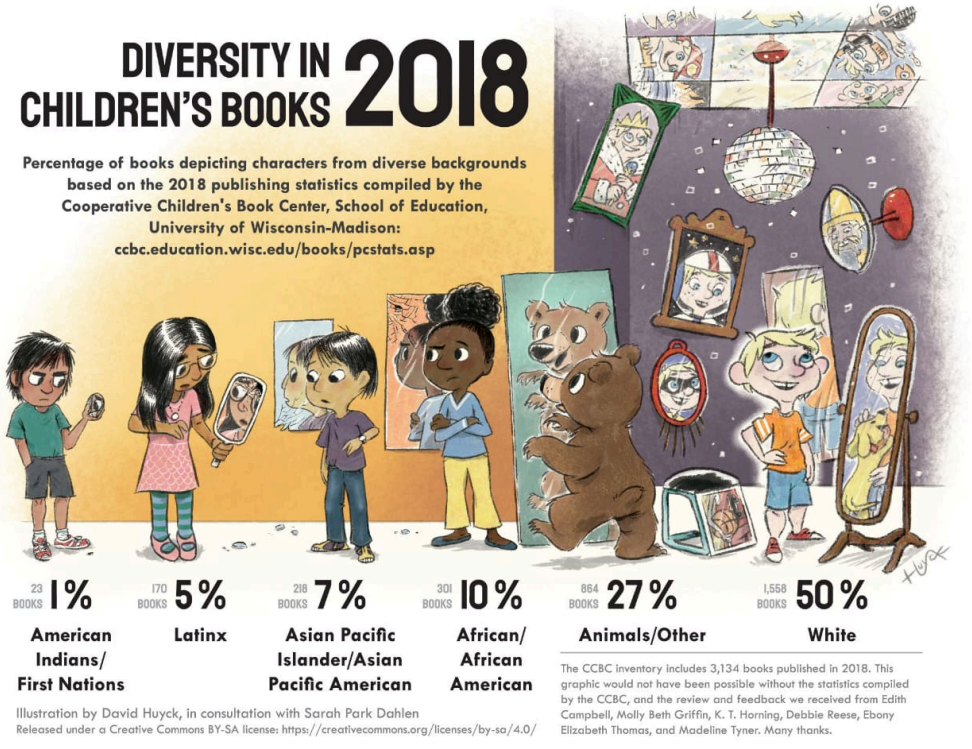
Text Analysis of Children's Books: Exploring DEI Themes Across Age Groups

Beatriz Radtke, Yulia Savine

I. Introduction

In the 1990s, Dr. Rudine Sims Bishop famously said that children's books must serve as "mirrors, windows, and sliding glass doors." Children must see stories that reflect their own experiences as well as those that expose them to communities other than their own in books that they read. The stories they read can serve to further stereotypes, or they can help children create positive mental models of their identities. So, now over 30 years later, does children's literature reflect this need, and how does it do so?



[1]

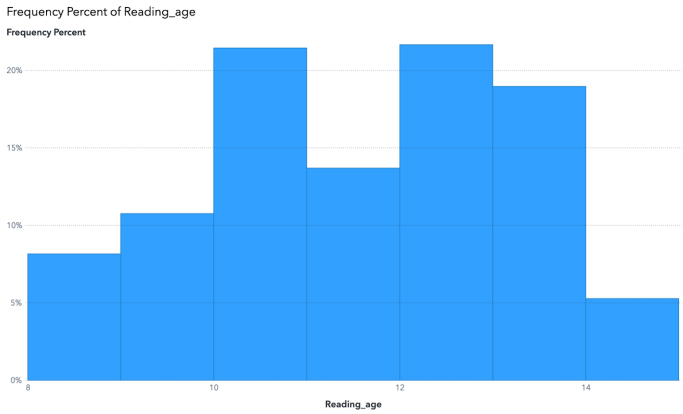
We note in *Image 1* that in 2018, children's literature was dominated by White and Animal characters, leaving children of color and of non-dominant cultures perhaps sidelined. This demonstrates a need for more diverse children's books, and, beyond

that, children's books that show positive examples of inclusivity, with characters children can look up to. ([1] https://socialjusticebooks.org/diversity-graphic/)

II. Data

For our data, we used the children's stories csv of Kaggle's of Highly rated Children Books And Stories dataset which includes 3269 children's books, their descriptions, and a reading/interest age. The data was sourced from Book Trust UK (www.booktrust.org.uk) which boasts over 8245 book entries across age levels.

