Humanizing Computational Literature Analysis Through Art-Based Visualizations

Alexandria Leto

University of Denver

Follow this and additional works at: https://digitalcommons.du.edu/etd

Part of the Data Science Commons, and the Other Computer Sciences Commons

Recommended Citation

https://digitalcommons.du.edu/etd/2130

This Thesis is brought to you for free and open access by the Graduate Studies at Digital Commons @ DU. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ DU. For more information, please contact jennifer.cox@du.edu,dig-commons@du.edu.
Abstract
Inequalities in gender representation and characterization in fictional works are issues that have long been discussed by social scientists. This work addresses these inequalities with two interrelated components. First, it contributes a sentiment and word frequency analysis task focused on gender-specific nouns and pronouns in 15,000 fictional works taken from the online library, Project Gutenberg. This analysis allows for both quantifying and offering further insight on the nature of this disparity in gender representation. Then, the outcomes of the analysis are harnessed to explore novel data visualization formats using computational and studio art techniques. Our results call attention to the need for several identified trends in this data set that require more cross-disciplinary investigation to unpack. Further, our combined approach demonstrate both the need for, and possible methods of, communicating data that warrants social change in an evocative and powerful way.

Document Type
Thesis

Degree Name
M.S.

Department
Computer Science

First Advisor
Scott Leutenegger

Second Advisor
Sarah Gjertson

Third Advisor
Kerstin Haring

Keywords
Computational literature analysis, Data visualization, Natural language processing, Sentiment analysis

Subject Categories
Computer Sciences | Data Science | Other Computer Sciences

Publication Statement
Copyright is held by the author. User is responsible for all copyright compliance.

This thesis is available at Digital Commons @ DU: https://digitalcommons.du.edu/etd/2130
Humanizing Computational Literature Analysis Through Art-Based Visualizations

A Thesis
Presented to
the Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science
University of Denver

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by
Alexandria Leto
November 2022
Advisor: Dr. Scott Leutenegger
Abstract

Inequalities in gender representation and characterization in fictional works are issues that have long been discussed by social scientists. This work addresses these inequalities with two interrelated components. First, it contributes a sentiment and word frequency analysis task focused on gender-specific nouns and pronouns in 15,000 fictional works taken from the online library, Project Gutenberg. This analysis allows for both quantifying and offering further insight on the nature of this disparity in gender representation. Then, the outcomes of the analysis are harnessed to explore novel data visualization formats using computational and studio art techniques. Our results call attention to the need for several identified trends in this data set that require more cross-disciplinary investigation to unpack. Further, our combined approach demonstrate both the need for, and possible methods of, communicating data that warrants social change in an evocative and powerful way.
Acknowledgements

I would like to extend my warmest thanks to my advisor, Dr. Scott Leutenegger, for guiding me through this journey. Not only was he exceedingly supportive, but he offered indispensable knowledge and insight through each step of the research process. He made it possible for me to pursue only those endeavors which I was most passionate about. I would also like to thank Professor Sarah Gjertson for her crucial contributions to my creative process and whose help made this work truly cross-disciplinary. I am also grateful to Dr. Kerstin Haring, whose enthusiasm and help were extremely encouraging.

I would be remiss in not mentioning my parents and sisters, who were always willing to lend a listening ear. I’d like to also thank my significant other, Alec, whose patience and kindness are unmatched. Finally, I’d like to extend my love to my dog, Momo, for her infinite capacity for emotional support.
Table of Contents

List of Figures ......................................................... vi

1. Introduction ...................................................... 1

2. Related Work ..................................................... 4
   2.1. Natural Language Processing and Sentiment Analysis ........ 4
   2.2. Computational Text Analysis of the Female Role in Fiction ... 6
   2.3. Data Visualization ........................................ 7

3. NLP Analysis ..................................................... 11
   3.1. The Data Set ............................................. 11
   3.2. Data Cleaning ............................................ 11
   3.3. Hand Tagged Data ....................................... 12
   3.4. Sentiment and Frequency Analysis ......................... 13
   3.5. Emotion Analysis ....................................... 14
   3.6. Adjective and Adverb Frequency Analysis ................. 15

4. Results .............................................................. 17
   4.1. Gendered Sentence Frequency ............................... 17
   4.2. Author Demographics ..................................... 18
   4.3. Sentiment Analysis ....................................... 20
   4.4. Emotion Analysis ......................................... 22
4.5. Frequent Adjective and Adverb Analysis .................................. 23

5. Art Based Visualization .......................................................... 25
  5.1. "Shift" Frequency Visualization ........................................... 25
  5.2. X-icon Representation ....................................................... 28
  5.3. Plotter Art ................................................................. 30
  5.4. Dyed Linen ................................................................. 35
  5.5. Kaleidoscope ............................................................... 36
  5.6. Cocktail Dress ............................................................. 38

6. Limitations ................................................................. 40

7. Future Work ............................................................... 41

8. Conclusion ................................................................. 44

Bibliography ................................................................. 46

A. Appendix - Frequent Adjective and Adverb Analysis Results . . . 50

B. Appendix - "Shift" Frequency Visualizations ............................ 69

C. Appendix - 100 Most Frequently Downloaded Gutenberg ebooks in
   the week of October 3, 2022 ................................................. 80
List of Figures

3. NLP Analysis

3.1. The majority of fictional books analyzed had authors with birth years between 1800 and 1900. ................................................. 12
3.2. A lexicon of gender-specific nouns and pronouns was used to identify male versus female sentences. ............................................. 16

4. Results

4.1. There was a significant drop in the number of female versus male sentences with author birth years between 1850 and 1940. ................. 18
4.2. The number of male authors whose work appears in the data set is significantly greater than the number of female authors. ...................... 19
4.3. There was a sharp drop in the percentage of books by female authors beginning in 1900. ......................................................... 20
4.4. Both female and male authors with birth years between 1850 and 1940 wrote about fewer female characters in their work. ....................... 21
4.5. There was a decrease in the positive sentiment in both male and female sentences between 1850 and 1940. ................................. 22
4.6. There was an increase in the negative sentiment in both male and female sentences between 1850 and 1940. ................................. 23

5. Art Based Visualization.
5.1. "Shift" frequency visualization displaying results from gender frequency analysis of *The Great Gatsby* ........................................... 26
5.2. "Shift" frequency visualization displaying results from gender frequency analysis of *Pride and Prejudice* ........................................... 27
5.3. "Shift" frequency visualization displaying results from gender frequency analysis of *Moby Dick* ....................................................... 28
5.4. X-icon representation of *The Great Gatsby* ........................................... 29
5.5. X-icon visualization for one hundred Gutenberg books .......................... 30
5.6. An AxiDraw pen plotter and paint palette were used to create visualizations with paint. ....................................................... 31
5.7. Acrylic on paper executed by a pen plotter aimed at expressing a surprised response to the sharp decrease in female representation within the data set. ....................................................... 32
5.8. Acrylic on paper executed by a pen plotter aimed at expressing surprised response to decrease in both female representation and positive sentiment. 33
5.9. Acrylic on paper executed by pen plotter comparing female (left) to male (right) representation within the data set. .............................. 34
5.10. Acrylic on linen executed by pen plotter comparing female (left) to male (right) representation within the data set. .............................. 35
5.11. Scraps of linen matched to ten-year buckets of books whose dyed portion is proportional to the percentage of gendered sentences that were labelled female. ....................................................... 36
5.12. Images whose fragments were used to create the inner (left) and outer (right) rings of the kaleidoscope visualization. .............................. 37
5.13. Stills from kaleidoscope visualization depicting maximum female representation (left) and minimum female representation (right). .............. 38
5.14. List of words that appeared on the list of 1,000 most frequent adjectives and adverbs in female sentences, but did not appear in the list of 1,000 most frequent adjectives and adverbs in male sentences for the 1880-1890 author year bucket and their attached frequencies. 39

