Do gender representations in textbooks reflect bias? Building a pipeline to analyze gendered person mentions in educational texts

Doruk Arisoy
University of Washington, ECE
arisod@uw.edu

Yuling Gu
University of Washington, ECE
yulinggu@uw.edu

Abstract

The way that characters are depicted in textbooks can influence the stereotypes formed by students reading them. In this work, we build a pipeline using various natural language processing (NLP) tools to allow for nuanced, automatic annotation of texts targeted towards the study of gender representations in educational texts. Our pipeline supports the detection of gendered mentions and entities, differentiating between real-world and hypothetical entities, as well as between named and generic entities. Using this pipeline, our analysis shed light on the different ways that female and male representations in textbooks reflect gender bias.

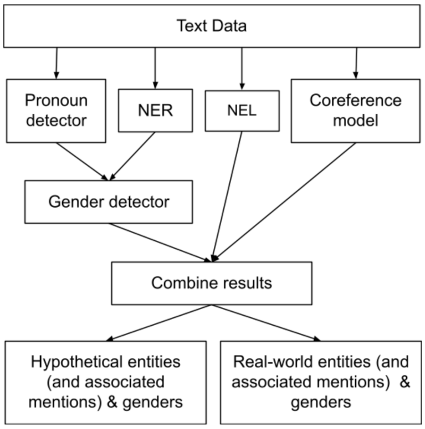

Text Data
Pronoun detector
NER
NEL
Coreference model
Gender detector
Combine results
Hypothetical entities (and associated mentions) & genders
Real-world entities (and associated mentions) & genders

Figure 1: Overview of our pipeline architecture.

1 Introduction

Women are not only underrepresented compared to men in Science, Technology, Engineering, and Mathematics (STEM) fields (Wang and Degol, 2017), studies like Leslie et al. (2015) show that such under-representation of women exists across the academic spectrum. One potential factor associated with this gender gap is the implicit messages conveyed through educational texts, which in turn affect students' developments in their fields.

Educational text, such as text in textbooks, often feature characters whenever a scenario or story is set up to convey the material in a contextualized manner. The choice of characters and how they are portrayed influences the students' impression of them, which in turn influences how students form stereotypes (e.g. gender stereotypes). The possibility that such depictions can involve stereotyped representations is a serious concern because negative stereotypes often results in diminished confidence, poor performance, and loss of interest (Rydell et al., 2010), having a detrimental effect on the students' academic performance.

According to the Global Education Monitoring Report by UNESCO (United Nations Educational and Organization, 2020), females are under-represented in textbooks across many countries, perpetuating stereotypes though omission. Through omission, it may leave the impression that female characters are less important. Many articles report insights on how this trend is found in different textbooks for instance, in economics (Flaherty, 2018) and history (Chiponda and Wassermann, 2011). However, existing research in the educational field looking into potential biases in textbooks often rely on expensive expert annotation and are thus limited in scale.

In this work, we conduct more scalable investigation in this direction using NLP tools. Our work investigates if there are gender biases in the portrayal of characters in both science and non-science textbooks by comparing male and female characters for the 1) number of times each character is mentioned and 2) number of characters discussed. We then focus on hypothetical characters to analyze authors' choices over 1) whether to use a male or female hypothetical character and 2) whether to give the character a name or present him/her as a generic character.

To allow for such analysis, we automatically annotate text in textbooks by introducing a pipeline (in Section 3) that builds upon different open-source NLP tools and utilizes a custom-designed algorithm to combine output from different components (Figure 1). With this, we can obtain person mentions and entities[1] with their associated genders, as well as differentiate real-world and hypothetical entities for our analysis. We evaluated different components of our pipeline with hand-annotated data (in Section 4) to compare its performance against existing baseline where possible, and guide our design choices. In Section 5, using the output from our pipeline, we present detailed analysis on different ways of how representation of females and males in textbooks reflect gender-bias.