Potential points of interest to explore in this dataset included gender representation, ethnic and cultural references, sentiment patterns towards marginalized groups, stereotyping languages, inclusive terminology, equity-oriented vocabulary, and intersections of identity. We are primarily focusing on gender themes and correlates with reading level / interest age.



Reading and interest ages for our data were centered around the 10-14 range, though with a good selection of documents for all groups.

Importantly, we are making the assumption that the description of a book is a reflection of the content. We also note that the age level recommendation might not be accurate to all readers. It may also be challenging to navigate the subjectivity of these analyses.

III. Methods

We are exploring the key differences in frequency and approaches to DEI themes across reading levels.

We began by cleaning our data. Notably, we removed table headers and modified age and interest level data to be numeric (e.g. changed “10-14” to 10).

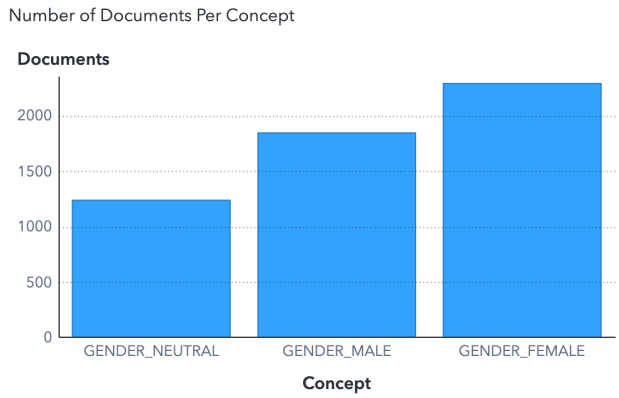
We then used the SAS Viya software to run CLASSIFIER, CONCEPT, C_CONCEPT, and CONCEPT_RULE rules on our data to extract varying elements of interest.

We proceeded with further analysis by creating graphs including word clouds and bar graphs to analyze vocabulary and to compare usage across genders.

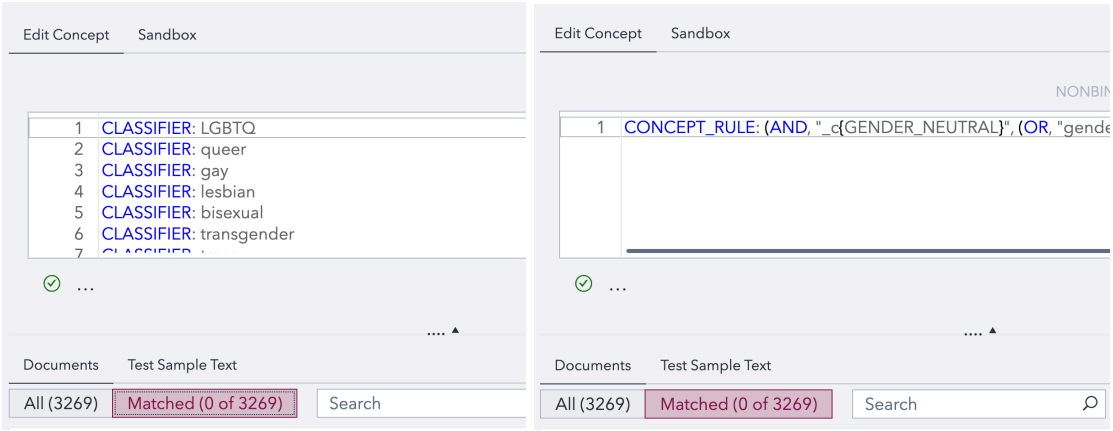
IV. Results, Analysis, and Discussion

A. Representation

We began by creating classifier rules GENDER_MALE, GENDER_FEMALE, and GENDER_NEUTRAL.