5.15. Front and back view of cocktail dress visualization. 39

B. Appendix - "Shift" Frequency Visualizations

B.1. Walden, and On The Duty Of Civil Disobedience by Henry David Thoreau (1798) 69

B.2. Pride and Prejudice by Jane Austen (1718) 70

B.3. Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley (1102) 70

B.4. The Adventures of Sherlock Holmes by Arthur Conan Doyle (930) 71

B.5. Alice’s Adventures in Wonderland by Lewis Carroll (832) 71

B.6. The Yellow Wallpaper by Charlotte Perkins Gilman (779) 72

B.7. Dracula by Bram Stoker (651) 72

B.8. You are forbidden! by Jerry Shelton (571) 73

B.9. The Picture of Dorian Gray by Oscar Wilde (569) 73

B.10. Moby Dick; Or, The Whale by Herman Melville (544) 74

B.11. A hitch in time by Frederik Pohl (541) 74

B.12. Weird Tales, Volume 1, Number 1, March 1923: The unique magazine by Various (521) 75

B.13. Grimms’ Fairy Tales by Jacob Grimm and Wilhelm Grimm (509) 75

B.14. A Tale of Two Cities by Charles Dickens (500) 76

B.15. A Modest Proposal by Jonathan Swift (488) 76

B.16. The Great Gatsby by F. Scott Fitzgerald (470) 77

B.17. The Snake’s Pass by Bram Stoker (465) 77
<table>
<thead>
<tr>
<th>Book Title</th>
<th>Author</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Iliad</td>
<td>Homer</td>
<td>78</td>
</tr>
<tr>
<td>The Scarlet Letter</td>
<td>Nathaniel Hawthorne</td>
<td>78</td>
</tr>
<tr>
<td>Ulysses</td>
<td>James Joyce</td>
<td>79</td>
</tr>
</tbody>
</table>
1. Introduction

In recent years, there has been an increase in Diversity, Equity and Inclusion (DEI) research in computer science ranging from identifying bias in machine learning to studying what makes different groups more comfortable with robots. This research is typically done to present information and inspire social change. Some key examples include a study by Li et. al in which reviews from a tourist website are analyzed using natural language processing techniques to explore the nature of racism as it relates to tourism [14]. In another important example, Leslie et. al turn the lens inward to the field of computer science, studying the race and gender biases that exist in facial recognition technologies [12].

For this important DEI research to most effectively accomplish goals of measurable change, findings must be presented to the general public. In most cases, traditional graphs and charts are the most efficient and effective mode of presenting experimental findings. If constructed properly, a basic line chart can summarize an experimental finding to someone previously unfamiliar with a study or even an entire field of science in a matter of minutes. But what if the goal is not necessarily efficiency? What if we hope to implore our readers to linger and fully consider the implications of our findings? What if we hope also to communicate our emotional responses, as researchers, to our audience? This is where a simple chart may leave something to be desired.

Art, for centuries, has been a universal mode for communication of complex emotions. As such, one way to humanize data visualization is to incorporate techniques
from visual art. To explore this idea, we designed and implemented a DEI-focused study that focused on issues that the research team related to personally. Then, we experimented with a range of art techniques to communicate these findings.

We chose to use existing techniques from Natural Language Processing (NLP) to study the lack of gender diversity of fictional characters. It is a well known fact that there are fewer female than male characters in fiction. But how have these inequities evolved over time? This work aims to answer this question for a set of 15,000 fictional works taken from Project Gutenberg, a free online library of ebooks, covering an author birth year range from 1750 to 1950. Beyond quantifying this difference in representation, we aim to further explore disparities by asking questions about how authors perceive the male versus female characters that they describe.

NLP is a field of computer science that can be used to extract meaning from or generate human language. NLP techniques can be used for a myriad of tasks from selecting appropriate resources to answer a Google search, to analyzing survey results, to poetry generation. One particular technique from NLP that is used heavily was sentiment analysis. Sentiment analysis is used to determine whether a writer or speaker feels positively or negatively about an object or person they are referencing. A classic application of sentiment analysis is to extract how a consumer feels about a product based on a review they wrote [9].

The results of our DEI study were then analyzed and visualized with a collection of traditional bar charts, scatter plots, histograms and line graphs. These helped us to narrow which results were most important to explore with art techniques. In addition to focusing on how to most effectively reach the audience with these results, we chose to incorporate an exploration of computational art, which we broadly defined as any art created with some form of computer augmentation. Our visualization techniques include a diverse range of types and amounts of computer input to the work.
These include computer-generated works, pen-plotter drawn work using paint on paper and on linen, up to purely studio based work inspired by the results of the NLP analyses.
2. Related Work

2.1. Natural Language Processing and Sentiment Analysis

Preprocessing is an imperative first step to many NLP tasks including frequency and sentiment analysis. Constructing the proper preprocessing pipeline, along with selecting the tools for each step, can have an impact on the effectiveness of the task itself [22]. In [1], Giulio et. al. present the effect that several preprocessing techniques including stop word removal and stemming had on the accuracy of a sentiment analysis task on twitter data.

The foundations used to design the NLP analysis portion of this work, including the preprocessing pipeline, were largely formed with two textbooks. The first source, *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics* [9] introduced concepts such as tokenization of words, case folding, lemmatization, sentence segmentation and part of speech tagging.

The second resource, *Natural Language Processing with Python—Analyzing Text with the Natural Language Toolkit* [3] discussed many of these same topics in light of the NLTK Python library, which has been referred to as “the mother of all NLP libraries”[10]. A more application-based source, we used many of the tools discussed throughout. In chapter three, processing raw text is discussed. We utilized the method for reading and decoding plaintext from a url with the urllib Python library. Further,
we used nltk’s sent_tokenize tool to divide the raw book text into sentences. We further divided these sentences into tokens with nltk’s word_tokenize tool as discussed in this chapter. Further, the chapter discusses normalizing text with the lower Python function to convert tokens into all lowercase letters. We used this method when identifying sentences that contained male and/or female-specific nouns and pronouns.

Chapter five of this book presents tools for categorizing and tagging words. They demonstrate nltk’s pos_tag function, which takes as input a list of tokens and returns a list of tuples where one element is the token and the other is a part of speech. We used this function to identify adjectives and adverbs within the male and female sentences.

Sentiment analysis is a constantly evolving and growing subfield in NLP. The goal of sentiment analysis is to determine whether a speaker’s opinion about a subject is positive, negative or neutral. Methods of sentiment analysis include lexicon-based approaches, machine learning-based approaches, deep learning-based approaches and transfer-learning based approaches [18]. Lexicon-based approaches, in which a dictionary of words with attached sentiment scores is used can be extremely effective and intuitive to apply to datasets that contain a wide range of writing styles [18]. SentiWordNet is an example of a lexicon that is commonly used in lexicon-based approaches [7]. Guidance for harnessing SentiWordNet can be found in [10].

