Unfortunately, this metadata is not naturally available in Project Gutenberg, although it is mentioned at the beginning of the plain-text file in a subset of texts. We wrote a customized rule-based script to detect and extract either the publication or copyright year (whichever one occurred first), when it was present in the first 20 lines of the plain-text file. Because these two years tended to be largely the same, or within a year of each other, we decided to use it as a single variable. A small discrepancy between Hypothesis 3 and the other two hypotheses is that, due to a small coding error, we only extracted the year if it fell within the range 1800-1919 (inclusive). Even with this conservative extraction technique, in total, we were able to extract 610 male-authored and 47 female-authored texts, each associated with a single year that we henceforth refer to as the publication year. Although the total number of texts being used for Hypothesis 3 is much smaller than the total number of texts used for Hypothesis 1 and 2, the proportion of female-authored texts for which we were able to extract the year (28.48%) is qualitatively similar to the proportion of male-authored texts (26.78%) for which the extraction succeeded.

Next, because our goal is to investigate whether there is a change over time in female-specific character prevalence, we used the same definition of prevalence as in Hypothesis 2 (i.e., the proportion of female-specific character mentions to total character mentions on a per-text level). To get an interpretable trend result, we also 'normalize' each year by converting it to an index that equals the year minus 1800. Hence, 1800 is set to 0, 1801 to 1, and so on. We treat this 'year index' as our independent variable for the purposes of plotting a scatter-plot of the prevalence measure, and fitting a trend-line based on least-squares linear regression. A similar number of plots and results are reported as for Hypothesis 2 i.e., for both male- and female-authored texts, we plot female-specific character prevalence over time for each of the three definitions of mentions. We evaluate whether the coefficient of the independent variable is positive and significant. Doing so enables us to assess whether time has a positive effect on the proportion of female-specific character mentions. It also allows us to compare the coefficients across female- and male-authored texts in equivalent experimental settings. Where applicable, we will also apply the Bonferroni correction for evaluating significance in the multiple comparisons setting. In the supplementary material, we provide additional quantitative results for each regression.

4 Results

In this section, we describe our key findings. For ease of exposition, we present the findings pertinent to each hypothesis in turn.

4.1 Hypothesis 1

Hypothesis 1 was the claim that female-specific character prevalence is lower than male-specific character prevalence in the Project Gutenberg corpus under study. In the previous section, we also presented three different ways in which gender-specific character prevalence can be computed. Using these three different measures, we plot in Figure 1 the differences between male and female character prevalence as bar graphs. Additional statistical details are enumerated in Table 2, with a more complete profile of statistical significance testing in the supplementary material.

The figure yields some important insights, the most important being that, no matter the