In order to attempt to understand whether our GENDER_NEUTRAL rule was extracting singular or plural they, we also created two rules: LGBTQ_REPRESENTATION and NONBINARY_REPRESENTATION



As we can see in the figures above, the two rules returned no matches across all of the book descriptions. We can thus hypothesize that there are no matches for 'they/them/theirs' in the context of a gender expansive character.

B. Stereotyping

On a broader scale, we thought it would be interesting to see which adjectives and action verbs were most common between male and female characters. We also hoped to explore non-binary or gender fluid characters, but were not able to find entries that reflected this type of representation. Additionally, we do not know if the descriptions of the books accurately captured information about all of the characters that would allow us to fully understand the gender representation over all characters.

GENDER VERBS

Figure: Word Cloud for C_CONCEPT: GENDER_FEMALE _c{:V}

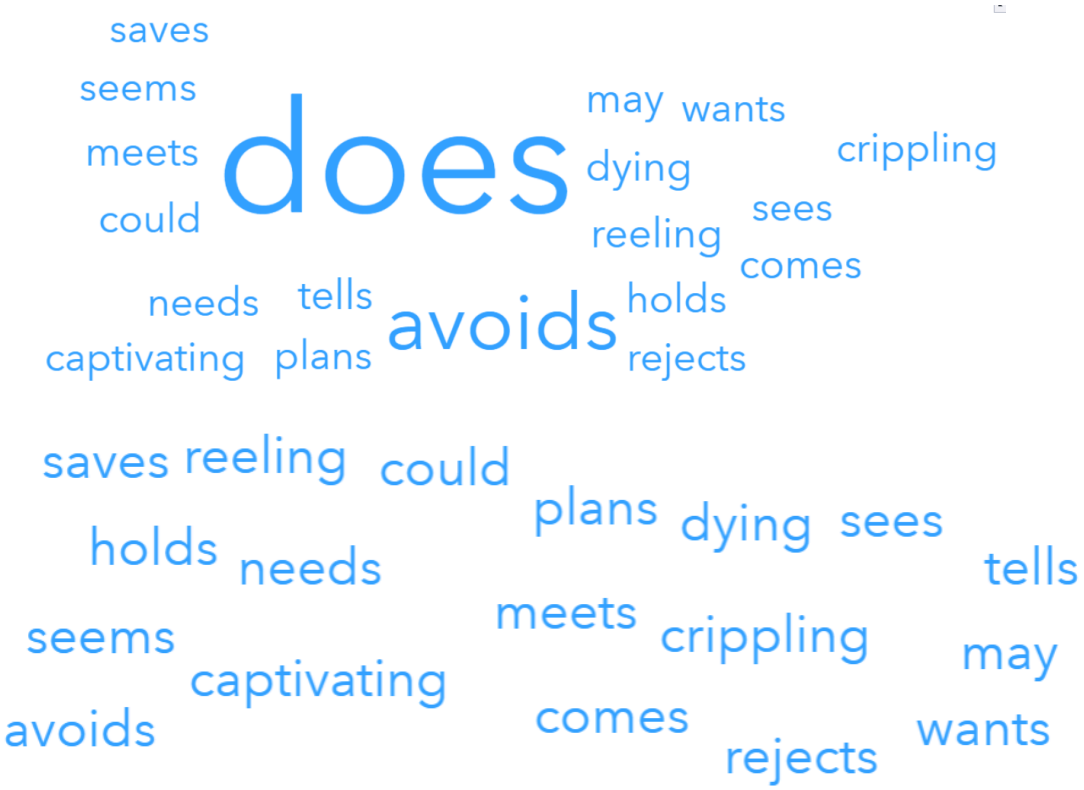


Figure: Word Cloud for C_CONCEPT: GENDER_MALE _c{:V}

**Note: In the processing of this data for both FEMALE and MALE -associated verbs we removed verbs such as "is" and "has" to focus on action words which may provide more insight into how female vs. male characters are represented in texts. Kept "does" in male due to very high frequency which I thought was interesting, and overall there was much lower variation in GENDER_MALE action verbs in context.*