Another resource, the article Identifying Emotions in Social Media: Comparison of Word-Emotion Lexicons was used to help design the methodology for the emotion detection portion of the text analysis. In the paper, the authors compare the performance of three different emotion lexicons: NRC, Depeche Mood and EmoSenticNet in detecting emotions within a set of twitter data. The results of the analysis using each lexicon were compared against ground truth obtained using a questionnaire survey where the authors asked participants to manually assign emotion scores to a set of data. They determined that, overall, the NRC lexicon was the most accurate lexicon[11].
2.2. Computational Text Analysis of the Female Role in Fiction

Resources were also used to obtain more specific background knowledge about the role of female characters during the time frame for which we gathered data.

The journal article, *The Transformation of Gender in English-Language Fiction* by Ted Underwood, David Bamman, and Sabrina Lee [21] pertained heavily to our work. In this article, Underwood et al. used computational analysis to examine gender representation in a library of 104,000 books mainly taken from the HathiTrust Digital Library. Along with frequency data, this article presented a measure of the difference in the characterization of male versus female characters by examining the accuracy of a machine learning model to predict whether sentences pertained to a male character or a female character. Results showed that the accuracy of the model steadily declined for literature through the 20th century, allowing the authors to conclude that gender barriers became less clearly defined during this time period. In an attempt to explain these findings, Underwood et al. randomly chose 10,000 words describing women and 10,000 words describing men. They subtracted these sets from one another. From these results, they chose words of interest and tracked the usage of these words across time.

We were able to find similar trends in the representation of male versus female characters across time in a set of fiction taken from Project Gutenberg, further addressing the scepticism that the trends found in the Underwood et al. study pertained only to the HathiTrust library or academic libraries in general. Rather than focusing on publication year of individual literary works, we chose to approach our study by examining the authors of these works; we include information in addition to author gender such as author birth year and birth country.

Where Underwood et al. chose to track the changes in the characterization of male versus female characters with a set of well-chosen words of interest, we present a more general sentiment analysis surrounding these characters. Further, we include in-
formation about which of the ten NRC emotions, including positive sentiment, negative sentiment, anger, fear, joy, sadness, disgust, surprise, trust, and anticipation [17] were more closely associated with each gender across time. While we also include information about individual words used in proximity to male versus female words, we chose to examine only the most frequent words associated with the two genders because we felt it was more generally representative of how authors approached their characters.

2.3. Data Visualization

A broad overview of the field of Data Visualization was required in gauging the effectiveness and novelty of our own art-based approaches and motivating the importance of developing them. “A Brief History of Data Visualization”, a section of the 2008 book, *Handbook of Data Visualization* written by Michael Friendly, highlights the key landmarks in the field of Data Visualization beginning with origins of cartography in ancient Egypt and ending with the interactive graphing methods of the present. Throughout this section, Friendly tracks the evolution of visualization techniques in conjunction with advancements in science, mathematics and technology. He states that new graphical methods throughout history “arose from concrete, often practical, goals: the need or desire to see phenomena and relationships in new or different ways”[8]. This idea further motivated our goal of creating new visualizations to better suite the emotional depth that can accompany DEI research. It also brings to mind the overarching idea that as advancements in fields such as NLP yield new data and directions of study, the ways in which we present these emerging ideas must too evolve.

An overview of the field of Data Visualization would be incomplete without the acknowledgement of Edward Tufte, whose books are credited with helping to make data visualization more widely considered in a variety of fields in recent years [5]. One such book, *The Visual Display of Quantitative Information*, presents what is essentially
a checklist for a high-quality graphic. He goes on to discuss a number of examples and non-examples of these requirements in use. Tufte, in breaking down these various graphics, pays careful attention to how and how well a viewer will be able process the presented data. These discussions provided us with the language and ideas we could use as the groundwork for evaluating our own data representations [20].

In another compelling portion of his 2001 book, Tufte discusses the idea of “graphical elegance” and the importance that aesthetics play in effective visualizations. The language used in this section is reflective of that used in discussing visual studio art, including considerations of color and balance. He also goes on to say that “[Attractive displays of statistical information] often have a narrative quality, a story to tell about the data” [20]. In the article Data visualization as creative art practice, the authors build upon these ideas, employing series of data-based artworks to motivate their argument that “the aesthetic consideration of data visualization in a design field thus could enhance the experience and emotional influence from data visualization and amplify the ability to obtain knowledge”[13]. These concepts hit at the very crux of goals for greater evocation in visualization of data involving potentially sensitive topics.

Looking beyond data-driven artworks, we examined other alternative forms of graphics. Namely, we explored two books by Manuel Lima: The Book of Circles: Visualizing Spheres of Knowledge [15] and The Book of Trees: Visualizing Branches of Knowledge [16]. Each book is a collection of visualizations from across time in the form of trees and circles, respectively. Aside from being excellent resources for exposure to hundreds of graphics, each book digs into the symbolism associated with each motif and how this symbolism may be harnessed to improve graphical representations. The 2012 David McCandless book, Information is Beautiful, is another collection of various “experiments” in visualization. The book contains dozens of graphs and charts concerning dozens of seemingly random topics that employ the use of symbolism, unique
aesthetics, and—perhaps most strikingly—comedy to create extremely effective visuals. Both of these resources allowed us to stretch our minds in regards to what a graphic can be and do.

Background research was also completed to explore the current landscape of specifically visualizing results from NLP analysis. In the paper, *Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization*, the authors created an interface to make text mining results more comprehensible to those in the humanities studying literature. They focus on visualizing frequent patterns within a given text. Aside from creating a useful tool, they justify the need to connect a wider audience to the results of NLP-based studies: “these text mining concepts can help the user to analyze the text, and to create insights and new hypotheses [6].”

In *How we draw texts: a review of approaches to text visualization and exploration*, the authors argue that “[t]ext visualization and exploration...constitute a subfield of data visualization”. This paper is a review of several important visualization methods for both single texts and text collections. It helped us to gain traction about what has been done in this field ranging from simple pie charts to word clouds to dynamic text visualizations[19].

A range of data-visualization and art techniques were experimented with in this study. Correspondingly, a range of resources were used to collect the skills needed to perform these experiments. The book *Generative Design: Visualize, Program and Create with Processing* by Hartmut Bohnacker, Benedikt Groß, and Julia Laub [4] helped us to learn the basic concepts of computational art in the context of Processing, a coding language used to create visual art. Though we primarily used p5.js, which is a javascript library marketed as an “interpretation” of Processing, there is a large overlap in the basic concepts of the two resources.
The Axidraw Pen Plotter was another tool that was used to create visualizations. Pen plotters are devices that copy a collection of user-input vectors onto a flat space using a pen or other utensil. When beginning to work with such a device, it’s easy to overlook the full impact this tool can have. Why not just use a traditional printer to create output? The article, *Biofeedback Painting: Let the Heart Lead the Brush* helped us to understand the best use cases for the pen plotter. The article presented a project where a plotter, in real time, created custom pieces of artwork based on a subject’s heart beat using input from a pulse sensor. A pen plotter is perhaps best used in cases in which the art-making process is an interactive experience for the viewer. They also present the idea that the medium used in conjunction with a plotter is important: “The type of brush, the texture of the paper and the movements of the pen can affect the physical result, which generates aesthetic forms of digital visualization that are beyond human or machine control” [23]. These concepts were crucial in designing the pen plotter-based visualizations included in this work.
3. NLP Analysis

3.1. The Data Set

Project Gutenberg is an online collection of over 60,000 eBooks in the public domain. The books are obtained through donations and selections are made with a focus on literature and other written works. The majority of work included in Project Gutenberg are those whose U.S. copyright has expired.

