Exploring semantic differences in book reviews on female-led and male-led Young Adult novels

Anonymous ACL submission

Abstract

This project explores the semantic differences in reader response to Young Adult (YA) novels with female versus male protagonists. More specifically, we apply the SemAxis framework to Goodreads reviews, comparing readers' associations between stereotypically gendered descriptors and personal referents. Overall, we find no strong gender bias.

1 Introduction

On the topic of gender bias, the words we use speak volumes. In an analysis of performance reviews, for instance, men were most commonly praised for being "analytical" and "competent" compared to their female counterparts, who were commended for being "compassionate" and "enthusiastic" (Smith et al., 2018).

While all four of these adjectives are undeniably approbatory, each set creates a narrative that pigeonholes individuals into certain ways of being. As a result, we see real-world effects in our society, such as interviewers preferring resumes with male names over identical ones with female names (Steinpreis et al., 1999). To counter these biases, one proposition is to increase representation in the media we consume, such as literature (Bal and Veltkamp, 2013). Indeed, if children learn gender roles from a young age—internally adjusting their behaviours to align with externally perceived norms—then promoting examples of multidimensional characters of all genders broadens our cultural horizons (Bem, 1983).

In this project, we analyze how readers respond to characters in Young Adult (YA) literature, a genre targeted at those in their formative years. By comparing the word associations in reviews of novels with female and male protagonists, we explore whether these books contribute to countering or reinforcing gender norms.

For all our code, see our Github repo.

2 Related work

Many before us have performed similar work centering literature and gender bias. Though most of these analyzed children's books (Brugeilles et al., 2002), plays (Copnehaver, 2002), and pre-modern fiction (Nagaraj and Kejriwal, 2022b), the last decade saw YA literature's emergence as an object of study with the genre's boom in popularity.

Citing its ability to re-form worldviews, some researchers have employed content analysis to find that YA protagonists demonstrate traits opposite their gender's stereotypes (Kimsey, 2019), while others identified stereotype-strengthening messages of female objectification (Brown, 2013). More recently, Natural Language Processing (NLP) techniques and AI bore distant reading approaches, resulting in insights such as male characters being four times more prevalent than their female counterparts and more associated with terms like "strong" and "power" (Nagaraj and Kejriwal, 2022a).

Though our project shares similar aims with the ones above, ours differs in that we focus not on published bodies of literature, but on readers' reactions to them. Our method, furthermore, considers word associations across a wide variety of text, and not just raw word counts in specific works. We thus take a step back from representational differences in fiction to instead consider how these character archetypes are perceived more generally, which in turn influences stereotypes in our culture today.

3 Method

To compare the language used to describe female-led versus male-led novels, we employ the SemAxis framework (An et al., 2018) on reviews originally shared on Goodreads, a popular social cataloguing website.

3.1 SemAxis

SemAxis is a lightweight framework designed for characterizing domain-specific semantics beyond sentiment. Using an embedding model and custom-defined semantic axes, this approach allows us to compare the language and implicit sentiments across a wide array of book reviews. The framework consists of three main parts: 1) constructing word embeddings, 2) defining semantic axes, and 3) mapping words onto these axes.

3.1.1 Constructing word embeddings

Word embeddings provide a mode for distilling large amounts of data into analyzable representations (word vectors). In our implementation, we use the standard Word2vec Skip-Gram model, which predicts surrounding words given the current word and is known to perform well on semantic tasks (Mikolov et al., 2013, Lucy et al., 2020). For our purposes, we create 3 different word embeddings: one trained on reviews of female-led novels, one trained on reviews of male-led novels, and one trained on all the reviews in our dataset. These separate embeddings account for the fact that many reviews do not explicitly reference the protagonist, but instead use pronouns that might refer to other characters or even the author.

