ADELPHI
UNIVERSITY

Methods on the Effects of Sexism in Academic Literature

Lylybell Teran

Bachelor of Science in Mathematics and Computer Science

A thesis submitted in fulfilment of the requirements for the degree of

Bachelor of Science

2022

Department of Mathematics and Computer Science

Supervised by Dr. Carl Giuffre and Dr. Monica Morales Hernandez

Adelphi University

Abstract

Over 10.8 million workers in the U.S. are involved in STEM occupations yet merely 27% are filled by women. Despite the increasing demand for workers, the issue remains that women are far less likely to enter a career in STEM compared to their male counterparts. In this study, we explored the underlying influences on gender inequality by processing textual analysis on academic literature. Textbooks often are used in conjunction with lectures to provide students with a comprehensive understanding about a branch of study, which allows for future independent development. These foundational influences may establish long-lasting effects on self-identification, social, and cultural ideology that may differ between men and women. Textual analysis consisted of frequency calculations of consecutive sequences of words, called n-grams. An existing sentiment dictionary was used to evaluate the emotional bias of the text with respect to gender pronouns. This study attempts to search for subtle sexism in textbooks via textual analysis.

Contents

List of Tables

List of Figures

Acknowledgements

I want to thank my supervisors, Dr. Carl Giuffre and Dr. Monica Morales Hernandez, whose expertise was invaluable in formulating the research questions and methodology. Their insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. They provided me with the support to choose the right direction and successfully complete my thesis. Thank you all for your patient support and the opportunity to further my research.

Chapter 1

Introduction

1.1 Literature Review

Studies reveal that storybook reading involves more than a mere entertaining time for children. In 2019, researchers analyzed 34 preschoolers' ability to process information about our environment [1]. The children were first exposed to a book containing cartoon depictions of the moon and talking animals. They were then introduced to a more realistic story that involved people and accurate lunar images. After each reading session, the children were interviewed and asked to draw pictures of the moon. The fictional story resulted in 33 of the children describing and illustrating the moon as a circle that hides behinds clouds, which is a misconception based on cartoon fiction [1]. After reading the factual story, there was a noticeable increase in their ability to formulate more scientific questions and illustrate accurate depictions.

Concluding remarks stated that young children are capable of processing information by applying prior knowledge, cultural information, text, and visuals to make their own definitions [1]. In addition, the scientific book was rated just as amusing and interesting as the cartoon book. Thus, revealing that both young boys and girls display similar interests, which offers an opportunity to close the gender disproportionately in the sciences [1]. As a result, researchers highly recommend educators and parents to consider the social conceptual concepts that books contain, since they can lead children to misconceptions [1].

1.2 Previous Work

The influences of gender stereotyping have previously been analyzed in children's books. According to Kneeskern and Reeder, female characters are typically depicted as childcarers while males go out to work [2]. This stereotypes females as submissive and gentle in contrast to the dominant roles exhibited by males. As seen with the 34 preschoolers, children can also be misinformed when it comes to developing moral beliefs on each other [1]. In this research, children were assessed before and after exposure to multiple fiction books that were counter-stereotypical. The results showed that consistent exposure to these books led to a shift in their perspectives [2]. In comparison to pre-exposure examination, males displayed a more significant change of gender beliefs about stereotypes [2]. This research supports the concept that literature influences early adolescent perspectives on gender roles.

Recent text analysis on **Wikipedia plots dataset** also depicts the instilled gender stereotyping across 100,000 plots in films, novels, and video games. The results located words that aid in reinforcing gender roles [3]. The top 200 gender-associated verbs after 'he' or 'she' revealed contrasting differences. Words like "she screams" and "she cries" are 3x as likely to occur, which supports the idea that women are more likely to play roles as victims [3]. By contrast, men commonly play aggressors since they are 4x more likely to be accompanied by "he kidnaps" and "he beats" [3]. Additionally, 'he' was 4x more likely to be followed by "saves" and "rescues," which also paints males as stereotypical protagonists as confirmed by Kneeskern and Reeder's research [2]. In conclusion, Robinson determined that typical gender roles are enforced throughout fictional scripts that we consume on a daily basis.

Chapter 2

Methods

2.1 Motivation

Through a student's life, academic literature is constantly encountered through articles, autobiographies, essays, and textbooks since early years of primary education. These are years that are crucial in the development of a child where they establish physical, social, and intellectual capacities [4]. Advancement of personal studies is reliant on interest formed throughout primary, secondary, and sometimes higher education. Therefore, it is only natural to question the content that is encountered across educational stages.

2.2 Proposed Problem: Assessing Academic Literature

Women are severely underrepresented in STEM positions while men consist of the majority of the workforce. Previous studies have determined the influences of books in developing knowledge and interest on certain topics, which may vary between males and females. Thus academic literature may be crucial to early formation of self-identity within the scope of academia [1].

2.3 Methodology

In response to this problem, I produced a text-mining algorithm to compare the significance of words surrounding pronouns. I approached this via text analysis on R using the **tidytext** package. Sentiment value was extracted from verbs and adjectives surrounding feminine and masculine pronouns. The frequency and the sentiment score of a word was then used to calculate sentiment severity. The word comparison between female and male pronouns were then visualized for facilitated analysis. Further ANOVA testing was performed to determine statistical significance of gender-associated words on sentiment severity.