GENDER ADJECTIVES

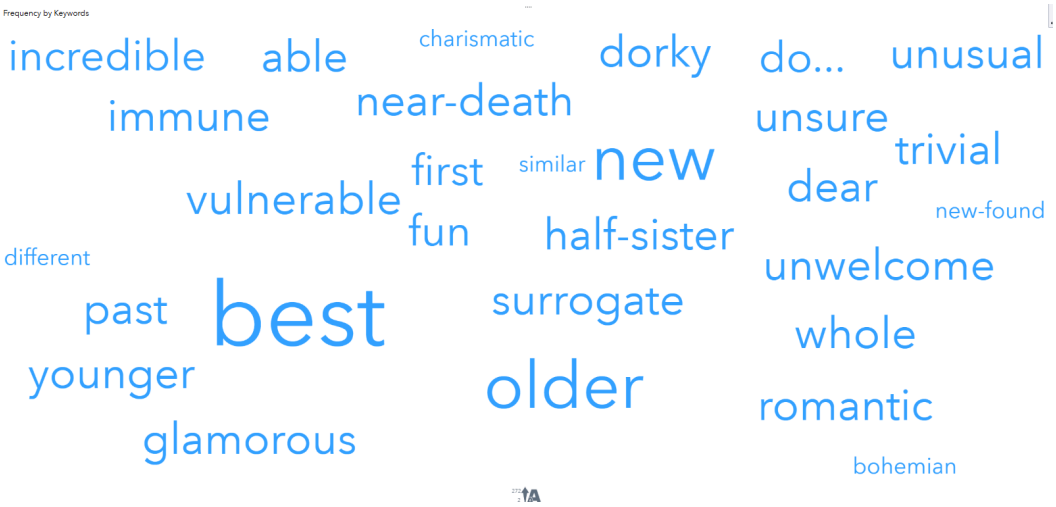
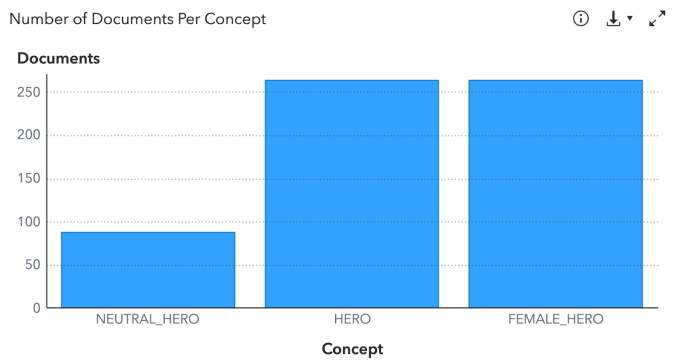


Figure: Word Cloud for CONCEPT_RULE: (DIST_2, "GENDER_FEMALE", "_c{:A}")
**Note: In the processing of this data we removed "Martian" and "snowball-throwing" as there were high match words.*

Figure: Word Cloud for CONCEPT_RULE: (DIST_2, "GENDER_MALE", "_c{:A}")
**Note: In the processing of this data we removed "Martian" and "snowball-throwing" as there were high match words.*

HERO

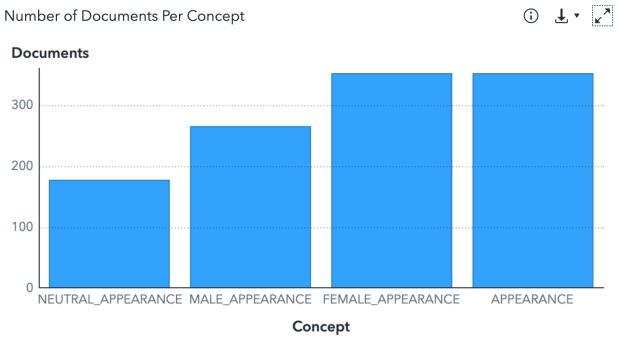
We wanted to examine the frequency of correlation between gender pronouns and mentions of "hero" or "heroine" or any expansion of these words. We thus created FEMALE_HERO, MALE_HERO, and NEUTRAL_HERO concepts.