3.1.2 Defining semantic axes

The next step in SemAxis is defining semantic axes, which are the vectors between select pole words. These pole words typically comprise sets of antonym pairs (Lucy et al., 2022) like “strong” and “ weak,” but can also be similar words with different connotations like “brave” and “reckless.” Either way, the purpose here is to select pairs of words that carry different meanings and to then calculate the vectors between their embedding representations. As outlined by An et al., 2018, the words vectors describing positive and negative pole words can be understood as $S^+ = \{v_1^+, v_2^+, \ldots, v_n^+\}$ and $S^- = \{v_1^-, v_2^-, \ldots, v_m^-\}$, and the vectors describing the pole words can be understood as $V^+ = \frac{1}{n}\sum_1^n v_i^+$ and $V^- = \frac{1}{m}\sum_1^m v_j^-$. The semantic axis from S^+ to S^- is thus

$$V_{\text{axis}} = \mathbf{V}^+ - \mathbf{V}^-$$

For our project, we define semantic axes by first establishing a list of 16 core descriptors wherein half are stereotypically female and the other half are stereotypically male (table 1). Our selection was informed by the Bem Sex-Role Inventory (BSRI), a list of socially masculine and feminine terms compiled by renowned psychologist Sandra Bem (Bem, 1974), as employed by similar studies in section 2.

Female descriptors	playful, supportive, organized, reserved, caring, optimistic, flexible, selfless
Male descriptors	decisive, aspiring, fearless, logical, tough, confident, charming, intelligent

Table 1: The 16 core descriptors used in our project

For each of the core 16 descriptors, we create one set of pole words with its antonym, and another set of pole words with a synonym that has negative connotations. An example of the former is “playful” and “serious”; an example of the latter is “playful” and “immature.” Through these antonymous and connotative axes, we are interested in which types of words reviewers use, and whether this changes based on the protagonist’s gender.

In total, we create 32 sets of pole words (one antonymous and one connotative for each of the 16 core descriptors), which yields 96 semantic axes, since we train three separate word embeddings.

3.1.3 Projecting words onto semantic axes

Next, we project certain words onto the semantic axes to distinguish their relation to said axes:

$$\text{score}(w)_{\mathbf{V}_{\text{axis}}} = \cos\left(v_w, \mathbf{V}_{\text{axis}}\right) = \frac{v_w \cdot \mathbf{V}_{\text{axis}}}{\|v_w\| \, \|\mathbf{V}_{\text{axis}}\|}$$

If a word has a positive score, then it is more associated with the positive pole word in the axis (and vice versa). In our work, we project words from a set of female, male, and gender-neutral protagonist references (table 2). The third set is used on the gender-specific embeddings, helping us target language applied to the protagonist. To obtain more robust projection scores, we record the average value across each set of references.

Female references	she, her, hers, woman, girl, heroine
Male references	he, him, his, man, boy, hero
Gender-neutral references	protagonist, mc, main, character

Table 2: The sets of words we project onto semantic axes

3.2 Dataset

To obtain reader reviews, we use the Young Adult subsection of the Goodreads Book Graph Dataset (Wan and McAuley, 2018). Scraped in 2017 for content recommendation research, this dataset comprises 2,389,900 instances in which a user reviewed a YA novel on the popular site Goodreads, where each record includes features like book ID, user ID, text-based review, and a rating out of 5. Though no arbiter of choice by any means, Goodreads provides a helpful picture of audience opinion insofar as readers are willing to engage with books beyond simply reading them.

To refine our data for better processing time, we first isolated for the novels with a minimum of 10,000 reviews, as demarcated in the dataset's provided meta-data. We then identified the gender of each novel's main protagonist(s) as follows:

1. To ChatGPT, input the list of novels and ask "For each novel, what is the gender of the main protagonist(s)? Please label them as 'F', 'M', or 'V' if various"

2. For each novel, cross-check ChatGPT's label using the relevant Goodreads summary or personal knowledge (i.e., if the authors read the book)

By combining computer-generated and human-sourced labels, we separate our data into reviews on female-led novels, male-led novels, and novels led by both or neither with a high degree of confidence. This process bore a total of 325,095 reviews covering 63 novels.

	Reviews	**Novels**
Female-led	206,528	36
Male-led	82,887	19
Various	35,680	8
Total	325,095	63

Table 3: Initial human-annotated dataset

	Reviews
Female-led	523,011
Male-led	145,343
Total	668,354

Table 4: Classifier-labelled supplemented dataset

After preliminary tests with the data in table 3, we noticed that there were not enough reviews on male-led novels to create accurate embeddings. We therefore extended this initial dataset by adding high-confidence samples labelled by Sklearn's Multi-layer Perceptron classifier on the rest of the data from Wan and McAuley. This model performed with 98.17% accuracy on 120k instances of training data and 40k instances of testing data (with no book IDs overlapping between the two). By extracting all reviews classified with confidence $\geq$ 0.99, we obtain the supplemented dataset in table 4.