Metadata for all entries in Project Gutenberg may be downloaded from the library’s website in the form of a csv catalog. The catalog includes the type of entry, title, language, author name(s), author birth year(s), subject(s), and the Project Gutenberg bookshelf that the work appears on. Also included is a text number that corresponds to the web address where the eBook resides. The books are available in HTML, EPUB, Kindle, or plain text format.

3.2. Data Cleaning

As we were not interested in all of the eBooks included in the Project Gutenberg catalog, we chose to construct a subset of this information and to put it into a form that was more useful to us. We completed this task by reading the catalog CSV into a Pandas DataFrame and cleaning the data using Python.

Because we were interested in studying the author relationship with male versus female characters, we filtered for all entries that included the word “fiction” but did not
include the word “nonfiction” in the subject column. We chose to eliminate eBooks not written in English and all sound files in the catalog. We removed entries with author birth years occurring before 1700 and after 1950. We removed translators from the authors column and filtered out all submissions with more than one author. Finally, we eliminated entries for which the author birth year was not included. Cleaning and filtering the data reduced the size of the catalog from 66,004 to the subset of 16,330 eBooks used for analysis.

![Gutenberg Fiction by Author Birth Year](image)

Figure 3.1.: The majority of fictional books analyzed had authors with birth years between 1800 and 1900.

### 3.3. Hand Tagged Data

For each of the 3,000 authors whose work was included in the final data set used for the sentiment and frequency analysis, we added hand tagged data. We recorded the author birth country and author gender. For authors born in the UK or Ireland we
also included which country they were born in: England, Scotland, Wales, or Ireland.
Because of the range of times we did not differentiate between the Republic of Ireland
and Northern Ireland, instead group all under the heading of Ireland. This added data
is not currently widely available for Project Gutenberg and enhanced the data analysis
significantly.

3.4. Sentiment and Frequency Analysis

After obtaining the desired set of data to be analyzed, for each entry in the
reduced catalog, we constructed the url where the plain text version of the book resides
using the text number included in the catalog. We used the urllib Python library to
pull the plain text into a single Python string. We used the gutenberg.cleanup Python
library to remove the header and footer that is included with every Project Gutenberg
eBook. We used the NLTK Python library to tokenize the single string into a list of
sentences from the text. Each sentence string was then tokenized into a list of words
using NLTK’s word tokenizer. The list of words was converted into a set of words and
set intersections with pre-determined sets of gender-specific words, which may be found
in Figure 3.2 were calculated. If the sentence intersected with the set of female-specific
words, it was labelled as a female sentence. If the sentence intersected with the male-
specific words, it was deemed a male sentence. If it intersected with both, it was labelled
as both a male and female sentence.

The tokens of all sentences deemed male, female, or both were tagged with
their part of speech using NLTK’s part of speech tagger. These tags were converted
into WordNet tags and the tokens paired with their tags were passed to the WordNet
lemmatizer function.

We collected sentiment scores for all tokens that were labelled adjectives or
adverbs. The sentiment scores were obtained from SentiWordNet, which is a lexical
resource that assigns a positive, negative and neutral score between zero and one to a word or group of words.

To obtain sentiment scores for each sentence, we averaged both the positive and negative scores obtained for the adjectives and adverbs found in the sentence.

We recorded the average positive and negative sentiment scores of all female sentences in the entry. We also recorded the number of female sentences divided by the total number of sentences in the eBook. We recorded the same information for male sentences in each eBook.

3.5. Emotion Analysis

Much like in the previous section, the plain text of each book was obtained and tokenized into sentences. Each sentence was tokenized into words to determine whether it contained any male or female-specific words. The sentences were labelled as male, female, or neither.

The NRC emotion lexicon focuses on ten categories: anger, fear, anticipation, trust, surprise, sadness, joy, disgust, positive sentiment and negative sentiment [17] [11]. For each of these emotions and sentiments, it contains a list of associated words. A raw sentence can be used to create an NRCLex object using the nrclex python library and the number of words in the sentence that pertain to each category may be tallied.

After a book’s sentences were labelled, all raw male and female sentences were used to create NRCLex objects. The tallies for each category were summed for all male and female sentences, respectively. Then, these counts were normalized by dividing by the total number of male or female sentences, respectively. The result was two dictionaries: one for “male” emotions and the other for “female”. The keys were the ten NRC Lexicon categories and the values were the average number of words associated with the category in each sentence.
3.6. Adjective and Adverb Frequency Analysis

We chose to supplement the findings of the sentiment and frequency analysis with more specific data about the adjectives and adverbs that appeared in the data set in association with male versus female characters. We grouped the set of eBooks on the decade in which the author was born. For each decade, we collected the 1000 most frequently used adjectives and adverbs used in female and male sentences. Then, we calculated a set difference between the female and male adjectives and adverbs to obtain a list that was specific to each respective gender.
<table>
<thead>
<tr>
<th>Female Words</th>
<th>Male Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>she</td>
<td>he</td>
</tr>
<tr>
<td>her</td>
<td>him</td>
</tr>
<tr>
<td>hers</td>
<td>his</td>
</tr>
<tr>
<td>herself</td>
<td>himself</td>
</tr>
<tr>
<td>woman</td>
<td>man</td>
</tr>
<tr>
<td>women</td>
<td>men</td>
</tr>
<tr>
<td>women’s</td>
<td>men’s</td>
</tr>
<tr>
<td>gal</td>
<td>guy</td>
</tr>
<tr>
<td>girl</td>
<td>boy</td>
</tr>
<tr>
<td>girls</td>
<td>boys</td>
</tr>
<tr>
<td>girlfriend</td>
<td>boyfriend</td>
</tr>
<tr>
<td>sister</td>
<td>brother</td>
</tr>
<tr>
<td>sisters</td>
<td>brothers</td>
</tr>
<tr>
<td>mom</td>
<td>dad</td>
</tr>
<tr>
<td>moms</td>
<td>dads</td>
</tr>
<tr>
<td>mother</td>
<td>father</td>
</tr>
<tr>
<td>mothers</td>
<td>fathers</td>
</tr>
<tr>
<td>grandmother</td>
<td>grandfather</td>
</tr>
<tr>
<td>grandmothers</td>
<td>grandfathers</td>
</tr>
<tr>
<td>grandma</td>
<td>grandpa</td>
</tr>
<tr>
<td>grandmas</td>
<td>grandpas</td>
</tr>
<tr>
<td>granddaughter</td>
<td>grandson</td>
</tr>
<tr>
<td>wives</td>
<td>husbands</td>
</tr>
<tr>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>mrs</td>
<td>mr</td>
</tr>
<tr>
<td>miss</td>
<td>master</td>
</tr>
<tr>
<td>ms</td>
<td>-</td>
</tr>
<tr>
<td>niece</td>
<td>nephew</td>
</tr>
<tr>
<td>nieces</td>
<td>nephews</td>
</tr>
<tr>
<td>daughter</td>
<td>son</td>
</tr>
<tr>
<td>daughters</td>
<td>sons</td>
</tr>
<tr>
<td>aunt</td>
<td>uncle</td>
</tr>
<tr>
<td>aunts</td>
<td>uncles</td>
</tr>
<tr>
<td>widows</td>
<td>widowers</td>
</tr>
<tr>
<td>lady</td>
<td>gentleman</td>
</tr>
<tr>
<td>ladies</td>
<td>gentlemen</td>
</tr>
</tbody>
</table>

Figure 3.2.: A lexicon of gender-specific nouns and pronouns was used to identify male versus female sentences.
4. Results

For each set of results from our computational text analysis, we used the python library plotly to generate traditional bar chart, pie chart, line plot and scatter plot visualizations. These visualizations allowed us to focus on the most interesting insights which were used as the focus for the art-based visualizations.