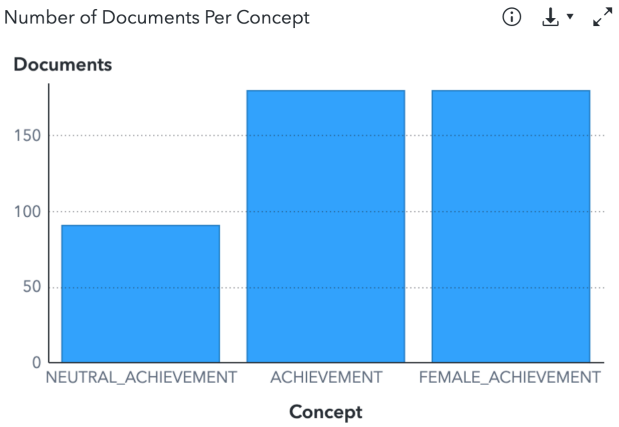


**Note: MALE_HERO exists, there were simply no matches for the rule*

ACHIEVEMENT, APPEARANCE

We thought it would be interesting to see whether there is a divide between achievement versus appearance focus in our data based on gender. We created appropriate concepts and ended up with the results in the tables below:

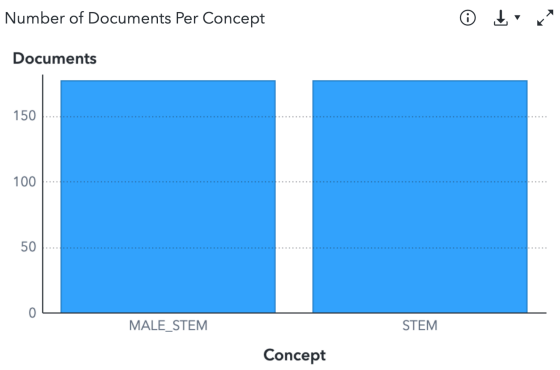




**Note: MALE_ACHIEVEMENT exists, there were simply no matches for the rule*

STEM

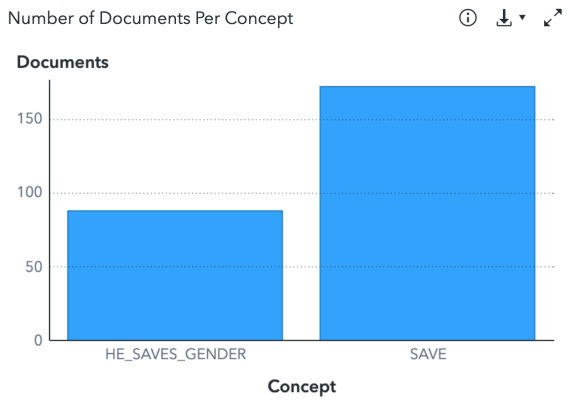
The gender divide in STEM is one that is also interesting to us as computer science majors. We created a CLASSIFIER rule that pulled STEM vocabulary then crossed that with references to gender pronouns.



**Note: FEMALE_STEM and NEUTRAL_STEM exists, but there were no matches for these rules*

SAVES

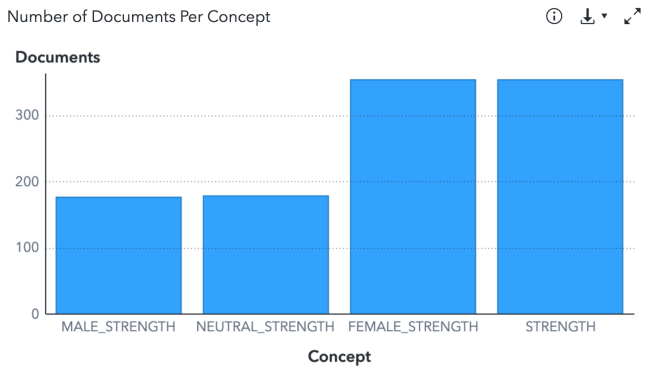
Another trope in stories is that male genders are often the ones doing the saving. We wanted to see if that was true in our data so we created a SAVE rule to pull expansions of the word “save” and looked for sentences of the form “[pronoun] save@ [pronoun].”