2.4 Evaluation Metrics

The sentiment severity of words was evaluated based on their respective sentiment value and the ratio of frequency.

- **Frequency:** a ratio of the specified word to the total number of words used with either feminine or masculine pronouns.

$$\frac{occurrence}{Total} \tag{2.1}$$

- **Severity:** calculation that defines the influence of a word based on its sentiment value and frequency to occur.

$$sentiment \times frequency \tag{2.2}$$

2.5 Assumptions

The following are assumptions I consistently maintained throughout the project.

- Although gender is non-binary we only considered male and female pronouns for simplicity and ignored neutral pronouns such as ‘they/them.’
- Only the top 50 popular names (proper pronouns) were considered.
- Literature was retrieved from only open-source sites and under biased selection (eliminates randomization).
- For literature dated pre-1960, dataset contained only articles of journals due to difficulty in retrieving full length textbooks.

Chapter 3

Initial Assessment: Fairy Tales

3.1 Research framework

Initially I analyzed popular fairy tale stories since previous studies have determined the influences of books and self-development amongst children. Fairy tales are commonly a child's first interaction with a fictional storybook. Thus I believed that fairy tales may play a crucial role in a child's formation of self-identity and life morals since many authors incorporate social and cultural concepts into their stories.

To assess fairy tale stories, I created a text-mining algorithms via two methods: creating text bigrams and creating GloVe word embeddings. We will discuss how the latter method demonstrated to be inefficient for sentiment analysis. However, the bigram method will be proven to successfully assess gender associated words in chapter 4.

3.2 Dataset

3.2.1 Fairy Tale Origins

Many fairy tale stories have been transformed to cater to children throughout the years. In the 1800s, the Grimm brothers from Germany published a collection of tales that consist of the majority of stories we read today [5]. Some of their works include "Hansel and Gretel," "Rumpelstiltskin," "Snow White and the Seven Dwarfs," etc.[5]. The authors did not write the stories with the intention of children reading them and were simply stories about children and families under difficult living conditions. These stories were meant to preserve traditions by encapsulating German morals within literature, which still hold today [5].

3.2.2 Data Pre-processing

The top 20 most popular children's fairy tales were compiled onto a text (**txt**) document consisting of romance, thriller, and magical short stories. Eleven of the stories were created by the Grimm brothers and a list of the fairy tale stories can be viewed on table 1. Initially we ignored proper nouns such as 'Jane' or 'John' to facilitate the text mining process. Instead we specified four different types of masculine and feminine pronouns: 'he', 'she', 'his', 'hers', 'him', 'her', 'himself', 'herself.' The specifics on pronoun types can be viewed on table 4. In addition we filtered common English stop words that were not significant in the project, which are highlighted in table 3.

3.3 Text-Mining Algorithm

As previously mentioned, I approached the text mining algorithm via two methods. We will discuss the failed attempt in creating GloVe word embeddings.

3.3.1 Failed Attempt: GloVe Word Embeddings & Text2vec

The GloVe algorithm collects words in the form of word co-ocurrence matrix X. Each element X_{ij} of the matrix represents how often word i appears in context of word j. By doing a simple normalization of the values for each row of the matrix, we can obtain the probability distribution of every context given a word. Using these probabilities, the relationship between words can be calculated by doing the ratio between the probabilities of the context given those two words. After pruning the data, I created a TCM matrix that was factorized via the **text2vec** algorithm, which is a parallel stochastic gradient descent algorithm for the model to learn two sets of word vectors - main and context.

To generate words vectors, we must specify related words. For instance, the following operation would result in a vector that is close to a vector for "queen."

$$vector(\text{“}king\text{”}) - vector(\text{“}man\text{”}) + vector(\text{“}woman\text{”})$$

This method generated word vectors along with a similarity value from the scale of 0.00 to 1.00. The greater the value, the more similar the words are used in context. Although this method located similar word vectors near male and female pronouns, it did not provide us with its significance in relation to sentiment value. In addition, most word vectors would not generate due to their lack of occurrence in the small dataset. Therefore this method was not used for further analysis. In order to proceed with the textual analysis, the text-mining algorithm we will discuss in chapter 4 was applied to the Fairy Tales dataset.

3.4 Sentiment Analysis

Table 3.1: Positive Word Frequencies With Their Sentiment Values

Word Frequencies			
word	**masculine**	**feminine**	**value**
beautiful	0.0064	0.0083	3
lovely	0.0088	0.0066	3
pretty	0.0052	0.0057	1
best	0.0076	0.0057	3
elegant	0.0052	0.0057	2

A sample indicating: positive words, frequencies to be either masculine or feminine, and sentiment value can be referred above in table 3.1. This figure supports the idea that female characters are likely to be described as “beautiful” and “pretty.” Similarly, since protagonists are likely to be played by male roles, they are nearly 2x more likely to display “best” or “lovely” behavioral traits.