4.1. Gendered Sentence Frequency

Our first objective in examining the collected data was to determine whether we observed the same counter-intuitive trend found by Underwood et al. in a collection of works taken from the Hathi Trust Library: the ratio of female to male representation steadily decreased in books published in the early 20th century [21]. As Figure 4.1 illustrates, the percentage of female-specific gendered sentences declined in books with author birth dates between 1850 and 1930 in the Project Gutenberg fiction collection.

Underwood et al. reported a decline in female character appearances in books “from the nineteenth century through the early 1960s...” [21]. Though this range is shifted forward in time from the trend we noted, we attribute this offset to the difference that we used author birth date to order works chronologically while Underwood et al. used original publication date. Having confirmed this trend was also present in the Gutenberg fiction-specific collection, we continued seeking further details and possible explanation.
Figure 4.1.: There was a significant drop in the number of female versus male sentences with author birth years between 1850 and 1940.

4.2. Author Demographics

Another first step in the analysis was visualizing the hand-tagged data that was added for the 3,000 authors whose work appeared in the library. Much like with the frequency of female sentences, we were unsurprised to find that the majority of the 3,000 unique authors were male. As shown in Figure 4.2, 60.4% of the authors were male while 25.3% were female. We were unable to identify the remaining 14.3%. In future work, we would like to invest time in decreasing the percentage of un-identified author genders.

Figure 4.3 illustrates the percentage of books with identified authors who were by females. From this graph we observe that the peak percentage of female authors
Figure 4.2.: The number of male authors whose work appears in the data set is significantly greater than the number of female authors.

occurred in the birth year range 1760-1770, where over 70% of books with known authors were by women. This percentage decreased significantly and hovered around 30% between 1820 and 1910. Following this period, the percentage decreased steadily from 1900 to 1940, hitting a minimum of about 5%. Because these results only necessarily hold for the Project Gutenberg collection we used for experimentation, it would be interesting to explore other corpora with more content to determine whether there is a similar initial high percentage of female authors and similar drop during this period.

We noted that though the trends did not perfectly align, there was a decrease in the percentage of books by female authors in the same time range in which the frequency of female sentences dropped. This observation lead us to explore the possibility that the drop in female authors may have contributed to the drop in female sentences.
There was a sharp drop in the percentage of books by female authors beginning in 1900.

From here we explored the ratio of female to male sentences in books by women versus books by men.

Examining this data revealed that, overall, female authors dedicated more sentences to female characters than male characters overall. However, this ratio dropped significantly for both male and female authors, helping us to understand that the shift in fiction away from female characters was not specific to one gender of author, though the drop in female authors certainly contributed to the drop in female representation.

4.3. Sentiment Analysis

We now examine how the sentiment of sentences containing male versus female subjects evolved over time. For each sentence labelled male, female, or both, SentiWordNet was used to assign a positive, negative and neutral sentiment score between...
Figure 4.4.: Both female and male authors with birth years between 1850 and 1940 wrote about fewer female characters in their work.

zero and one. Books were divided into ten-year increments, or buckets. For each bucket, male and female positive and negative sentiment scores were averaged. These averages were visualized in chronological order in a line chart to be analyzed.

The positive sentiment scores, shown in 4.5 revealed that female characters were mentioned in a more positive context than male characters from around 1750 to 1910. During this time frame, changes in male and female positive sentiment closely mirror one another. One such notable change is the steady decline in positive sentiment from 1860 to 1910. Perhaps the most important observation, though, is the change that occurred around 1910: the positive sentiment in female sentences ceased to be greater than in male sentences.
In examining the negative sentiment scores, we observe that the negative scores remained extremely close between male and female sentences from 1750 to 1880. However, there is a steady divergence between 1880 to 1930 where the negative sentiment scores for female sentences became increasingly higher than that of the male sentences. There is a large overlap between this divergence and the steady decrease of female sentences.

Figure 4.5.: There was a decrease in the positive sentiment in both male and female sentences between 1850 and 1940.

4.4. Emotion Analysis

In analyzing the data obtained using the NRC Lexicon to study the emotions of fear, anger, anticipation, surprise, sadness, disgust and joy as they appeared in male versus female sentences, we gained further intuition about the results of the sentiment analysis. Interestingly, we observed a decrease in the presence of every included emo-
Figure 4.6.: There was an increase in the negative sentiment in both male and female sentences between 1850 and 1940.

Figure 4.6.: There was an increase in the negative sentiment in both male and female sentences between 1850 and 1940. However, this drop was more significant in female sentences for fear, anger, trust, surprise, and disgust. As the emotional language in the library decreased, it decreased more surrounding female characters.

4.5. Frequent Adjective and Adverb Analysis

Observations from the results of the frequent adjective and adverb analysis, which may be found in Appendix A, helped to provide more detail about how authors described their characters. The first observation is that the words appearing in the respective sets feel extremely in-line with typical gender stereotypes. For example, the 1850-1860 bucket of adjectives and adverbs occurring frequently in male but not in female sentences include, to name a few, the words “honourable”, “financial” and “sturdy”. The corresponding female bucket includes “maternal”, “dainty”, “hysterical”
and “sweetly”. This trend of words that reaffirm typical gender norms appears throughout the set.

The second observation is that the words used to describe women involve physical appearance more often than do the words to describe men. The female sets include “beautifully”, “desirable”, “slim” and “supple”. While the male set includes words such as “handsome” and “bearded” in later years, “military”, “good-natured” and “tough” are far more frequently-occurring. Though no specific conclusions may be drawn from looking at these lists of words, we foresee them being useful and informative in future computational analysis of this body of text.

To summarize our keys results, we identified a steady decrease in the representation of females in fictional works whose author birth years occurred between 1850 and 1930. Further, we found that the positive sentiment in sentences with both female and male characters decreased during this time while negative sentiment increased. However, the decrease in positive sentiment and increase in negative sentiment was greater for “female” sentences. While by themselves these results are interesting, they open up more questions that require further interdisciplinary study to answer. In response to the sentiment analysis results, we ask how the decrease in positive sentiment and increase in negative sentiment, which is more significant in female than in male sentences may have been related to current events during the time this shift occurred.
5. Art Based Visualization

Following data analysis, we used visual art techniques to create a portfolio of pieces that seek to give appropriate emotional weight to these findings. In addition to evoking an emotional response from the viewer, thus enabling the viewer to see data in another light, these pieces explore the relationship between an artist and technology through varied artistic approaches and tools. We ask viewers to consider the broader theme of what role technology can play both in revealing new details about DEI issues and in artistic creativity.

5.1. "Shift" Frequency Visualization

One of the first pieces created in this portfolio focused on frequency analysis findings for a singular work. To create it, a novel was divided evenly into three parts, meant to symbolically represent the rising action, climax, and falling action. While a detailed literary analysis to find the exact location of these three sections could be used, we use this simplistic approach to allow the technique to be applied to any text with minimal user effort and to clearly define boundaries in an unambiguous way. Frequency values were obtained for the male and female sentences. The piece was constructed in p5.js and includes six rectangular segments. The top row of rectangular segments symbolically represents the rising action, or the first of the three divisions. The second row represents the climax and the third the falling action. The left column represents, in each of the three divisions, the percentage of all gendered (male and female) sen-
tences that are male-specific. Conversely, the right column represents the percentages of gendered sentences that are female-specific. Each row of the figure contains the same number of dots and these percentages dictate how many of the dots are randomly generated into the left-hand rectangle versus the right.