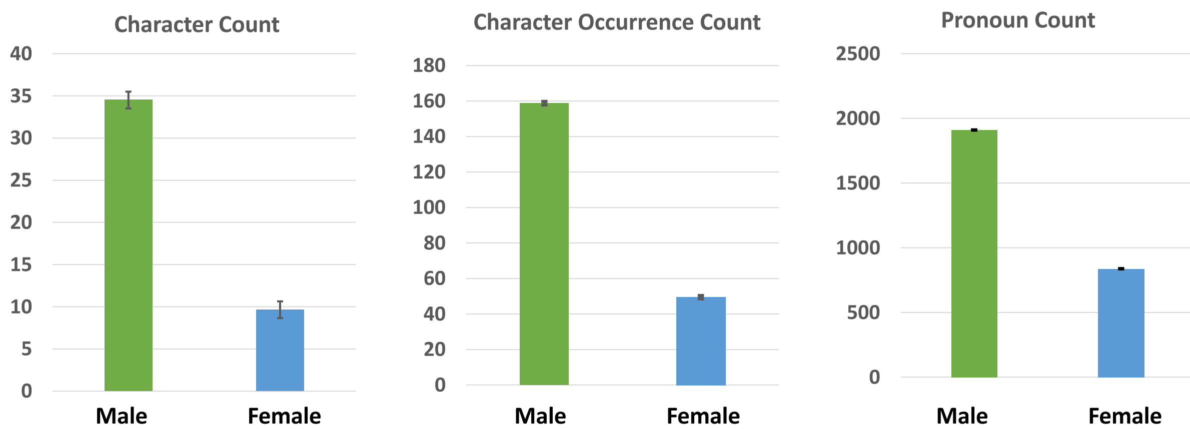

Figure 1: Bar graphs, with 95% confidence intervals, characterizing differences between male-specific and female-specific character prevalence using the three measures defined in Section 3.3.

Table 2: Statistical measures and summary statistics in support of Hypothesis 1. In all cases, the sample size (n) was 2,443 texts, and for computing t-test statistics, the degrees of freedom equals $n - 1 = 2{,}442$. Bonferroni correction has not been applied to the P values shown. All results are reported to two significant digits.

Statistic	**Character Count**	**Character Occurrence Count**	**Pronoun Count**
Mean (with 95% Confidence Interval) for male characters	34.51 (32.85 to 36.17)	158.82 (151.09 to 166.55)	1910.50 (1839.21 to 1981.79)
Mean (with 95% Confidence Interval) for female characters	9.65 (9.13 to 10.18)	49.51 (45.98 to 53.04)	838.14 (791.29 to 884.99)
One-sided P value (paired Student's t-test)	9e-252	4.3e-187	7.5e-283

definition of mentions used for quantifying character prevalence, female-specific character prevalence in the texts under study is significantly lower than male-specific character prevalence. While the magnitude of the difference is striking in all cases, it can depend on the specific measure employed. For instance, greater relative differences are observed for the Character Count and Character Occurrence Count measures, compared to the Pronoun Count measure. However, the confidence intervals are much smaller for the Character Occurrence Count and Pronoun Count measures, compared with the Character Count measure. Regardless, in all cases, the error bars suggest significant differences between female-specific and male-specific character prevalence across the corpus.

In testing for significance by using the one-sided independent Student's t-test, we find that the difference between the means of male-specific and female-specific character prevalence is highly significant with a reported P value (for each of the three measures) that is near zero. Even after applying the Bonferroni correction, therefore, each of the differences remains significant. The overall statistics suggest that there is a significant difference between male-specific and

Table 3: Mean proportion (represented as percentage) of female-specific characters to all characters, measured using the three definitions of character mentions, across male-authored and female-authored texts. The one-sided P value, calculated using the unpaired Student's t-test, tests the null hypothesis (discussed earlier as part of experimental methodology) that this proportion in female-authored texts is greater than in male-authored texts. Bonferroni correction has not been applied to the P values shown. Results are reported to two significant digits.

Statistic	**Character Count**	**Character Occurrence Count**	**Pronoun Count**
Mean proportion (with 95% Confidence Interval) in male-authored texts	21.38% (20.82% to 21.95%)	21.79% (20.98% to 22.61%)	25.03% (24.36% to 25.70%)
Mean proportion (with 95% Confidence Interval) in female-authored texts	36.25% (34.08% to 38.42%)	38.23% (35.15% to 41.31%)	46.42% (43.48% to 49.35%)
One-sided P value (unpaired Student's t-test)	1.77e-28	5.62e-20	4.70e-31

female-specific character prevalence across the corpus.

4.2 Hypothesis 2

Hypothesis 2 was the claim that the difference between male-specific and female-specific character prevalence, found to be significant and large in the previous investigation, could decline if we stratify by the gender of the authors. As a first step, we plot in Figure 2, bar graphs that are the equivalent of those presented earlier for Hypothesis 1 but separately for male-authored and female-authored texts. The figure suggests that, while the mean of male-specific character mentions does not change much (for any of the three measures) between male- and female-authored texts, the mean of female-specific character mentions rises significantly.