Table 3.2: Negative Word Frequencies With Their Sentiment Values

Word Frequencies			
word	**masculine**	**feminine**	**value**
cried	0.0052	0.0057	-2
weeping	0.0052	0.0057	-3
tears	0.0052	0.0083	-2
corpse	0.0064	0.0048	-1
dead	0.0080	0.0049	-3

By contrast, in table 3.2 we have a sample of negative words and their frequencies throughout the dataset. Female roles tend to be portrayed as distressed characters alongside words such as “cried,” “weeping,” and “tears.” Although these words contain a negative sentiment value, they do not match the negativity value of some words found near male pronouns. In addition to men being portrayed with noble characteristics, they are also likely to die and thus are commonly followed by “corpse” and “dead”. According to Kneeskern and Reeder, this is a common scenario since heroes tend to die honorably at the end of a story [2].

3.5 Sentiment Severity

The contribution of a word i.e. the sentiment severity variable was determined using the equation 2.2, where the ‘masculine’ and ‘feminine’ frequency values were multiplied by their sentiment value for each word. This calculation allows for a better understanding of how significant the specified word plays for developing the characterization of gender in fairy tales.

Figure 3.1: Sentiment Severity for Males

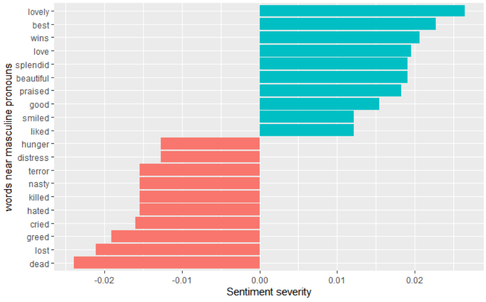
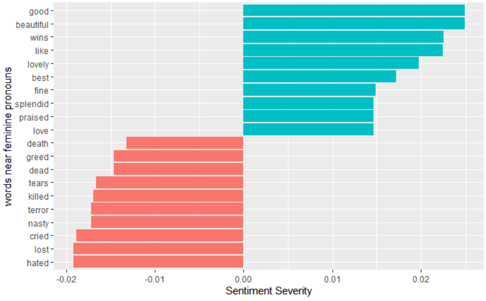


Figure 3.2: Sentiment Severity for Females



The figure located in the appendix 3.1 and 3.2 compares the sentiment severity of the 20 most contributing words for male and female roles respectively.

3.5.1 Discussion

At glance, figures 3.1 and 3.2 seem to display similar levels of severity for both male and female roles. Upon further analysis, the y-axis of the figures differ in verbs and adjectives, which help compose of gender roles. In addition, the sentiment severity for each word, with respect to gender, also varies. Overall, males displayed negative and positive words with slightly higher severity levels in comparison to that of word-associations for females.

Male Characterization

The most contributing positive words for male roles are “lovely” and “best” with a sentiment severity of over 0.02. As mentioned before, males often play protagonists with heroic and glorified personalities. For example, the dataset included stories such as “Rapunzel” and “The Sleeping Beauty”, which involve classic plots where a prince comes to save the princess in the tower. These princesses are portrayed as helpless and distressed until the heroic prince comes to the rescue. Additionally, the most contributing negative words for males are “dead” and “lost,” with a sentiment severity value of less than -0.02. In support of male leads, many die courageously for the sake others.

Female Characterization

The most contributing positive words for female roles are “good” and “beautiful”, with a sentiment severity of approximately 0.025. These words help illustrate women with grace and elegance which is typical with princesses like in “Snow White and the Seven Dwarfs” and “The Princess and the Pea.” Additionally, these characters tend to be more delicate and prone to distressed behaviour as seen by the words “cried” and “tears.” By contrast, what is *also* typical of plots is the characterization of female villains like Ursula from “The Little Mermaid.” This is supported by the frequent use of “hated” and “nasty” in the dataset.

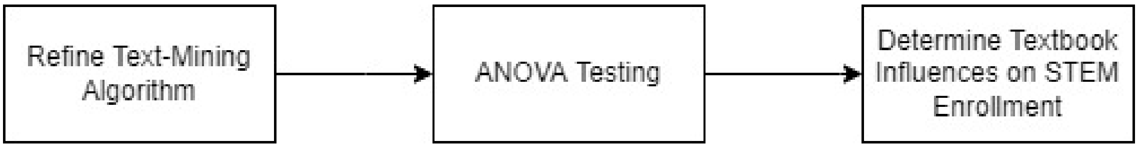
Self Characterization

The results show that fairy tales help establish gender bias, which could be affecting our children’s ability to self characterize themselves. If children are exposed to stereotypical beliefs during such a crucial stage of development, these ideas may influence an individual’s psychological features which cultivate identity [6]. For example, if a girl reads stories where woman are constantly depicted as weak and emotional, then they might develop low-self esteem. With such lack of confidence, it may be harder for woman to have a psychological sense of belonging to a rigorous and demanding field like STEM [6]. In the following chapter, I examined underlying influences within academic literature that may play a role in nurturing a STEM identity.

3.6 Summary

Through a text-mining algorithm, that I will discuss in-depth in the following chapter, I extracted sentiment and calculated its contribution within the top 20 fairy tale stories. The sentiment severity of words belonging to masculine and feminine pronouns supports typical characterizations of gender roles. The results also align with previous literature concerning stereotypical assertions in fiction that were written under different social and cultural normalities. My transition to academic literature was influenced by the sexist gender roles that were initially discovered in fairy tale stories. The figure below are steps that will be detailed in the folllowing chapter.