Figure 5.1.: "Shift" frequency visualization displaying results from gender frequency analysis of The Great Gatsby

Though “Shift” gender frequency visualizations for 20 of the most commonly downloaded ebooks on the Project Gutenberg website, may be viewed in Appendix B, we have selected the three diagrams that appear in this section to demonstrate the strengths of this method and present a variety of outcomes from the analysis. F. Scott Fitzgerald’s The Great Gatsby, yields interesting results because, in the first third of the book, the gender representation is nearly equal. However, as the story progresses, it becomes more male-centered. Figure 5.1, the visualization corresponding to The Great Gatsby, captures this shift.
Jane Austen’s 1813 novel, *Pride and Prejudice* exemplifies near-equal gender representation throughout. The resulting “Shift” diagram shown in Figure 5.2 is almost completely symmetrical and felt, to the research team, satisfying to view. In stark contrast, Herman Melville’s *Moby-Dick; or The Whale* features vastly more male representation than female. Figure 5.3 makes this fact immediately obvious. It lacks symmetry and is overall visually unappealing, feeling chaotic and off-balance. This is intentional and meant to express the sentiment that the vastly male-dominated representation is, put simply, ugly.

When displayed on a web page, there is an added feature such that when the piece is clicked, the random location of points within the rectangular segments is regenerated, resulting in a slight shifting effect. This feature, along with the success of
this method in communicating how gender representation within a novel can change, was inspiration for the name of the collection of frequency visualizations: “Shift”.

5.2. X-icon Representation

While the “Shift” series is successful in visualizing the evolution of gender representation throughout a single book, we wanted to include a visualization technique that could concisely capture this same information for a large number of books. Simply placing a set of the “Shift” pieces in rows and columns was visually disorienting, as it was difficult to discern which rows and columns signified a single work.

Our solution was to use a single shape to visualize all information for a book. We chose to use Xs. The letter X is made up of two intersecting diagonal line segments. For the ease of explaining the visualization, we will denote the “first” segment of the X as the line whose endpoints appear on the upper left hand side and lower right hand
side. It follows that the “second” segment is the line whose end points appear on the lower left and upper right. In our visualization, the first segment of the X represents female representation while the second, the male. The two segments are divided into three equal parts. The first third of the book is represented by the leftmost third of the X, the second third of the book by the middle third of the X and so on. The width of the individual thirds of each segment are proportional to the percentage of total gendered sentences that are labelled with the respective gender within the corresponding division of the book.

This arrangement is displayed in Figure 5.4 for the results of the analysis of “The Great Gatsby”. We know from observing Figure 5.1, that the gender representation in the “The Great Gatsby” begins relatively equally, but becomes gradually more male-centered. In the X-icon representation, this trend manifests as, from left to right, the “second” segment of the X becoming wider while the “first” segment becomes narrower.

Xs were chosen, not because of their success in visualizing a singular work, but because of their ability to convey a summary of a large number analysis results when lined up in rows and columns. As is shown in Figure 5.5, which visualizes the hundred most downloaded ebooks from Project Gutenberg in the week of October 3, 2022, it’s quickly obvious to the eye that the second segments of the X are consistently wider.
than their corresponding first segments. Though a simple observation, it’s powerful in conveying the idea that a large number of the depicted works feature male-dominated gender representation. Note that a complete list of the depicted works may be found in Appendix C.

5.3. Plotter Art

Transitioning from purely digital art into digitally-controlled physical art, the next pieces were created using a pen plotter. A pen plotter is a device that takes a collection of vectors as input and moves a writing utensil to replicate these vectors onto paper or another flat surface. We had access to an AxiDraw V3, which is marketed by
Evil Mad Scientist as a “High Performance Personal Writing and Drawing Machine”. While this type of machine is most commonly used with a pen or pencil, we found that the output was most interesting when we used it to create acrylic paintings.

To do so, we designed the individual pieces in Adobe Illustrator layer by layer. Each layer began with a line in the space to the right of the drawing surface. This line was designed to drag the paintbrush through the paint palette (which would be placed to the right of the drawing surface) thus loading the paintbrush with sufficient paint to execute the following lines within that layer. These designs were exported as SVGs and opened in Inkscape, which is the default software used to communicate with the AxiDraw. Once in Inkscape, the drawings were executed one layer at a time by the pen plotter, allowing the operator to reload the paint palette as needed. This setup, with the paintbrush attached to the plotter and the paint palette placed to the right of the drawing surface is shown in Figure 5.6

This technique resulted in an interesting interplay between the ability of the plotter to create precisely straight lines and the inconsistent, sometimes messy, texture of the paint. Figure 5.7 is an example of this. This piece was created in reaction to
the trend shown in 4.1. The color blocks represent each of the fifty year blocks of time that author birth years fell within in the data set. Moving from left to right, the first column represents the first fifty year block of author birth years in the data set, ranging from 1750 to 1800. The rightmost column represents data from books with author birth years from 1900 to 1950. We anticipated the resulting data from analyzing books with author birth years from 1750-1900; there was greater male representation than female. Blue and purple were selected for the first three columns of blocks to represent our unsurprised, though unhappy, reaction when the frequency data for the corresponding spans of time matched our assumptions. The red color block representing 1900 to 1950 is meant to be abrupt and striking, reflecting our puzzlement when female representation in the data set abruptly dropped off. Aside from creating a stark contrast with the previous blue blocks, red was also selected to represent the anger and frustration in finding such retrograde.

The acrylic painting pictured in Figure 5.8, extends ideas from the previously discussed painting by augmenting with reactions from the sentiment analysis, whose
Figure 5.8.: Acrylic on paper executed by a pen plotter aimed at expressing surprised response to decrease in both female representation and positive sentiment.

results are shown in 4.5. Like the previously described piece, the four columns represent the fifty year blocks of time. In this rendition, however, the top row illustrates findings from the frequency analysis while the bottom row represents results from the sentiment analysis. Once again, blue and purple were selected to reflect the first three half-centuries because the results were relatively unsurprising. Red was selected to represent the sudden drop in positive sentiment surrounding both male and female characters and, perhaps more surprisingly, the fact that the drop in positive sentiment was greater for female characters than male. We ask the viewer to consider these trends more deeply. Is there a causal relationship between the drop in positive sentiment and lack of female representation? What was happening in the world that may have contributed to the shift in literature?

The painting in Figure 5.9 is also a piece of work designed with p5.js and executed by the AxiDraw pen plotter with black paint. Instead of focusing on a trend over time,
this piece depicts the number of female versus male sentences in the entire data set. A total of 100 lines were randomly generated in the complete piece. The number of lines that appear on the right side versus the left side represent the ratio of female to male sentences throughout the data set. Once again, we rely on the lack of balance in the piece to encourage viewers to digest the large disparity in gender representation.

Though impactful when executed on paper, we found the piece to be more visually interesting on linen, as shown in Figure 5.10. The raw edges of the material and inconsistent texture of the paint lend an earthier feel to the piece. In addition to being more visually interesting, using linen as a medium calls upon the domesticity or femininity typically associated with textiles, based on the idea that crafts such as sewing, knitting, etc. are traditionally considered “women’s crafts” [2]. The juxtaposition of calling attention to unequal representation and a medium loaded with gender stereotypes are intended to add layer of evocation.
5.4. Dyed Linen

The visualization shown in Figure 5.11 features ten scraps of linen, each representing a ten year block of novels based on author birth year. The first strip represents 1850-1860, the second 1860-1870 and so on. The final strip represents 1940-1950. The dyed portion is proportional to the amount of female representation while the undyed material is proportional to the male representation within each block of time. These specific dates were chosen because they fully captured the decline of female representation that we witnessed in the data set.