Table 3 provides precise quantitative information putting the statistical hypothesis, discussed earlier in Section 3.3 to the test. We find that, even after applying the Bonferroni correction (which would effectively triple each of the reported P values), female-specific character prevalence in female-authored texts is significantly higher than in male-authored texts, regardless of the mention definition used.

In the supplementary material, we provide additional statistical details on the means visualized in Figure 2, again using the Student's t-test without assuming equal variance. Interestingly, female-specific character prevalence (using the Hypothesis 1 definition of prevalence as means) in female-authored texts was found to be significantly lower than male-specific character prevalence, although the magnitude of the difference (18.75) was smaller than in male-authored texts (25.3). The significance of the effect (P value) was also diminished by more than two orders of magnitude, and in the case of the Pronoun Count measure, was only moderately significant. The most important takeaway is that, as the previous results indicated, the proportion of female-specific character mentions, as compared to all character mentions (across any of the three measures) was found to be significantly higher in female-authored texts than in male-authored texts. The general increase in character mentions in female-authored texts may be a consequence

Figure 2: Bar graphs, with 95% confidence intervals, of the means of male-specific and female-specific character mentions after stratifying by the gender of the text's author.

of female authors writing more character-driven stories, or emphasizing characters more in their stories, but we leave an investigation of this cause for other research.

In summary, there is good empirical support in favor of Hypothesis 2: based on all three gender-specific character prevalence measures, stratifying by the gender of the author shows that the differences noted in the Hypothesis 1 results are diminished, with female character mentions (as a proportion of all character mentions) being significantly more prevalent in female-authored texts, compared to male-authored texts. A detailed tabulation of significance results and statistics is also provided in the supplementary material.

4.3 Hypothesis 3

Hypothesis 3 was the claim that female-specific character prevalence changes significantly in the corpus under study between 1800 to 1920. The definition of prevalence adopted was the same as the (proportion) definition adopted for Hypothesis 2. Figure 3 demonstrates the results by plotting the proportion of female-specific character mentions as a ratio of the total number of character mentions, using all three prevalence measures. Unlike the other two hypotheses, we do not see empirical support for Hypothesis 3 in Figure 3. Particularly, the null hypothesis that the coefficient in each regression in Figure 3 is different from zero cannot be rejected even at the 90% confidence level. Clearly, applying the Bonferroni correction would make the results even weaker. Equivalently, the F-statistic for the regression has a high P value. Full quantitative details of the regressions are provided in the supplementary material. The results are consistent with those reported by Digital Humanities Lab, MIT (2022), although they did not conduct significance testing or regression-based trend analysis for their findings.

It is theoretically possible that, with additional independent factors or using more data, a positive and significant relationship may emerge, but we leave for future work to discover such factors. In Section 5, we cite some recent work suggesting that finding such a positive association, at least with sufficient statistical strength, may be unlikely. Also, we plot the regression by doing log transformations on both the x and y-axis variables in turn, but the conclusions did not change appreciably (hence, these auxiliary regressions are not reported herein).

Although the texts used for Hypothesis 3 are only a subset of the texts used for investigating Hypotheses 1 and 2 (as explained earlier in Section 3.3), the intercepts of the regressions were found to be significantly different than 0 at the 99% confidence level or above, and we also note that the intercepts for the results corresponding to female-authored texts are considerably higher (although still below 50%) than for the male-authored texts. Hence, these results do provide an additional robustness check on the results shown earlier for Hypothesis 2.