Figure 3.3: Phase II Steps


Refine Text-Mining Algorithm
ANOVA Testing
Determine Textbook Influences on STEM Enrollment

Chapter 4

Research Design

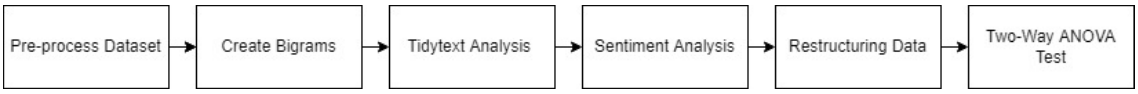
4.1 Research Objectives

A text-mining algorithm was applied to uncover stereotypes that establish gender roles within academic literature. Our objective was to determine if there are distinct differences in the content catered towards females versus males. We aimed to discover if exposure to sexist differences within academics, influences their interests in pursuing a STEM career.

4.2 Research framework

My text-mining algorithm locates adjectives and verbs that surround masculine and feminine pronouns. This was approached by creating text bigrams in preparation for sentiment analysis using the **AFINN** lexicon from the **tidytext** library. Further data restructuring was done to assess varibles such as the gender association of words, the book genre and publication year through two-way ANOVA testing. See the figure below for a summary of the steps:

Figure 4.1: Summary of Steps



4.3 Dataset

4.3.1 Academic Literature History

The world's first scientific journal called *Philosophical Transactions*, was published in 1665 by Henry Oldenburg. The popularity of this journal lead to the enforcement of systematic procedures such as expert peer review in order to maintain transparency and legitimacy in academic literature [7]. In 1970 the North American Feminists created a publishing house called the Feminist Press [8]. These movements led to the creation of guidelines, such as popular publisher McGraw Hill that established regulations for employees and collaborators to prepare non-sexist material. The changes are detailed in an 11-page set of guidelines "for equal treatment of the sexes." The guidelines were prepared by a 12-person editorial committee and distributed to McGraw-Hill's 8,000 authors and editorial employees [7].

Along with the rise of the internet, the early 1900's led to the trend of open access publishings specifically with respect to journals in the sciences. Open access literature is made available free for all on the web by the publisher at the time of publication [8]. Open source content permits their free use and re-purposing by others, which was enforced within the dataset.

4.3.2 Data Pre-processing

In order to analyze textual content, the dataset was re-structured before processing into the text analysis algorithm. Thirty open-source literature, consisting of 20 textbooks and 10 articles from journals, were converted from **PDf** to **TXT** files. The open-source websites included: OpenStax, ResearchGate, and Sci-Hub. The textbooks consisted of Natural Sciences subjects such as Mathematics, Physics, and Biology. Additionally, Social Sciences subjects like Humanities, History, and Business were also assessed. A detailed list of the literature can be viewed on table 2. To decrease run-time of the algorithm and facilitate analysis, we removed numbers and filtered common English stop words that were not significant in the project, which are highlighted in table 3.

4.3.3 Pronouns and Proper-Nouns

Four different types of masculine and feminine pronouns were instantiated: ‘he’, ‘she’, ‘his’, ‘hers’, ‘him’, ‘her’, ‘himself’, ‘herself’ and specifics on pronoun types can be viewed on table 4. To further the complexity of the algorithm, proper nouns were included into our text analysis of academic literature. Fifty of the most common female and male names were accounted for within a dictionary. These names were labeled as part of masculine or feminine pronouns accordingly and a list of the names can be viewed in table 5.

4.4 Text-Mining Algorithm

As previously mentioned, I approached this via two methods. We will detail the successful text-mining process using tidytext analysis via bigrams. As previously mentioned, the core process of this algorithm was also used to analyze fairy tale literature (with the exception of the mining of proper nouns).

4.4.1 Bigrams & Tidytext Analysis

In this method I generated text bigrams that were analyzed via the **tidytext** package. Using this package, I tokenized the text into bigrams which are consecutive pairs of words, which can be seen in the source code located in figure 1. Tokenization is a process that facilitates data handling by turning text into ‘tidy data’ that consists of the following structure:

- Each variable is a column
- Each observation is a row

For tidy text mining, the token that is stored in each row is most often a single word, in this project I instead stored them into an n-gram where n was set to 2. This evaluates pairs of two consecutive words, often called “bigrams.” For instance, “she cries” would store them into two columns that consists of the pronoun ‘she’ and the verb ‘cries.’ Thus, I filtered for cases where a word occurred after each pronouns such as ‘he’, ‘her’, etc.

In order to determine the frequency of a word, I used equation 2.1 for each respective pronoun. The respective source code for each pronoun calculation is available in figure 2. For example below is the equation for ‘he’ where the numerator is the occurrence of the *specified* word plus a constant 1. The denominator uses the sum operation in **R** to add the total number of times that words appear after ‘he’ and adds a constant of 1. The following calculates the frequency of a certain word following ‘he’:

$$he = \frac{(he+1)}{sum(he+1)}$$

After calculating the tendency for every word in the dataset, I was left with 58 columns and 11,481 rows. The columns consisted of the eight pronoun tendencies and the 50 most common proper nouns, along with the word it is followed by. The bigram method also extracted 11,481 rows each being a single word like an adjective or verb. For simplicity, a sample containing words that occur at least 20 times throughout the dataset can be viewed in the appendix 7.