Because we found the use of cloth as a medium for the visualizations interesting and evocative in previously discussed pieces, we chose to expand on this idea through this visualization by adding in the process of dying, which is also traditionally considered a female-specific craft. The process of precisely dying the correct amount of each scrap in order of the date they represented opened a new level for us as artist to connect more
deeply with the disparity in gender representation we are conveying. These statistics were not simply numbers or randomly-generated points and lines as they appeared to us in previous visualizations. Now, they were palpable and diminishing slivers of dyed material.

Though fundamentally reminiscent of a typical bar chart, this textile work is made more interesting by the attached symbolism and idea that it is a physical manifestation of something as cold and mathematical as a ratio. Beyond being a successful visualization, it was an outlet for fully processing the implications of our work.

5.5. Kaleidoscope

“Rotating through Time” is an animated piece created with p5.js with the central idea that the viewer is twisting a kaleidoscope. As the kaleidoscope turns, the viewer is being transported through time within the data set. The area of the inner ring, which is made up of warm colors and rotates clockwise, is proportional to the percentage of female sentences. The image whose fragments make up this ring is shown to the right in Figure 5.12 and is meant to be more stereotypically feminine. The outer ring features cool colors and is proportional to the percentage of remaining gendered sentences, or the male sentences. The image used to construct the outer ring is shown on the left.
in Figure 5.12. The choice to use stereotypically feminine versus masculine images, though invisible to the viewer, was made specifically because the images are fragmented to create the final piece. It is a hidden layer to the piece aimed to express the desire to dismantle or destroy such norms or stereotypes.

By abstracting away any dates or numerical ratios, the viewer is more able to lose themselves within the animation. This may allow them to fixate on the outward and inward motion of the inner ring and, as a result, the fluctuation of female representation. More specifically, it brings attention to the shrinking of the inner ring that begins near the middle of the animation. The inner ring shrinks until nearly disappearing, before it begins to expand once again seconds before the end of the video. This growing, brought on by the slightly higher relative female representation in books written by authors born between 1940-1950 feels almost encouraging or hopeful. While the female representation may have approached zero, there is hope for more inclusion in years following those present in this data set.
5.6. Cocktail Dress

The final piece that was created for this portfolio is a sculpture constructed with a female mannequin and a cocktail dress. The originally black mini dress, featuring a halter top and ruched side details was coated in canvas primer and white paint. Then, it was decorated with text. The dress is pictured in Figure 5.15 The words painted on the dress were taken from the list of words that appeared on the list of 1,000 most frequent adjectives and adverbs that were found in female sentences, but did not appear in the list of 1,000 most frequent adjectives and adverbs in male sentences for the 1880-1890 author birth year bucket. The words were selected from the list, which appears in Figure 5.14 and in Appendix A, because they felt, to the research team, applicable to expectations or pressures still felt by women today. These words, paired with the dress, make a statement about these expectations and pressures.

The cocktail dress, being a resounding image and possessing layers of symbolism, is felt to be the most powerful of the visualizations in the portfolio. In light of this, future work may include creating a complete collection of cocktail dress visualizations aimed at transporting a viewer through the process of uncovering and processing the results of this human-centered computational literary analysis.
Figure 5.14.: List of words that appeared on the list of 1,000 most frequent adjectives and adverbs in female sentences, but did not appear in the list of 1,000 most frequent adjectives and adverbs in male sentences for the 1880-1890 author year bucket and their attached frequencies.

Figure 5.15.: Front and back view of cocktail dress visualization.
6. Limitations

Perhaps the most glaringly obvious exclusion in our NLP analysis is the limitation to studying binary gender. While this posed little issue in the data set we used due to the lack of representation of other genders in the time periods in which the works were written, further considerations would be necessary if completing a similar study on more modern literature.

We also acknowledge that the lexicon of gendered words is not extensive and, as such, some gendered sentences may not have been identified. Further, we made no attempt to classify proper nouns as male or female because we felt it would introduce unnecessary error. Instead of ensuring that all gendered sentences were identified, we focused on ensuring that each word in the male set had an equivalent in the female set (i.e. father and mother).

Other limitations lie in the sentiment analysis portion of the work. Certain literary techniques that may alter the results of sentiment or emotion scores of a sentence were unaccounted for. For example, if a hyperbole, in which an author or speaker purposefully exaggerates, is used within a gendered sentence, the sentiment score of the sentence may be swayed drastically. Further, sarcasm, which could result in a completely opposite sentiment score than the true sentiment was not detected. Being no suitable solution to sarcasm detection for this broad collection of writing, which seems to fall hand-in-hand with detection of other literary devices, we acknowledge that appearances of these types of language would lead to inaccurate sentiment analysis results.
7. Future Work

The results of the computational analysis portion of this work—namely the decrease in female representation and positive sentiment beginning in books with authors born the mid-nineteenth century—raise a number of questions. We consider the state of the world in the early 1900’s and wonder whether events such as the great depression and world wars may have shifted the trajectory of fiction, increasing overall negative sentiment. We ask whether the more significant drop among female-defined sentences may have arisen from resentments related to the first wave of feminism or achievement of women’s suffrage in the United States. We also wonder whether the trend may be related to a shifting of author interests away from topics such as romantic courting as seen in the Victorian era, to those less centralized around women, such as science fiction. We fully acknowledge and embrace our lack of expertise on these specifics, but present our findings as an opportunity for others to explore these questions.

Keeping cross-disciplinary research in the forefront of our minds, we propose a number of directions for further investigation. Enlisting the help of a literature expert to help develop lexicons for an expanded NLP analysis is one such direction. This could help to further identify how female versus male characters are described. Observations from the frequent adjective and adverb analysis, which includes language that reinforces gender stereotypes, suggest that it may be fruitful to develop a list of words that implies a domestic context. Then, one could determine how many of the female versus male sentences contain one of these words, thus allowing us to have a conversation about how
often female versus male characters appear in a domestic context and how this evolved over time. It may also be interesting to study the sentiment scores of these sentences. Such studies would help to add context to the results of the general sentiment analysis as well as observations from the frequent adjective and adverb analysis.

In the vein of studying the author-character relationship, it would be useful to hand-tag each of the fictional works with their original publication date. This could help to determine whether the age of the author when the book was written affected characterization of the two genders. It would also be useful to conduct case-studies with the help of a historian or literary expert to explore the life of authors whose work exemplifies some of the noted trends. This method could help to gain further insight about the author’s point of view and help to direct further study.

Perhaps less directly related to these specific trends, we bring attention to the large number of female authors within the dataset that used male pseudonyms. We are curious to study whether female authors writing under male names approached gender in any way that was distinctly different from those writing under female names.

In addition to the results of the computational literary analysis, results from the data visualization constitute further exploration. We ask whether the techniques we used allow the viewer to better understand the data. Do they intrigue the viewer to learn more about the subject? Do these types of visualizations evoke the appropriate emotional response? In the future, we hope to explore these questions and more from a Human-Computer Interaction perspective. At the 2022 The Society for Literature, Science, and the Arts conference, we presented our art pieces. Due to the positive response of the viewers, we feel encouraged that these methods could be of interest.