5 Discussion

Our experiments indicate the severe imbalance of gender-specific character prevalence in the sample (available in the Project Gutenberg corpus) of books written during the pre-modern period. In particular, the experiments illustrate robust support for both Hypothesis 1 and 2. The former is supported by the fact that, when gender-specific character prevalence is compared using the means of each of the three measures, on average, three out of four character mentions (or character occurrences, including using pronouns) in the books are found to be male. The latter is supported by the fact that the proportion of female-specific character mentions is significantly higher in female-authored books than in male-authored books. Although not statistically studied in this article, there is significant disparity also in the numbers of male and female authors, the

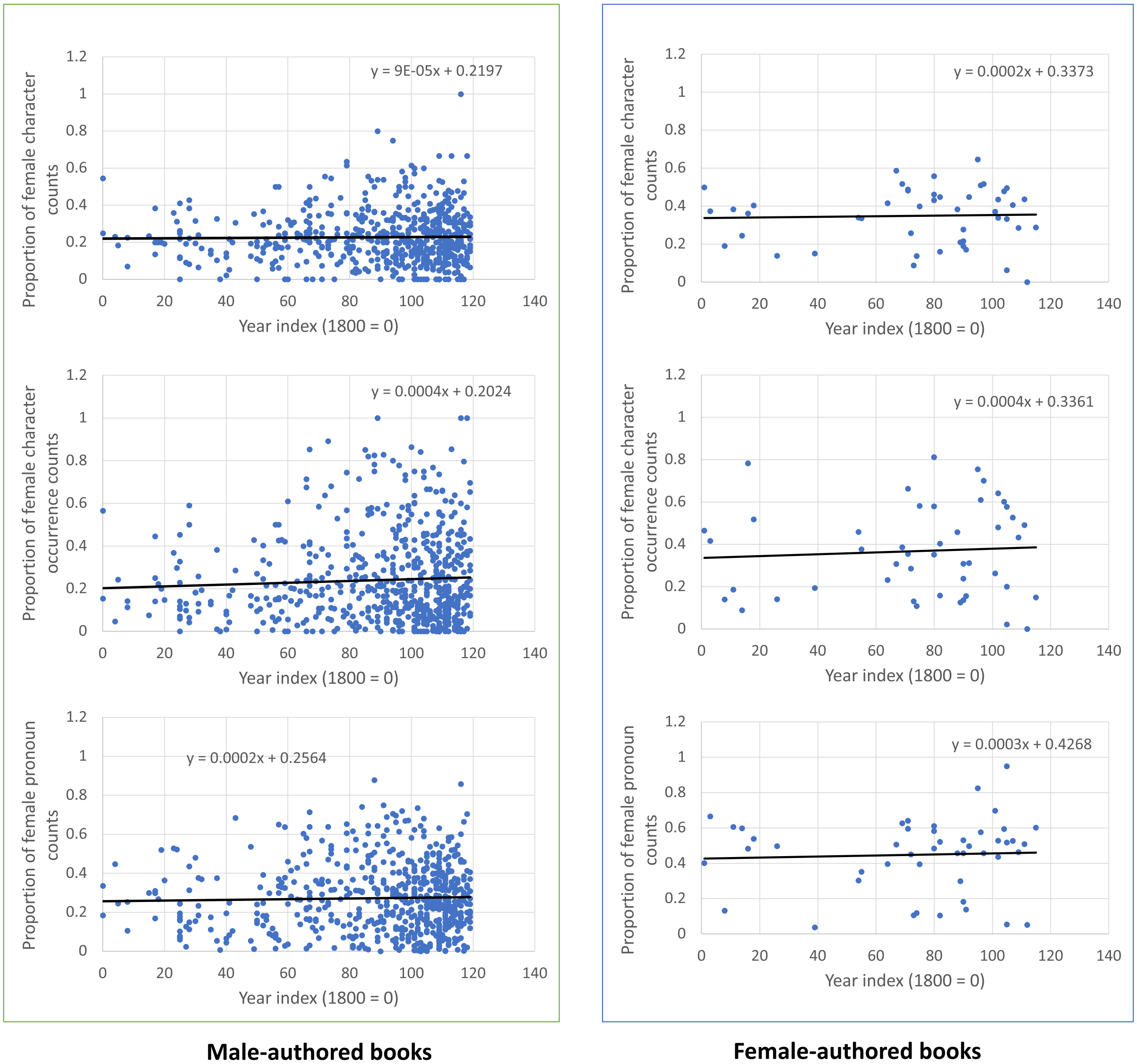

Figure 3: Scatter plots, with trendlines (derived using standard least-squares linear regression), in support of Hypothesis 3, after stratifying by the gender of the text's author. The horizontal axis denotes the year index (starting from 1800, which is assigned an index of 0) when the book was published, while the vertical axis represents the proportion of female-specific character mentions to the total, as measured using the three different definitions of mentions earlier described in Section 3.3.