3.3 Preprocessing

To prepare the raw reviews for the word embeddings, we converted all text to lowercase and removed numbers and punctuation, with the exception of hyphens to keep compound words and adjectives (e.g., "ice-cream" and "red-spotted"). We opted not to remove stop words to preserve pronouns such as "she" and "his," which we project onto the semantic axes. Finally, we tokenized each review into its component words using NLTK.

4 Results

4.1 Qualitative evaluation of embeddings

Since no domain-specific analogy test exists for our dataset, we verified our embeddings by manually checking nearest neighbouring words, as done in similar domain-specific evaluation cases (Lassner et al., 2023, Tilbury, 2018).

In the word embeddings for our entire dataset, for instance, the closest word to "weak" is "bella-like," reflecting common associations of *Twilight*'s Bella Swan with frailty. In the embeddings for reviews on female-led novels, one of the closest words to "tough" is "ladyballs," an older slang term for female demonstrations of socially male courage. Our models, in short, successfully learned the domain-specific language of the Goodreads community. Having extended our dataset to improve the male-led novel review embeddings (table 4) and verfied that the closest words to our 16 core descriptors (table 1) and referent words (table 2) made sense, we proceed with high confidence in our embeddings' ability to reasonably capture semantics.

4.2 Axis results

In this section, we project the referent words from table 2 onto our semantic axes to evaluate their associations in the reviews. Note that in all graphs, the negative pole words (the ones to the left of the hyphen) in the top 8 axes are the stereotypical female descriptors from table 1 whereas the next half are the stereotypical male ones.

We begin by exploring, separately, the embeddings for the reviews on female-led novels, and those representing reviews on male-led novels. Onto their respective semantic axes, we project gender-neutral protagonist references, such as "protagonist" and "mc" (a popular shorthand for "main character"). By taking the average projection value of these references for each semantic axis, we explore the different descriptors used by readers when reviewing female-led versus male-led novels.

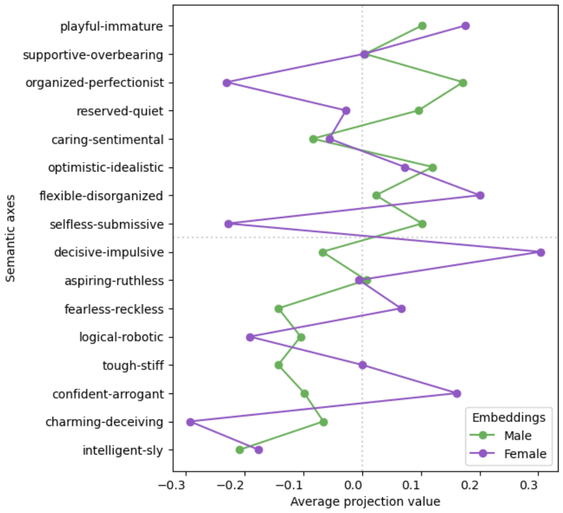

Figure 1: Gender-neutral protagonist references projected onto connotation axes

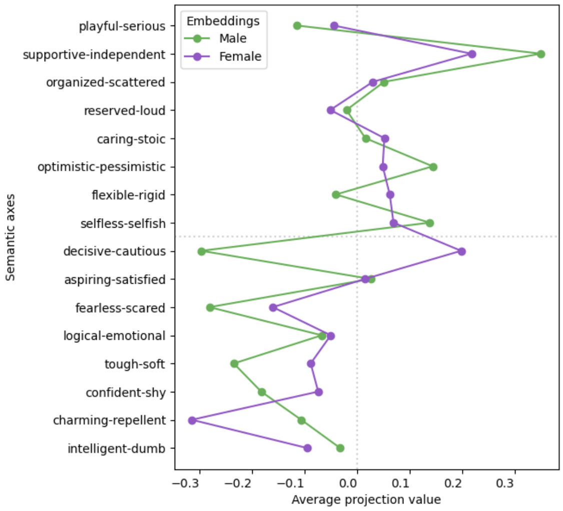

Figure 2: Gender-neutral protagonist references projected onto antonym axes

Interestingly, fig. 1 and fig. 2 display a weak decreasing trend, with the projection values drifting from positive to negative as one moves down the semantic axes list. This is especially apparent for the male embedding projections: for the top 8 axes, the average value (of the average projection values) is 0.7 in both graphs, compared to -0.10 and -0.15 on the latter 8 axes for fig. 1 and fig. 2, respectively.