This paper has two main contributions. First, we introduce a pipeline for automatically annotating textbooks to allow for analysis of potential gender bias in text at a large scale. Second, our analysis shed light on how the choice of characters in textbooks and the way they are portrayed can perpetuate negative gender-stereotypes – that females are represented as less important because they appear less frequently in textbooks, and even when authors do choose to use a female character, it is often portrayed as a generic "passer-by".

2 Data

The data used in this work comprises open source textbooks and articles for K-12 (Michigan, 2014; Siyavula, 2014; CK12, 2007) (compiled for linguistic complexity classification in Nadeem and Ostendorf (2018)). This includes both text from science and non-science textbooks, with grades ranging from elementary school to high school as detailed in Table 1. There are $33,575$ samples where each sample is a variable length test question or passage from a textbook. Some samples contain a single sentence while others take the form of a multi-sentence paragraph, giving a total of $178,054$ sentences. Figures 2 and 3 provide examples of such sample. Each sample has a science or non-science label referring to the type of textbook it was extracted from.

3 Approach

For each sample, our pipeline (Figure 1) 1) detects person mentions and 2) their associated genders, 3) determines if the person mention corresponds to a real-world person in a database, and 4) groups mentions referring to the same person into a cluster (i.e an "entity"). We then use a custom algorithm to combine the output of these sub-tasks. For our analysis, we retain only gendered mentions and the final stage of our pipeline separates entities into real-world and hypothetical ones.

Science textbook	
Grade	**Number of textbooks**
Elementary school	10
Middle school	23
High school	8

Non-science textbook	
Grade	**Number of textbooks**
K-8	9 (history)
High school	2 (history and economics)

Table 1: Grades and subjects of the textbook data used.

3.1 Detecting Person Mentions

We used spaCy's (Honnibal and Montani, 2017)[2] name-entity-recognition (NER) model to detect named person mentions. We detected other gendered mentions using hand-crafted lists[3] of pronouns and nouns for male, female and neutral genders ("*Pronoun detector*" in Figure 1).

3.2 Getting Genders of Mentions

Gendered pronouns and nouns are labeled male, female or neutral based on which list they were found in. To detect the gender of a named person mention, the Gender Guesser Python package which utilizes a database of names with associated gender (Michael, 2007) was used [4].

3.3 Searching the DBpedia database

We perform named entity linking (NEL) to map person mentions in text to real-world entities, using the dbpedia-spotlight (Olieman, 2017) tool. For each named person mention, the NEL component rank candidate resources in DBpedia (Auer et al., 2007) according to the similarity score between the context vector for this mention and context vectors for the candidates. Cosine similarity is used and the similarity score threshold (above which a real-world entity is considered to be found) is a hyperparameter (see section 4's evaluations).

3.4 Coreference Resolution

To identify which mentions refer to the same person, we experimented with two coreference resolution models: 1) NeuralCoref (Clark and Manning,

[1] For the purposes of discussion in this paper, we use the term "entity" to refer to each unique character discussed, where there can be multiple mentions associated one entity. For instance, in the example in Figure 3, there are four person mentions associated with the one entity (which is "Darwin").

[2] Available at `https://spacy.io/`

[3] Based on Nadeem (2020)

[4] If the model was unsure of the gender, for example of a gender-neutral first name, we interpreted it as neutral

2016) by Huggingface (Wolf et al., 2020) and 2) AllenNLP's (Gardner et al., 2017) coreference resolution model that uses SpanBERT embeddings (Joshi et al., 2019). Based evaluations in section 4, we used NeuralCoref in our final pipeline. These models output clusters of mentions, where each cluster contains mentions in the text that refer to the same entity. However, if a person is only mentioned once and not referenced again, the coreference component does not pick that person up. In such cases, we rely on the NER and pronoun detector components to get all person mentions.

3.5 A combined pipeline where parts complement each other

Finally, we combine outputs from all components using Algorithm 1 to generate the pipeline output. We first get all the coreference clusters, person mentions along with their genders and which mentions are in the database. For each coreference cluster, if any mention within the cluster is found in the database, we add that cluster with its gender to the real-world person set, otherwise the cluster and its gender are added to the hypothetical set.