4.5 Skewed Distribution of Gender-Associated Words

Initial assessment included a frequency calculation of words in order to indicated if sexist undertones exist within academic literature. The figures below demonstrate a simple analysis of the occurrence of gender-associated words.

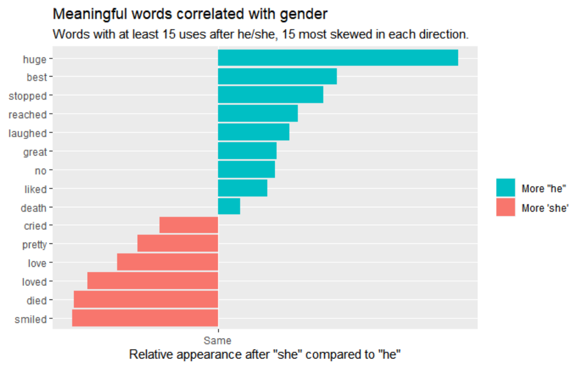


Figure 4.2: Distribution Based On Frequency

Based on figure 4.2, some words such as ”best” and ”great” appear 4x more frequently after masculine pronouns. Feminine pronouns are often followed by words such as ”pretty” and cried.” These gender stereotypes were also prevalent in fairy-tales stories, which furthered our initiative to pursue sentiment analysis followed by a statistical study.

4.6 Sentiment Analysis

To reduce the dimensionality of the dataset, I added the female and male frequencies to create ‘feminine’ and ‘masculine’ columns for each respective words. Then I extracted sentiment from all the words using the pre-established **AFINN** lexicon. This lexicon has an integer rating between -5 and +5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. Since the lexicon only has sentiment assigned to specified words, the number of rows (words) significantly decreased from 11,481 to 898.

4.7 Sentiment Severity

Following the same sentiment severity calculation from chapter 3, this value allows for a better understanding of how significant the specified word plays for developing the characterization of gender within literature. Figures 4.3 and 4.4 compare the sentiment severity of the 20 most contributing words for male and females roles in academic literature, respectively.

Figure 4.3: Sentiment Severity for Males

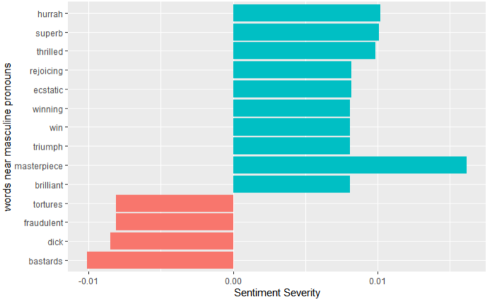
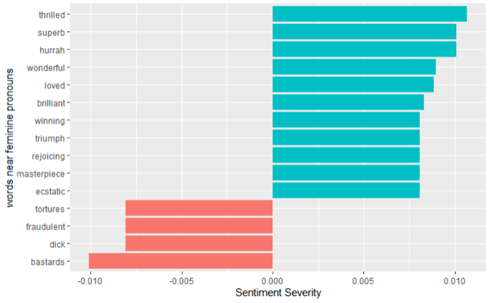


Figure 4.4: Sentiment Severity for Females



The figures above depict the same verbs/adjectives and similar levels of sentiment severity. Some words like thrilled and ecstatic are mere synonyms that create the characterization of gender within academic literature. These words fail to suggest sexist bias contrary to the sentiment severity analyzed within fairy tales. Figures 4.3 and 4.4 show that the most contributing words do not consist of sexist narratives.

4.8 Two-Way ANOVA Testing

A two-way ANOVA test is a statistical test used to determine the effect of two independent variables on a dependent variable. In our research, the response variable was sentiment severity and the independent variables were book subject, publication year, and gender association of the word. ANOVA testing results in F-tests to determine whether two populations with normal distributions share variances or a standard deviation, on the significance of the regression formula overall. Two tests were conducted on all positive and all negative words. Figures 3 and 4 in the appendix demonstrate source code used to conduct ANOVA testing. The results of the two-way ANOVA tests on all positive and negative words are located in the figures below.

Table 4.1: Positive Words

ANOVA Variables		
Variable	**F-Value**	**P-Value**
Gender	0.3085	0.5787
Post 1960	0.1159	0.7336
Hard vs. Soft	0.1926	0.6609

Table 4.2: Negative Words

ANOVA Variables		
Variable	**F-Value**	**P-Value**
Gender	0.1850	0.6672
Post 1960	0.4008	0.5269
Hard vs. Soft	2.8854	0.0898

The alpha-value, also known as the level of significance, was set to 0.05. In both the positive and negative words, the p-values were significantly greater than 0.05 for each independent variable. This indicates that gender, year, and subject are not statistically significant with respect to sentiment severity. In other words, none of the independent variables influenced sentiment severity of words associated with gender pronouns. Additionally, the publication year had did not affect use of words despite the regulations added to avoid existing sexists undertones in academic literature.

4.8.1 Discussion

Figures 4.3 and 4.4 display similar levels of sentiment severity for the same words with respect to *both* male and female characterization. The lack of gender differentiation was further supported by the ANOVA test results in table 4.1 and 4.2. The variables analyzed were not significant enough to differentiate gender characterization within academic literature.