latter of which is higher than the former by more than an order of magnitude. Hence, one remedy that is suggested (but cannot be proven or causally situated) by the study is to increase representation of gender among authors to achieve better parity. Recent studies have highlighted the importance of representation and diversity for increasing participation and equity in various fields (although a clear study in art and literature has been lacking, possibly due to lack of data availability) (Hoekstra, 2010; Nielsen et al., 2018), and along various dimensions, including both race and gender (Stathoulopoulos and Mateos-Garcia, 2019; Richard, 2000; Greider et al., 2019;

Phillips and Malone, 2014).

The overall results are also consistent with some analogous findings in modern film. For example, as shown by some recent authors such as Yang et al. (2020), until very recently, meaningful representation of women in movies has been low using a number of metrics. In that paper, the authors used a prevalence measure called the *female cast ratio*, which is similar to the proportional measure of character prevalence we used in both Hypothesis 2 and 3. Comparing it to another commonly used metric (for film) based on the Bechdel test (Agarwal et al., 2015), the authors found that female representation is generally very low compared to male representation along various dimensions. Similar to our methodology in Hypothesis 3, Yang et al. (2020) also relied on a model based largely on linear regression to identify predictors of female representation. However, they considered a variety of predictors that were available in their movie database, such as the revenue of the movie, and the popularity of the movie. Unfortunately, such predictors are not straightforward to derive for the texts in our corpus, but it may be possible for future research to do so by combining a variety of historical data sources.

More specifically, there is also some support for the modern equivalent of Hypothesis 2 in the study by Yang et al. (2020). Their study showed that, in movies involving female screenwriters and filmmakers, female representation and character development on screen is higher. This is analogous to our Hypothesis 2 finding. This may also provide a partial explanation for Hypothesis 3, since over the period under study, female representation in fictional literature seems to have stayed largely stagnant. The number of female authors was also considerably lower than the number of male authors. Hence, female-specific character prevalence also stayed largely flat over this period. We note, however, that the experiments do not necessarily support a causal claim.

Even in other cultural areas (such as art), as well as fields like archaeology, a gap in female representation is evident e.g., in works by Nixon (1994) and Miller (2016). An equivalent study based on modern literature is not feasible at this time, since the text files of major literary works published in the last few decades (mostly after the post-war period) are not in the public domain and available for study, unlike the texts in the Project Gutenberg corpus.

Finally, our results are largely consistent with those of the Gender Novels Project (Digital Humanities Lab, MIT, 2022), at least when considering the Pronoun Count mention definition. In that project, the investigators also found a high difference between male- and female-specific character prevalence in the overall corpus, although it was not tested for statistical significance. Stratifying by the author's gender did make a (statistically significant) difference, and is consistent with the reported results for Hypothesis 2. Our results for Hypothesis 3 are also confirmed by their independent analysis: although they did not do a trendline or regression-based analysis, they did not see a difference when plotting bar-graph results by publication date or location.

An important limitation of this study, described and contextualized in more detail in the next section, is that we only study the disparity between male and female characters. Unfortunately, the disparity of non-binary and trans characters compared to traditional genders could not be accurately studied due to a lack of computational tools for extracting characters whose genders do not fall in the dichotomous categories of male and female. We believe that this highlights a pressing need to develop such tools. Within the supplementary material, we provide a more detailed description of the limitations of the study and potential ethical issues.