One advantage of this pipeline is that errors in the output of a component can be potentially fixed by output from another component. For example, if for gender-neutral name, the Gender Guesser will mark its gender as neutral. However, if that person is mentioned later with a gender-specific pronoun, then our pipeline labels all of the references for that person with that gender. We elaborate more on such corrections in evaluations (subsection 4.3).

3.6 Example input-output

Given example input, "A driver is stopped at a stoplight. He waits for the light to turn green. The driver looks to his right. He sees an interesting store. Honk, honk, is the sound heard from behind. Why would another driver beep their horn?", the pipeline will output that the mentions "*A driver*", "*He*", "*The driver*", "*his*" and "*He*" all refer to the same hypothetical entity and this is a male entity.

4 Evaluation on pipeline components

4.1 Evaluation on identifying female/male/neutral person mentions

We first evaluate the pronoun detector, gender detector and NER components of our pipeline. Our evaluation examines how well our pipeline detects person mentions and the associated gender (female, male or neutral) in comparison to a previous work examining gender representations in educational text. Specifically, we refer to the approach used in Nadeem (2020) as our baseline.

Algorithm 1: Combining components/sub-tasks

```
Input: $sample$
Output: real person mentions and their genders,
        hypothetical person mentions and their
        genders
$coref$ = run $sample$ through coreference model
$person\_mentions$ = run sample through NER model
  and pronoun detector
$genders$ = get genders of all $person\_mentions$
$known\_people$ = run sample through NEL model
initialize $real$, $real\_genders$, $hypothetical$,
  $hypothetical\_genders$
for every $mention$ in $person\_mentions$ do
    $gender$ = $genders[mention]$
    if $gender$ is neutral then
        continue // skip
    end
    if $mention$ in $real$ or $hypothetical$ then
        continue // skip
    end
    $person\_set$ = get references for $mention$ from
      $coref$
    if $mention$ in $known\_people$ then
        add $person\_set$ to $real$
        add $gender$ to $real\_genders$
    else
        add $person\_set$ to $hypothetical$
        add $gender$ to $hypothetical\_genders$
    end
end
return $real$, $real\_genders$, $hypothetical$,
  $hypothetical\_genders$
```

4.1.1 Baseline

To identify person mentions in the text, Nadeem (2020) used spaCy (Honnibal and Montani, 2017) to perform: 1) part-of-speech (POS) tagging, 2) named entity recognition (NER), and 3) dependency parsing. Person mentions were identified by using POS tagging and NER to look for pronouns, proper nouns, nouns, and names that indicate people. Person mentions appearing as the nominal subject and direct object (using dependency parsing) were retained and assigned gender labels. Gender Guesser (Michael, 2007) was used together with lists of pronouns as well as nouns to assign gender labels. Compared to this approach in Nadeem (2020), we retained all detected mentions instead of only those appearing as the nominal subject and direct object. We also experimented with both using the original hand-crafted lists of pronouns and nouns in Nadeem (2020) and an expanded version of that (see Appendix A for the lists).

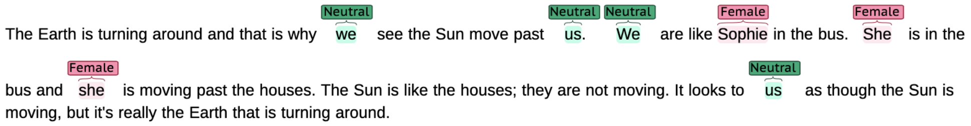

The Earth is turning around and that is why we see the Sun move past us. We are like Sophie in the bus. She is in the bus and she is moving past the houses. The Sun is like the houses; they are not moving. It looks to us as though the Sun is moving, but it's really the Earth that is turning around.

Figure 2: Example of an annotated sample for evaluation on identifying female/male/neutral person mentions.

Approach Metric	**Baseline**	**Baseline + expanded list**	**Our pipeline**
Precision	78.71%	80.18%	**84.16%**
Recall	83.56%	91.44%	**92.81%**
F-score	81.06%	85.44%	**88.27%**

Table 2: Our pipeline outperforms previous work in detecting person mentions in general.