Chapter 5

Conclusion

5.1 Summary

By creating a text-mining algorithm via tidytext analysis, I extracted sentiment and calculated its contribution within 30 academic literature. This content included natural and social sciences textbooks from 1940 to 2022. Initial discovery of the skewed occurrences of gender associated words, as seen in figure 4.2, established our research initiative to proceed. However, two-way ANOVA tests reveal that the variables described in table 6 are not statistically significant. Therefore, the results support the null hypothesis that the variables such as book subject, publication year, and gender association of the word do not influence the sentiment severity of words within academic literature. As a result, the research was inconclusive in determining textual influences on STEM interests between female and male genders.

5.2 Limitations

The sentiment analysis approach in the previous chapter simply counted the appearance of positive or negative words, according to a reference lexicon. Therefore a word's context can matter nearly as much as its presence, which can be problematic. For example, the words “happy” and “like” will be counted as positive, even in a sentence like “I'm not happy and I don't like it!” This may explain some of the words that did not fit the expected gender stereotyping.

Additionally, many textbooks include graphics to help students understand a concept. I did not consider the characterization of gender that is illustrated for a student's comprehension. Based on Altun's research proving misconception development in preschoolers after being exposed to picture books, this variable certainly affects adolescent's perspectives. Furthermore, acquiring text from open-source websites was limiting and instilled bias since many came from a common publisher. An increase in the dataset, beyond open-source, would provide an array of characterizations that may help uncover sexist bias as seen in previous research.

5.3 Future Work

As previously mentioned, the **AFINN** lexicon significantly reduced the number of words that were mined, since the existing dictionary only accounts for 2,750 words. Further development or expansion of the lexicon is needed to account for more descriptive words that may be vital to the results of this research. Complementary to the development of the lexicon, I plan on increasing the n-gram tidy text analysis to consider more words that may influence the meaning of a sentence. This will improve the overall context of a sentence and thus solidify characterization of gender within academic literature. The influences of the disproportion of gender in STEM enrollment may be determined with improvements of the text-mining algorithm.

Appendix

Initial Dataset

Table 1: Top 20 Fairy Tales

Fairy Tale Stories
Little Red Riding Hood
The Three Little Pigs
The Gingerbread Man
Hansel and Gretel
The Ugly Duckling
Snow White and the Seven Dwarfs
Beauty and the Beast
Cinderella
Jack and the Beanstalk
Pinocchio
Rumpelstiltskin
Sleeping Beauty
The Tale of Peter Rabbit
Goldilocks and the Three Bears
The Little Mermaid
The Pied Piper of Hamlin
The Frog Prince
Rapunzel
The Princess and the Pea
The Emperor’s New Suit

Updated Dataset

Table 2: Open-Source Academic Literature

Textbooks and Articles
A Guide to Naturalist
American Government
American Historian
AP Biology
AP Physics
Astronomy
Biology Laboratories
Business and Benevolence
Calculus
Calculus I
Career and Technical Education
Chemistry in the French Enlightenment
Climate Science
Economic Geography
How Arguments Work
Introduction to American Literature 1865 to Present
Introduction to Business
Introduction to Anthropology
Key to Humanization
Mental Growth and Personality Development
Physics in Pre-Nazi Germany
Principles of Management
Science and Human Survival
Textbooks and School Curricula
The Human Biology Council
The Humanities and Education for Humanity
Transistor Physics
Women Writing on Writing Women
World History to 1648
Writing Guide

Stop-Words

Table 3: Insignificant Stop Words Removed

English Stop Words	
a	you
the	I
on	at
with	is
it	by
to	when
and	will
in	the
was	did
as	that
how	had
can	for
of	if
but	so
not	what
would	could

Pronoun Specifics

Table 4: Pronouns Analyzed

Pronouns				
Gender	**Subjective**	**Objective**	**Possessive**	**Reflexive**
Masculine	He	Him	His	Himself
Feminine	She	Her	Hers	Herself

Proper Noun Specifics

Table 5: Top 50 Most Common Proper Nouns

Proper Nouns	
Male	**Female**
James	Mary
Robert	Patricia
John	Jennifer
Michael	Linda
William	Elizabeth
David	Barbara
Richard	Susan
Joseph	Jessica
Thomas	Sarah
Charles	Karen
Christopher	Nancy
Daniel	Lisa
Matthew	Betty
Anthony	Margaret
Mark	Sandra
Donald	Ashley
Steven	Kimberly
Paul	Emily
Andrew	Donna
Joshua	Michelle

Variables Assessed

Table 6: Variables Analyzed

ANOVA Test Variables		
Variable	**Levels**	**Description**
Gender	Male or Female	gender association of word
Post 1960	True or False	book was published before or after the feminist movement
Hard vs. Soft	Hard or Soft	book is a natural or social science