Recall that the top 8 axes in all graphs have stereotypical female descriptors as their negative pole words, and the next 8 have stereotypical male descriptors as their negative pole words. Aside from certain outliers, then, this decreasing trend suggests that overall, reviews of male-led novels associate character referents with stereotypical male descriptors and the antonyms and negative connotative synonyms of stereotypical female descriptors. To test whether readers indeed perceive characters as stereotypically male but not stereotypically female, we next project male, female, and gender-neutral protagonist references onto the semantic axes of our entire dataset's embeddings.

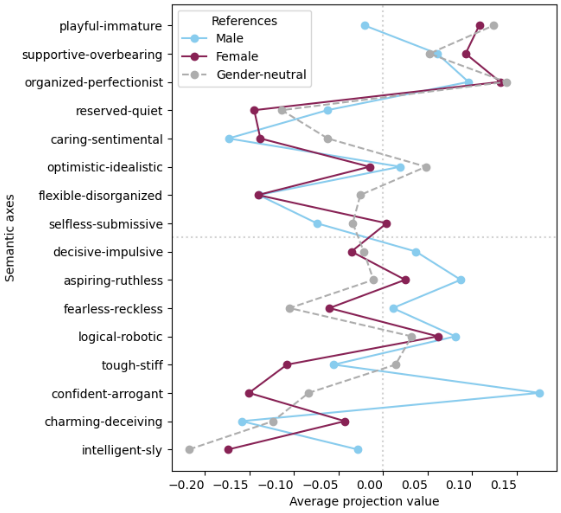

Figure 3: All references projected onto connotation axes using full dataset embeddings

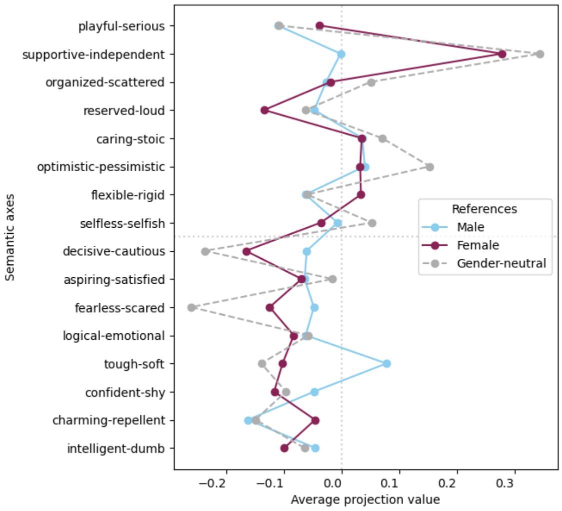

Figure 4: All references projected onto antonym axes using full dataset embeddings

In fig. 3 and fig. 4, there are no decreasing trends. However, one sees that most projection values are weakly negative. On the connotation axes, the average of all female referent projections is -0.04 compared to -0.01 for male projections; on the antonym axes, the average of all female projections is -0.03 compared to -0.05 for all male projections. In other words, all referents tend to be more associated with the sixteen stereotypical female and male descriptors, though very weakly so.

5 Discussion

5.1 Findings

Overall, we find no strong gender bias in the word associations of reviews on female-led and male-led novels. In fig. 1 and fig. 2, there is only a weak decreasing trend wherein gender-neutral protagonist referents are more associated with stereotypical male descriptors and non-stereotypical female descriptors. Even so, this was most apparent in the male embeddings, which were created using less review data than the female embeddings.

In fig. 3 and fig. 4, by contrast, the only visible trend is slightly negative projection values in general, with most clustered around 0.0. That said, certain outliers seem to counter gender biases on a micro scale. Among the connotative axes, for example, male references are much more associated with "arrogant" than female references, which are more associated with "confident." Similarly, among the antonymous axes, female references are much more associated with "independent" than male references, which demonstrate a neutral average projection value.

This might speak to the idea that YA authors are increasingly crafting female protagonists who are intelligent, driven, and confident (Peterson, 1996), or that readers are increasingly perceiving female characters as such. Nevertheless, these outliers are tempered by groups of axes which contribute to stereotypical notions, such as the protagonist referent projections on the male embeddings in fig. 2 being closer to "decisive," "fearless," and "tough." Given these varying observations, we cannot definitively assert strong gender bias in our experiments.