Threshold Metric	**0.7**	**0.8**	**0.9**
Precision	83.71%	87.60%	**89.39%**
Recall	68.42%	**67.80%**	**67.80%**
F-score	75.38%	76.44%	**77.11%**

Table 3: The NEL component works the best on our data when the threshold for similarity score is 0.9.

4.1.2 Annotated data

The data we used (section 2) does not have labels for person mentions and their gender. For our evaluation in this section, we hand labeled 100 samples, annotating for person mentions and whether it is a female, male or neutral mention. We show an example of such annotated sample in Figure 2. The 100 samples labeled contain 607 sentences, a total of 292 person mentions, out of which there are 64 male mentions and 17 female mentions.

4.1.3 Evaluation results

We evaluate the systems based on 1) overall performance for person mention detection, 2) detecting male mentions and 3) female mentions. To assess overall performance, person mentions identified by the systems were compared against human annotations, ignoring the gender labels. In Table 2, we report precision, recall and F-score for 1) the approach in Nadeem (2020), 2) the approach in Nadeem (2020) but with expanded hand-crafted lists of pronouns and nouns, as well as 3) our pipeline (using expanded lists). Of all the person mentions detected, we then focus our evaluation on the gendered mentions, revealing similar results [5]. Across all these evaluations, the expanded lists improves performance on detecting person mentions. Further, mention detection output from our pipeline consistently outperforms the baseline system (even when the baseline is augmented with expanded lists) in precision, recall and F-score.

4.2 Evaluation for NEL and coreference

We use evaluations in this subsection to help us make design choices for NEL and coreference resolution components of our pipeline.

4.2.1 Annotated data

We hand labeled 300 samples containing gendered mentions[6] for 1) whether each named person mention corresponds to a real-world person and 2) coreference relations between mentions. An example of annotated sample is in Figure 3. The 300 samples contain 3202 sentences, 1359 gendered person mentions, 323 person mentions corresponding to real-world people and 575 gendered entities (based on coreference relations). We report the performance of our pipeline under different design choices, using the annotated data as reference labels. Due to limited amount of hand labeled data, we did not have a separate test set to evaluate the final threshold chosen for the NEL component or the coreference model chosen for our final pipeline.

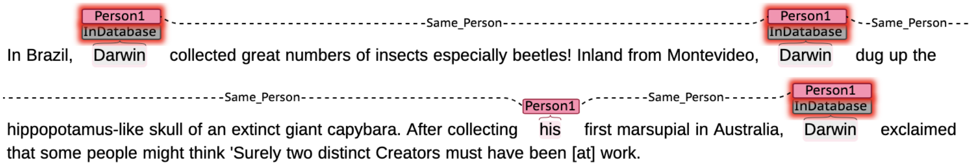
4.2.2 Evaluation results

NEL component: We evaluate for how well person mentions in the given text are linked to real-world entities in database. We use this evaluation to choose a similarity score threshold for the NEL component in our final pipeline. From Table 3, setting the threshold to 0.9 gives the best performance in terms of precision, recall and F-score.

Coreference resolution component: We assess how well person mentions associated with the same person are grouped into the same cluster. For evaluating entity clusters, we use the B-cubed, mention-based approach (Bagga and Baldwin, 1998). We report the precision, recall and F-score, and also compute time needed.Based on the comparison in Table 4, the NeuralCoref model was chosen for our final pipeline as it works better on our data considering precision, F-score, and compute time.

[5] Due to space constraints, we present the tables for these in Appendix B

[6] Since we focus on gendered mentions for our analysis, this set of evaluations is centered on gendered mentions.



In Brazil, Darwin collected great numbers of insects especially beetles! Inland from Montevideo, Darwin dug up the hippopotamus-like skull of an extinct giant capybara. After collecting his first marsupial in Australia, Darwin exclaimed that some people might think 'Surely two distinct Creators must have been [at] work.

Figure 3: Example of an annotated sample for evaluation on NEL and coreference resolution.

Coreference model Metric	**NeuralCoref**	**AllenNLP**
Precision	**74.46%**	64.95%
Recall	62.31%	**65.24%**
F-score	**67.85%**	65.09%
Compute time	**1.6s/item (faster)**	9.0s/item (slower)

Table 4: NeuralCoref model works better for our data.