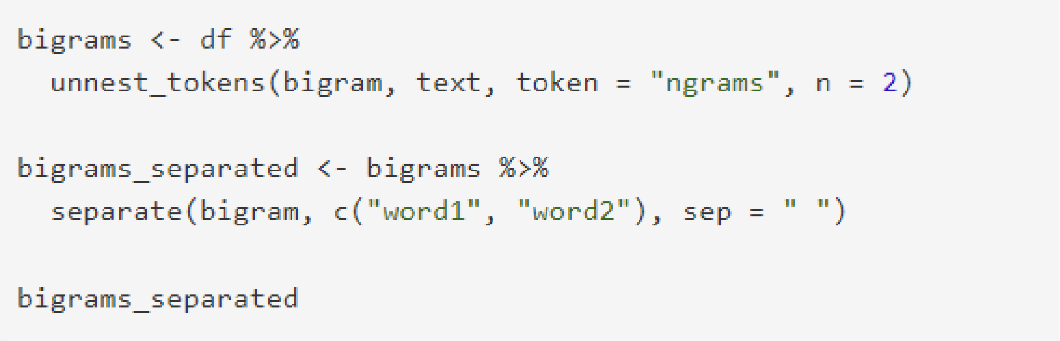
Sample of Word Frequencies: Academic Literature

Table 7: Word Frequencies For Each Pronoun

Word Frequencies								
word	**he**	**him**	**his**	**himself**	**she**	**her**	**hers**	**herself**
said	0.0264	0.0014	0.0012	0.0016	0.0183	0.0009	0.0017	0.0016
came	0.0264	0.0014	0.0012	0.0033	0.0099	0.0009	0.0017	0.0016
eyes	0.0009	0.0014	0.0110	0.0016	0.0008	0.0111	0.0017	0.0016
thought	0.0123	0.0014	0.0012	0.0016	0.0137	0.0009	0.0017	0.0016
went	0.0141	0.0014	0.0012	0.0016	0.0160	0.0009	0.0017	0.0016
head	0.0009	0.0014	0.0097	0.0016	0.0007	0.0128	0.0017	0.0016
saw	0.0160	0.0014	0.0012	0.0016	0.0236	0.0009	0.0017	0.0016
knew	0.0057	0.0014	0.0012	0.0016	0.0130	0.0009	0.0017	0.0016
heart	0.0009	0.0014	0.0049	0.0016	0.0008	0.0154	0.0017	0.0016

Bigram Text-Mining Source Code

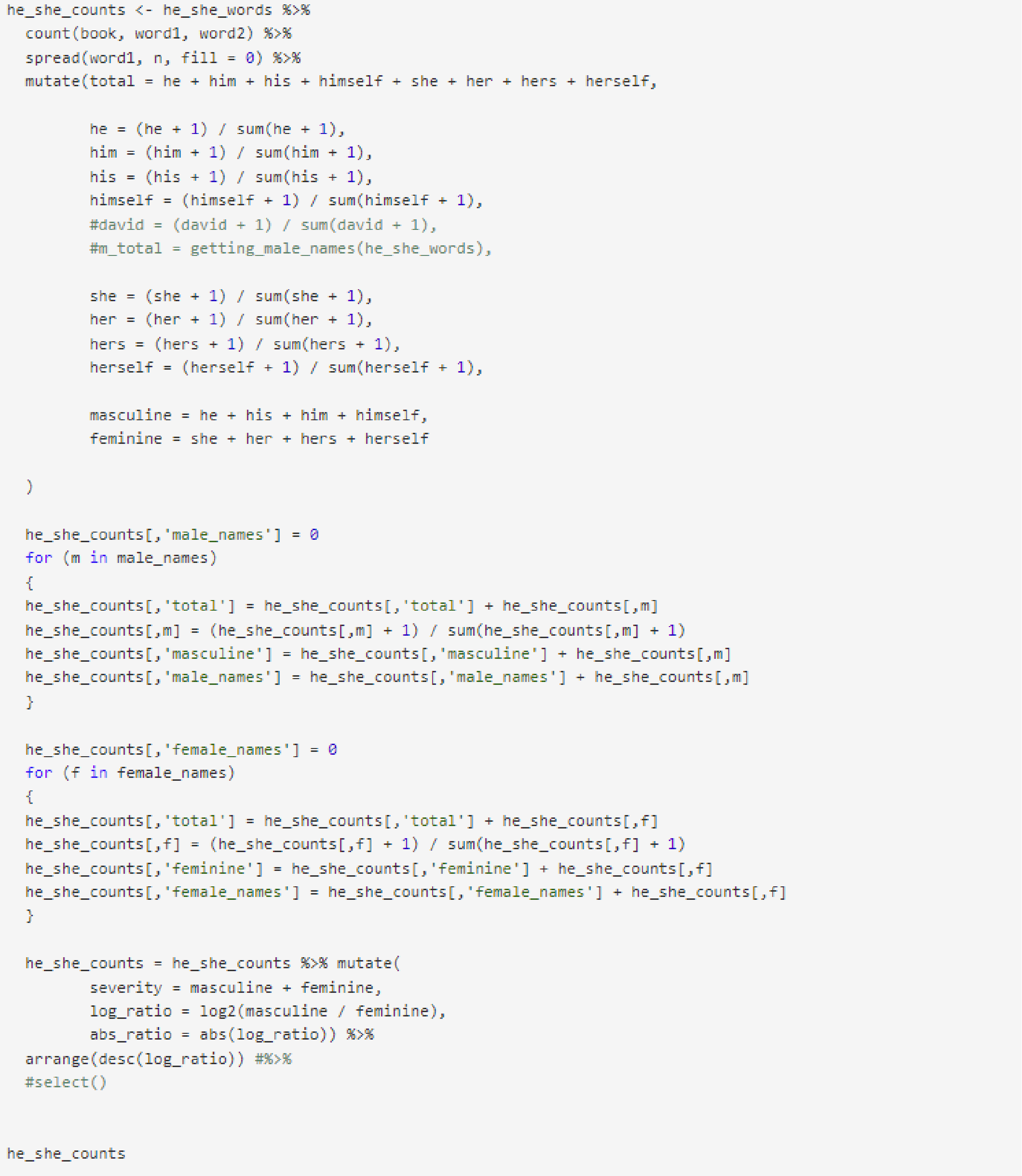
Figure 1: Bigram Separation Source Code



```
bigrams <- df %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)

bigrams_separated <- bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_separated
```