6 Future Work and Conclusion

In this article, we defined and measured the differences between male character prevalence and female character prevalence using three robust measures of character mentions, on a corpus of pre-modern, copyright-expired literary texts from the Project Gutenberg English-language corpus. Using computationally replicable methodologies relying on modern natural language processing tools, we found that female-specific character prevalence is significantly lower than that of male character prevalence, although the difference declines (while still being significant) when stratifying by the gender of the author. We also found that, when stratifying by the author's gender, the proportion of female-specific character mentions is significantly higher in female-authored texts than in male-authored texts. Unfortunately, this proportion (which is generally well beneath 50%) has increased very little over a 120-year timespan within our sample, regardless of whether the author was male or female. Recent results, discussed further in Sections 2 and 5, are consistent with this finding. More broadly, we hope that our findings serve as a case study illustrating the promise of using open-source tools and data, in conjunction with careful quality control and statistical analysis, to objectively explore gender-specific issues of socio-cultural inequality. There are also many promising directions for future research, including consideration of different time periods, use of genres as stratifying variables, and study of texts in different languages.

Supplementary Material

The supplementary material contains details on: data preprocessing, character extraction and gender classification; additional quantitative details, including complete statistical significance results, for Hypotheses 1 and 2; quantitative linear regression results (including supporting statistics such as the analysis of variance); methodological details and results for the secondary analysis noted in Section 1.1 wherein we seek to use computational techniques from NLP to qualitatively assess the *kinds* of words associated with male and female character occurrences; and, a detailed description of some limitations of the study that were briefly discussed in the main text. Additionally, code, data and workbooks for replicating the analyses in this paper are also provided separately as supplementary material.

References

Adams JE (2012). *A History of Victorian Literature*, volume 10. John Wiley & Sons.

Agarwal A, Zheng J, Kamath S, Balasubramanian S, Dey SA (2015). Key female characters in film have more to talk about besides men: Automating the Bechdel test. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 830–840.

Asghari F (2016). Methodological considerations in gender studies. *Interdisciplinary Studies in the Humanities*, 7(4): 105–127.

Belkhyr S (2013). Disney animation: Global diffusion and local appropriation of culture. *Études Caribéennes*, (22).

Bonferroni C (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8: 3–62.

Budzise-Weaver T (2016). Developing a qualitative coding analysis of visual artwork for humanities research. *DHQ: Digital Humanities Quarterly*, 10(4): 33–45.

Burke RJ, Mattis MC (2013). *Women on Corporate Boards of Directors: International Challenges and Opportunities*, volume 14. Springer Science & Business Media.

Burley T, Humble L, Sleeper C, Sticha A, Chesler A, Regan P, et al. (2020). Nlp workflows for computational social science: Understanding triggers of state-led mass killings. In: *Practice and Experience in Advanced Research Computing*, 152–159.

Cabrera D, Roy D, Chisolm MS (2018). Social media scholarship and alternative metrics for academic promotion and tenure. *Journal of the American College of Radiology*, 15(1): 135–141. https://doi.org/10.1016/j.jacr.2017.09.012

Devlin J, Chang MW, Lee K, Toutanova K (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint: https://arxiv.org/abs/1810.04805

Digital Humanities Lab, MIT (2022). The gender novels project. http://gendernovels.digitalhumanitiesmit.org/info/gender_novels_overview. Accessed: 2022-09-29.

Fine L (1998). Gender conflicts and their "dark" projections in coming of age white female southern novels. *Southern Quarterly*, 36(4): 121.

Glavaš G, Nanni F, Ponzetto SP (2017). Cross-lingual classification of topics in political texts. *Association for Computational Linguistics (ACL)*.