4.3 Evaluation on how parts of the pipeline complement each other

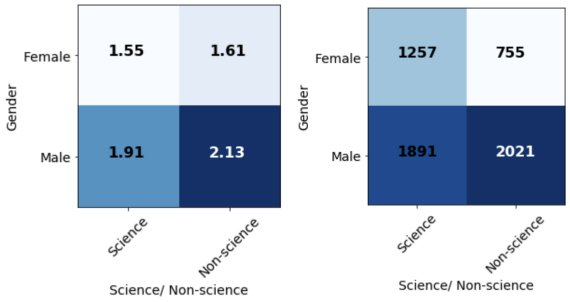
By examining the output for the 300 samples evaluated in subsection 4.2.1, about 46% had a type of sub-task correction. We identified four different ways the pipeline makes such corrections: 1) 21% of the time when a non-person mention is detected (as a person mention), Gender Guesser labels the gender as "neutral" and removes it from the pipeline; 2) 7% of the time when a gender-neutral person is later referred to with a gender specific pronoun, the person is kept and that gender is assigned; 3) 18% of the time the entity linker identifies a non-person mention to be a person in the database, the pipeline later corrects this (e.g. through gender detection); 4) 5% of the time, when the mention detection components did not detect one of the mentions in a cluster as a person, the pipeline fixes this by including it.

5 Analysis of results

Using the pipeline described in section 3 and the design choices made according to evaluations in subsection 4.2, we analyze gender representations in the textbook data for potential bias [7].

5.1 Gendered person mentions

Our pipeline detected a total of 11,094 gendered person mentions[8] in the 33,575 samples of textbook data. The number of male mentions (n = 7932) is more than twice that of female mentions (n = 3162). Across both science and non-science textbooks, two-sided binomial tests show that there are significantly more ($p < 0.001$) male mentions compared to female mentions.



(a) Mentions-to-entity ratio (b) Number of mentions

Figure 4: Analysis of gendered entities.

5.2 Gendered entities

Having observed more male mentions than female mentions, we investigate whether this is due to 1) each male character being mentioned more times compared to if it is a female or 2) there is a greater number of male characters being discussed. We perform analysis on gendered entities[9], analyzing both the mentions-to-entity ratio and the distribution of gendered entities.

5.2.1 More mentions per entity for males

Figure 4a shows the mentions-to-entity ratios for both genders across science and non-science textbooks. T-tests for independent samples, show that mentions-to-entity ratios in both science and non-science textbooks is significantly larger ($p < 0.001$) for males compared to females. Therefore, the observation in subsection 5.1 that there are more male mentions than female mentions can be partially attributed to longer discussions about male characters where each male entity is mentioned more times compared to if it were a female entity.

5.2.2 More male entities

In Figure 4b, across both science and non-science textbooks, two-sided binomial tests show that there are significantly more ($p < 0.001$) male entities compared to female entities. This suggests that

[7] Refer to Appendix C for details of statistical tests.

[8] See Appendix D.1 for distribution of gendered mentions.

[9] See Appendix D.2 for distribution of gendered entities.

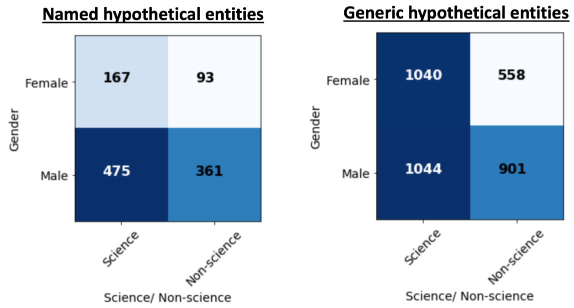


Figure 5: Generally more hypothetical male entities.

male characters are more often discussed in textbooks compared to female ones.