We propose displaying the artworks in a variety of settings. For example, formatting the work as an art exhibit in a gallery or as an informational display in an engineering and computer science hall. Then, results in the form of survey responses
may be collected from viewers. This kind of study could help to answer questions about whether the portfolio accomplished the goals we set in creating it. It may also help us to understand in what setting and for what group of people art-based visualizations are most effective and why.
8. Conclusion

This work focused on the inequality of gender representation among characters in a set of fiction taken from the Project Gutenberg library. To quantify this imbalance, we constructed a lexicon of common male and female-specific nouns and pronouns and, after appropriate text cleaning including tokenization and lemmatization, determined whether each sentence contained one or more of these words. Based on these results we tagged individual sentences as containing mention of a female character, male character, both or neither.

Upon analysis of these sentence taggings, we observed a predictably higher number of male versus female sentences. More interestingly, we observed a significant decrease in the ratio of female to male sentences among books authored by those born in the late 19th and early 20th century. This pattern matched with a similar pattern found in fiction taken from the Hathi Trust Library [21], which we found validating in reinforcing the viability of our analysis approach.

In addition to seeing patterns in the ratio of male to female sentences, we aimed to characterize these patterns. To do so, we conducted a sentiment score analysis, in which an average positive sentiment score for all male and all female sentences in a book were recorded. Analyzing the sentiment data revealed that, as the ratio of female to male sentences decreased, the positive sentiment scores decreased while the negative sentiment scores increased. The drop in positive sentiment in female sentences was greater than that of the male sentences.
After the analysis results were collected, we created an art portfolio to communicate the results in emotionally evocative and unique ways. The art pieces utilize techniques from computational art; some were created with p5.js and harnessed randomization, others were manually designed and executed using an AxiDraw pen plotter. Others were created using pure studio art techniques. In creating these pieces, we aimed to explore novel modes of data visualization that are more emotionally charged than a traditional chart or graph. Upon creation and presentation of the pieces, we believe that an art-based approach to visualization could be more suited to data surrounding human issues, such as the one examined in our NLP study. It may help to humanize collected data, which can feel cold and sterile, making it more palatable to the general viewer.
Bibliography


A. Appendix - Frequent Adjective and Adverb Analysis

Results

<table>
<thead>
<tr>
<th>Year</th>
<th>Male Adjectives and Adverbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Period</td>
<td>Words</td>
</tr>
<tr>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>Year Range</td>
<td>Top Words</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Time Period</td>
<td>Words Marked</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Period</td>
<td>Words</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Period</td>
<td>Words</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
</tr>
<tr>
<td>Date Range</td>
<td>Common Words</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Period</td>
<td>Words</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>Year</td>
<td>Female Adjectives and Adverbs</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>Time Period</td>
<td>Words</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
</tr>
<tr>
<td>Era</td>
<td>Words by Year</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------</td>
</tr>
</tbody>
</table>
B. Appendix - "Shift" Frequency Visualizations

Figure B.1.: Walden, and On The Duty Of Civil Disobedience by Henry David Thoreau (1798)
Figure B.2.: Pride and Prejudice by Jane Austen (1718)

Figure B.3.: Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley (1102)
Figure B.4.: The Adventures of Sherlock Holmes by Arthur Conan Doyle (930)

Figure B.5.: Alice’s Adventures in Wonderland by Lewis Carroll (832)
Figure B.6.: The Yellow Wallpaper by Charlotte Perkins Gilman (779)

Figure B.7.: Dracula by Bram Stoker (651)
Figure B.8.: You are forbidden! by Jerry Shelton (571)

Figure B.9.: The Picture of Dorian Gray by Oscar Wilde (569)
Figure B.10.: Moby Dick; Or, The Whale by Herman Melville (544)

Figure B.11.: A hitch in time by Frederik Pohl (541)
Figure B.12.: Weird Tales, Volume 1, Number 1, March 1923: The unique magazine by Various (521)

Figure B.13.: Grimms’ Fairy Tales by Jacob Grimm and Wilhelm Grimm (509)
Figure B.14.: A Tale of Two Cities by Charles Dickens (500)

Figure B.15.: A Modest Proposal by Jonathan Swift (488)
Figure B.16.: The Great Gatsby by F. Scott Fitzgerald (470)

Figure B.17.: The Snake’s Pass by Bram Stoker (465)
Figure B.18.: The Iliad by Homer (462)

Figure B.19.: The Scarlet Letter by Nathaniel Hawthorne (461)
Figure B.20.: Ulysses by James Joyce (459)
C. Appendix - 100 Most Frequently Downloaded Gutenberg ebooks in the week of October 3, 2022

Dubliners by James Joyce (2253)
Wuthering Heights by Emily Brontë (2095)
The Brothers Karamazov by Fyodor Dostoyevsky (2083)
Little Women by Louisa May Alcott (2060)
Autobiography of Benjamin Franklin by Benjamin Franklin (2023)
The Prophet by Kahlil Gibran (1972)
The Romance of Lust: A classic Victorian erotic novel by Anonymous (1946)
Anne of Green Gables by L. M. Montgomery (1940)
Les Misérables by Victor Hugo (1847)
Don Quixote by Miguel de Cervantes Saavedra (1839)
Narrative of the Captivity and Restoration of Mrs. Mary Rowlandson by Mary White Rowlandson (1836)
Carmilla by Joseph Sheridan Le Fanu (1830)
The Time Machine by H. G. Wells (1814)
A Study in Scarlet by Arthur Conan Doyle (1794)
Emma by Jane Austen (1778)
The Hound of the Baskervilles by Arthur Conan Doyle (1776)
The Tragical History of Doctor Faustus by Christopher Marlowe (1758)
Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley (1736)
Peter Pan by J. M. Barrie (1717)
The Wonderful Wizard of Oz by L. Frank Baum (1716)
The Works of Edgar Allan Poe Volume 2 by Edgar Allan Poe (1705)
Old Granny Fox by Thornton W. Burgess (1697)
Anna Karenina by graf Leo Tolstoy (1679)
The Kama Sutra of Vatsyayana by Vatsyayana (1661)
Essays of Michel de Montaigne Complete by Michel de Montaigne (1657)
Oliver Twist by Charles Dickens (1647)
International cartoons of the war by (1630)
Incidents in the Life of a Slave Girl, Written by Herself by Harriet A. Jacobs (1622)
Anthem by Ayn Rand (1615)
The War of the Worlds by H. G. Wells (1605)
How to Tell a Story, and Other Essays by Mark Twain (1598)
The Life and Adventures of Robinson Crusoe by Daniel Defoe (1580)
Sense and Sensibility by Jane Austen (1568)
Beyond Good and Evil by Friedrich Wilhelm Nietzsche (1543)
Thus Spake Zarathustra: A Book for All and None by Friedrich Wilhelm Nietzsche (1526)
The King James Version of the Bible by (1517)
Gulliver’s Travels into Several Remote Nations of the World by Jonathan Swift (1487)
Andersen’s Fairy Tales by H. C. Andersen (1477)
Winnie-the-Pooh by A. A. Milne (1468)
The Jungle Book by Rudyard Kipling (1465)
The silver blade: The true chronicle of a double mystery by Charles Edmonds Walk
The Call of the Wild by Jack London (1442)
Common Sense by Thomas Paine (1439)
The Turn of the Screw by Henry James (1425)
Persuasion by Jane Austen (1405)
The Awakening, and Selected Short Stories by Kate Chopin (1398)