5.2 Limitations

There are several limitations which may have impeded our results.

First and foremost, we were restricted by limited processing power, resulting in a dataset that could definitely be extended further. The male-led novel review embeddings, for example, were created with much less reviews than the female-led novel review embeddings. By using Yarowksky's algorithm, for instance, one could perform bootstrapping to create a larger dataset.

Furthermore, one could obtain more robust, statistically significant results from embedding-models trained on permutations of the data, as in Antoniak and Mimno, 2018 and Liu et al., 2021.

We could also have been more thorough in isolating for language pertaining to the novels' protagonists. Because our dataset includes reviews that do not explicitly mention the protagonist, the gendered referents projected on the full-dataset embeddings (e.g., "she" and "his") may capture references to other individuals. Though we rationalize this with the idea that readers' perceptions (and, subsequently, their reviews) are likely coloured or influenced by the gender of the protagonist through which they consume the narrative, we could have taken a more thorough route, such as by using NLP techniques to filter for descriptors directly applied to the protagonist in both name and referring pronoun.

6 Conclusion

In this study, we continue prior work investigating gender biases in relation to YA literature, but through readers' responses rather than in the works themselves. Following the novel SemAxis framework, we define several semantic axes by pairing stereotypical female and male descriptors with their antonyms and negatively connoted synonyms. By projecting female, male, and gender-neutral protagonist referents onto these axes, we explore readers' associations between gendered references and stereotypes.

Overall, we find no strong gender biases in our analysis. Though certain semantic axes seem to even counter gender stereotypes at a micro level, these outliers, along with the weak trends perceived, are insufficient to definitively assert overall bias either way.

In the future, this process can be improved by incorporating contextualized word embeddings, increasing training set size while pruning low-quality reviews, and implementing bootstrapping.

7 Statement of contributions

Our work was generally divided as follows:

- Lucas: made functions to analyze embeddings using SemAxis, extended dataset using high confidence classifier, created embeddings on dataset using Word2vec, wrote scripts to train and load data on HPC

- Faith: wrote scripts to load and pre-process data; created semantic axes based on academic research; plotted projection graphs; investigated related work; edited spelling, grammar, and flow throughout entire paper

- Alex: analyzed dataset breakdown, plotted dataset graphs, validated semantic axes using closest words in embedding, formatted paper to match ACL 2023 Proceedings template

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

PM Bal and Martijn Veltkamp. 2013. How does fiction reading influence empathy? an experimental investigation on the role of emotional transportation. *PloS ONE*.

Sandra Bem. 1974. Bem sex role inventory. *APA PsycTests*.

Sandra Bem. 1983. Gender schema theory and its implications for child development: Raising genderaschematic children in a gender-schematic society. *Signs: Journal of Women in Culture and Society*, 8.

Candy Brown. 2013. Gender stereotyping in contemporary bestselling, young adult fiction books.

Carole Brugeilles, Isabelle Cromer, and Sylvie Cromer. 2002. Male and female characters in illustrated children's books. *Population*, 57.

Bonny Copnehaver. 2002. A portrayal of gender and a description of gender roles in selected american modern and postmodern plays. *Electronic Theses and Dissertations*.

Christina Kimsey. 2019. Gender bias and stereotypes in dystopian young adult literature.

David Lassner, Stephanie Brandl, Anne Baillot, and Shinichi Nakajima. 2023. Domain-Specific Word Embeddings with Structure Prediction. *Transactions of the Association for Computational Linguistics*, 11:320–335.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. Statistically significant detection of semantic shifts using contextual word embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas u.s. history textbooks. *AERA Open*, 6.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 3477–3494.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Akarsh Nagaraj and Mayank Kejriwal. 2022a. Dataset for studying gender disparity in english literary texts. *Data in Brief*.

Akarsh Nagaraj and Mayank Kejriwal. 2022b. Robust quantification of gender disparity in pre-modern english literature using natural language processing.

Janet Peterson. 1996. Gender bias and stereotyping in young adult literature. *Children's Book and Media Review*, 17.

David Smith, Judith Rosenstein, Margaret Nikolov, and Darby Chaney. 2018. The power of language: Gender, status, and agency in performance evaluations. *Sex Roles*, 80.

Rhea Steinpreis, Katie Anders, and Dawn Ritzke. 1999. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 42.

Kyle Tilbury. 2018. Word embeddings for domain specific semantic relatedness.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.