Greider CW, Sheltzer JM, Cantalupo NC, Copeland WB, Dasgupta N, Hopkins N, et al. (2019). Increasing gender diversity in the stem research workforce. *Science*, 366(6466): 692–695. https://doi.org/10.1126/science.aaz0649

Han J, Wang H (2021). Transformer based network for open information extraction. *Engineering Applications of Artificial Intelligence*, 102: 104262. https://doi.org/10.1016/j.engappai.2021.104262

Hoekstra V (2010). Increasing the gender diversity of high courts: A comparative view. *Politics & Gender*, 6(3): 474–484. https://doi.org/10.1017/S1743923X10000243

Homans M (1993). Dinah's blush, maggie's arm: Class, gender, and sexuality in george eliot's early novels. *Victorian Studies*, 36(2): 155–178.

Hovy D, Volkova S, Bamman D, Jurgens D, O'Connor B, Tsur O, et al. (2017). Proceedings of the second workshop on nlp and computational social science. In: *Proceedings of the Second Workshop on NLP and Computational Social Science*.

Hu L, Kearney MW (2021). Gendered tweets: Computational text analysis of gender differences in political discussion on twitter. *Journal of Language and Social Psychology*, 40(4): 482–503. https://doi.org/10.1177/0261927X20969752

Hu M, Kejriwal M (2022). Measuring spatio-textual affinities in twitter between two urban metropolises. *Journal of Computational Social Science*, 5(1), 227–252.

Jarynowski A, Paradowski MB, Buda A (2019). Modelling communities and populations: An introduction to computational social science. *Studia Metodologiczne*, 39: 123–152.

Jockers ML (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

John J (2016). *The Oxford Handbook of Victorian Literary Culture*. Oxford University Press.

Jordan CE, Clark SJ, Waldron MA (2007). Gender bias and compensation in the executive suite of the fortune 100. *Journal of Organizational Culture, Communications and Conflict*, 11(1): 19.

Katz E (1999). Theorizing diffusion: Tarde and sorokin revisited. *The Annals of the American Academy of Political and Social Science*, 566(1): 144–155. https://doi.org/10.1177/000271629956600112

Keuschnigg M, Lovsjö N, Hedström P (2018). Analytical sociology and computational social science. *Journal of Computational Social Science*, 1(1): 3–14. https://doi.org/10.1007/s42001-017-0006-5

Lebert M (2009). A short history of ebooks. http://www.gutenberg.org/files/29801/29801-0.txt. Accessed: 2023-03-14.

Legal Information Institute, Cornell Law School (2020). *Gender Bias*. https://www.law.cornell.edu/wex/gender_bias. Accessed: 2022-09-29.

Liu Y (2019). Fine-tune bert for extractive summarization. arXiv preprint: https://arxiv.org/abs/1903.10318

Mason W, Vaughan JW, Wallach H (2014). Computational social science and social computing.

May Alcott L (1868). *Little Women. Project Gutenberg*.

Miller DL (2016). Gender and the artist archetype: Understanding gender inequality in artistic careers. *Sociology Compass*, 10(2): 119–131. https://doi.org/10.1111/soc4.12350

Milli S, Bamman D (2016). Beyond canonical texts: A computational analysis of fanfiction. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2048–2053.

Montasseri Z, Khaghaninejad MS, Moloodi A (2020). Gender representation in american movies: A corpus-based analysis. *The International Journal of Humanities*, 27(4): 42–53.

Montjoye YAd, Quoidbach J, Robic F, Pentland AS (2013). Predicting personality using novel mobile phone-based metrics. In: *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 48–55. Springer.

Nadeau D, Sekine S (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1): 3–26. https://doi.org/10.1075/li.30.1.03nad

Nagaraj A, Kejriwal M (2022). Dataset for studying gender disparity in english literary texts. *Data in Brief*, 41: 107905. https://doi.org/10.1016/j.dib.2022.107905

Napierala MA (2012). What is the bonferroni correction? *Aaos Now*, 40–41.

Naseem U, Razzak I, Musial K, Imran M (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113: 58–69. https://doi.org/10.1016/j.future.2020.06.050

Nath R, Murthy N (2004). A study of the relationship between internet diffusion and culture. *Journal of International Information Management*, 13(2): 5.