**Note: SHE_SAVES_GENDER and THEY_SAVE_GENDER both exist, but there were no matches for these rules*

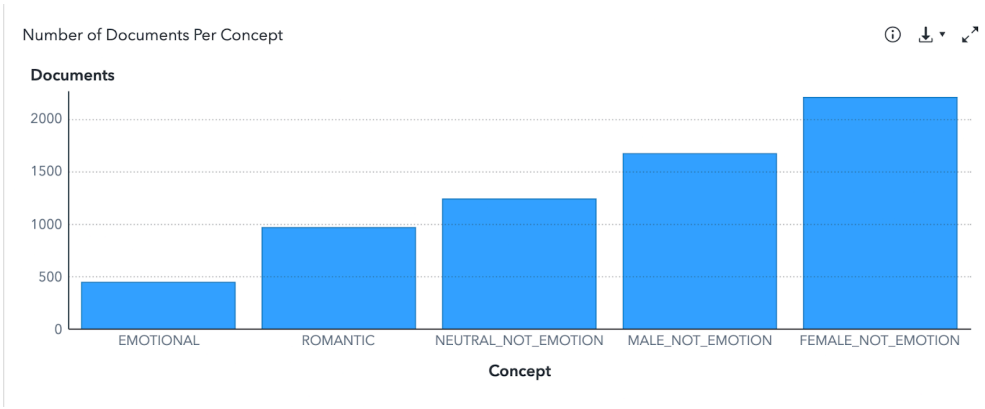
STRENGTH

We also looked for any trends in the pronouns used in conjunction with vocabulary that hinted at strength, both moral and physical.



EMOTIONAL, ROMANTIC

Finally, we looked for similarities between the gender pronouns used in association with vocabulary we defined as EMOTIONAL or ROMANTIC.



We used negation in these rules to pull the number of documents that were associated with a gender pronoun but did not have any mention of vocabulary we defined in rules EMOTIONAL and ROMANTIC.

C. Tone Differences when Discussing Gender/DEI Concepts

Tone differences blah blah

V. Limitations

One significant limitation is the ambiguity of neutral pronouns in English, particularly 'they/them/theirs'. These pronouns are often used in plural contexts, which may skew our data analysis. Our current dataset lacks explicit mentions of

non-binary genders, potentially leading to an overrepresentation of plural 'they' usage rather than singular gender-neutral applications.

Additionally, when analyzing Diversity, Equity, and Inclusion (DEI) themes, especially concerning stereotypes, we face two important constraints:

a) Time limitations: The depth and breadth of our analysis are restricted by the time available for the study.

b) Conceptual boundaries: Our analysis is inherently limited to the themes and stereotypes we can identify and articulate. This leaves room for undetected patterns or concepts within the data that may be significant but remain unexplored in our analysis.

VI. Future Work

It may be of interest to apply these rules on entire texts from children's books. Additionally, considering correlates like acceptance in schools and parent/teacher reviews of books may provide insight into the DEI patterns correlates with popularity, how they are valued among different demographics.

A deeper dive into this dataset could also reveal further information about gendered themes and language throughout. Of note would be to address some of the limitations noted above. It could be interesting to create more robust rules to search for gender expansive ‘they’ or to expand the LGBTQ representation search to more subtle expressions of the theme.

It would also be beneficial to look into other DEI categories such as race, religion, ability, culture, … to understand how those are characterized in books that our children are reading. It could also be interesting to apply all of these same rules and analyses to datasets that are meant to be more diverse or inclusive.