Pronoun Calculations Source Code

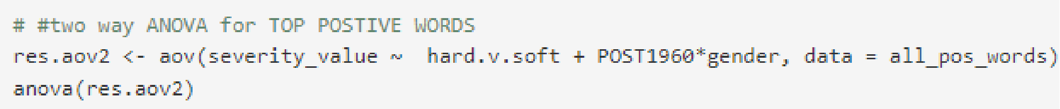
Figure 2: Pronoun Calculations Source Code



```
he_she_counts <- he_she_words %>%
  count(book, word1, word2) %>%
  spread(word1, n, fill = 0) %>%
  mutate(total = he + him + his + himself + she + her + hers + herself,

         he = (he + 1) / sum(he + 1),
         him = (him + 1) / sum(him + 1),
         his = (his + 1) / sum(his + 1),
         himself = (himself + 1) / sum(himself + 1),
         #david = (david + 1) / sum(david + 1),
         #m_total = getting_male_names(he_she_words),

         she = (she + 1) / sum(she + 1),
         her = (her + 1) / sum(her + 1),
         hers = (hers + 1) / sum(hers + 1),
         herself = (herself + 1) / sum(herself + 1),

         masculine = he + his + him + himself,
         feminine = she + her + hers + herself

  )

  he_she_counts[,'male_names'] = 0
  for (m in male_names)
  {
  he_she_counts[,'total'] = he_she_counts[,'total'] + he_she_counts[,m]
  he_she_counts[,m] = (he_she_counts[,m] + 1) / sum(he_she_counts[,m] + 1)
  he_she_counts[,'masculine'] = he_she_counts[,'masculine'] + he_she_counts[,m]
  he_she_counts[,'male_names'] = he_she_counts[,'male_names'] + he_she_counts[,m]
  }

  he_she_counts[,'female_names'] = 0
  for (f in female_names)
  {
  he_she_counts[,'total'] = he_she_counts[,'total'] + he_she_counts[,f]
  he_she_counts[,f] = (he_she_counts[,f] + 1) / sum(he_she_counts[,f] + 1)
  he_she_counts[,'feminine'] = he_she_counts[,'feminine'] + he_she_counts[,f]
  he_she_counts[,'female_names'] = he_she_counts[,'female_names'] + he_she_counts[,f]
  }

  he_she_counts = he_she_counts %>% mutate(
         severity = masculine + feminine,
         log_ratio = log2(masculine / feminine),
         abs_ratio = abs(log_ratio)) %>%
  arrange(desc(log_ratio)) #%>%
  #select()


he_she_counts
```

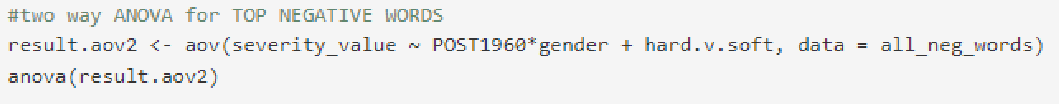
ANOVA Testing Source Code

Figure 3: ANOVA Test Source Code for All Positive Words



```
# #two way ANOVA for TOP POSTIVE WORDS
res.aov2 <- aov(severity_value ~  hard.v.soft + POST1960*gender, data = all_pos_words)
anova(res.aov2)
```

Figure 4: ANOVA Test Source Code for All Negative Words



```
#two way ANOVA for TOP NEGATIVE WORDS
result.aov2 <- aov(severity_value ~ POST1960*gender + hard.v.soft, data = all_neg_words)
anova(result.aov2)
```

Bibliography

[1] D. Altun, "From story to science: The contribution of reading fiction and hybrid stories to conceptual change with young children," *Children & Society*, vol. 33, no. 5, pp. 453–470, 2019.

[2] P. A. Ellen E., "Examining the impact of fiction literature on children's gender stereotypes," *Current Psychology*, vol. 23, 03 2020.

[3] D. Robinson, "Gender and verbs across 100,000 stories: a tidy analysis," Apr 2017.

[4] D. Holloway, L. Green, and S. Livingstone, *Zero to Eight. Young Children and Their Internet Use*. 01 2013.

[5] J. Zipes, D. Heitman, and L. W. Scanlan, "How the grimm brothers saved the fairy tale."

[6] A. Y. Kim, G. M. Sinatra, and V. Seyranian, "Developing a stem identity among young women: A social identity perspective," *Review of Educational Research*, vol. 88, no. 4, pp. 589–625, 2018.

[7] J. McDougall-Waters, "History of philosophical transactions," Nov 2014.

[8] "Academic publishing," Mar 2022.