Nielsen MW, Bloch CW, Schiebinger L (2018). Making gender diversity work for scientific discovery and innovation. *Nature Human Behaviour*, 2(10): 726–734. https://doi.org/10.1038/s41562-018-0433-1

Nixon L (1994). Gender bias in archaeology. In: *Women in Ancient Societies*, 1–23. Springer.

O'Connor SD (1996). History of the women's suffrage movement. *Vand. L. Rev.*, 49: 657.

Oh D, Dotsch R, Porter J, Todorov A (2020). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology. General*, 149(2): 323. https://doi.org/10.1037/xge0000638

Peters K, Chen Y, Kaplan AM, Ognibeni B, Pauwels K (2013). Social media metrics–a framework and guidelines for managing social media. *Journal of Interactive Marketing*, 27(4): 281–298. https://doi.org/10.1016/j.intmar.2013.09.007

Phillips JM, Malone B (2014). Increasing racial/ethnic diversity in nursing to reduce health disparities and achieve health equity. *Public Health Reports*, 129(1_suppl2): 45–50. https://doi.org/10.1177/00333549141291S209

Pilcher J, Whelehan I (2016). *Key Concepts in Gender Studies*. Sage.

Prabhumoye S, Choudhary S, Spiliopoulou E, Bogart C, Rose C, Black AW (2017). Linguistic markers of influence in informal interactions. In: *Proceedings of the Second Workshop on NLP and Computational Social Science*, 53–62.

Project Gutenberg (1971). Project gutenberg. https://www.gutenberg.org/. Accessed: 2022-09-29.

Reagle J, Rhue L (2011). Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5: 21.

Reddy S, Chen D, Manning CD (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7: 249–266. https://doi.org/10.1162/tacl_a_00266

Richard OC (2000). Racial diversity, business strategy, and firm performance: A resource-based view. *Academy of Management Journal*, 43(2): 164–177. https://doi.org/10.2307/1556374

Rochon TR (2000). *Culture Moves: Ideas, Activism, and Changing Values*. Princeton University Press.

Rodriguez MY, Storer H (2020). A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data. *Journal of Technology in Human Services*, 38(1): 54–86. https://doi.org/10.1080/15228835.2019.1616350

Rose A (2009). *Gender and Victorian Reform*. Cambridge Scholars Publishing.

Rosenmann A (2016). Alignment with globalized western culture: Between inclusionary values and an exclusionary social identity. *European Journal of Social Psychology*, 46(1): 26–43. https://doi.org/10.1002/ejsp.2130

Setzler M (2019). Measuring bias against female political leadership. *Politics & Gender*, 15(4): 695–721. https://doi.org/10.1017/S1743923X18000430

Siblini W, Pasqual C, Lavielle A, Cauchois C (2019). Multilingual question answering from formatted text applied to conversational agents. arXiv preprint: https://arxiv.org/abs/1910.04659

Stathoulopoulos K, Mateos-Garcia JC (2019). Gender diversity in ai research. https://media.nesta.org.uk/documents/Gender_Diversity_in_AI_Research.pdf. Available at SSRN 3428240.

Stevenson RL (1883). *Treasure Island*. Cassell & Co.

Tusan ME (2004). Performing work: Gender, class, and the printing trade in victorian britain. *Journal of Women's History*, 16(1): 103–126. https://doi.org/10.1353/jowh.2004.0037

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. (2020). Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wood-Doughty Z, Smith M, Broniatowski D, Dredze M (2017). How does twitter user behavior vary across demographic groups? In: *Proceedings of the Second Workshop on NLP and Computational Social Science*, 83–89.

Yang L, Xu Z, Luo J (2020). Measuring female representation and impact in films over time. *ACM Transactions on Data Science*, 1(4): 1–14. https://doi.org/10.1145/3411213