5.3 Focusing on gendered hypothetical entities

Sometimes the content of a textbook may dictate the real-world character mentioned. For instance, when a physics textbook introduces Newton's laws of motion, it probably has to mention Isaac Newton, a male scientist. Hypothetical entities[10] present an interesting case for analysis because this is where authors enjoy much more freedom over 1) whether to use a male or female hypothetical entity and 2) whether to give the character a name or present him/her as a generic character.

5.3.1 More male hypothetical entities

We first compare whether a male or female hypothetical entity is more frequently used. We analyze this for named and generic hypothetical entities, and both in science and non-science textbooks (Figure 5). In three out of the four cases (columns), we observe a gender bias where even when authors had a choice of the gender of the hypothetical mention they use, they are significantly more likely ($p < 0.001$) to use a male hypothetical character rather than a female one. However, the female-male entities gap closes for generic entities in science textbook, indicating that authors are almost equally likely to use a female generic person like "the girl" (n = 1040) compared to a male generic person like "the boy" (n = 1044) in science textbooks. This presents an interesting case for future work on why and how this is so.

5.3.2 Females more often as generic entities

We next compare whether a named or generic hypothetical entity is more frequently used (using the output of our NER component). We analyze this for female and male entities, and for both science and non-science textbooks. From Figure 5, generic hypothetical entities are significantly more likely ($p < 0.001$) to be used across all these cases compared to real hypothetical entities. The result is even more interesting when we analyze the generic-to-named hypothetical entities ratio, which can be obtained by dividing the number in each cell in the contingency table on the right in Figure 5 with the corresponding number in the cell from the contingency table on the left.

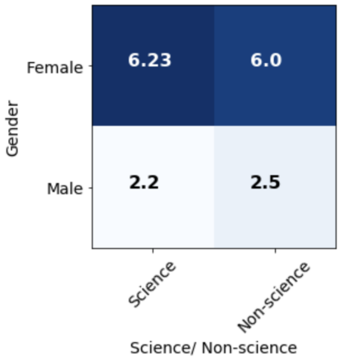


Figure 6: Generic-to-named ratios for hypothetical female entities are much larger.

We summarize the generic-to-named entities ratios obtained in Figure 6, and observe that this ratio is much higher for females. And in fact, the generic-to-named entities ratios for females (6.23 and 6.0 for science and non-science textbook respectively) is more than twice that for males (2.2 and 2.5, correspondingly) across both science and non-science textbooks. This indicates that hypothetical female characters are much more frequently portrayed as a generic entity without an individualized name compared to males. More concretely, females are more likely presented with a passerby-like identity like "that girl" rather than with a name like "Sophie" in comparison to males who are more likely given identities through names.

6 Summary

Drawing various NLP tools, we introduced a pipeline to allow for nuanced, automatic annotation of texts targeted towards the study of gender representations in educational texts. Our analysis suggests that across both science and non-science textbooks, compared to their female counterparts, 1) each male character is mentioned a greater number of times 2) the number of male characters discussed is greater. Even when authors had choices

[10] Appendix D.3 compares real-world versus hypothetical entities, and further analysis on hypothetical entities in this subsection may be found in Appendix D.4

over the hypothetical characters they use and how they are depicted, authors are 1) more likely to use a male hypothetical character and 2) give male characters a name while presenting females as a generic “passer-by” with no individualized identity.

We acknowledge that reducing gender to a male-female binary is simplistic, and hope to explore a more complex understanding of gender in future works. Analyzing how gender representations in textbooks change over the years (like in Garg et al. (2018)) and across different grade levels are also interesting directions for future work.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *LREC*.

Annie Chiponda and Johan Wassermann. 2011. Women in history textbooks: what message does this send to the youth? *Yesterday and Today*, pages 13 – 25.

CK12. 2007. CK-12 Free Online Textbooks, Flashcards, Adaptive Practice, Real World Examples, Simulations.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Colleen Flaherty. 2018. Gender bias, by the numbers.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265.

Jörg Michael. 2007. 40000 namen, anredebestimmung anhand des vornamens. *c't*, (17):182–183.

Michigan. 2014. Michigan Open Book Project.

Farah Nadeem. 2020. Automatic analysis of language use in k-16 stem education and impact on student performance.

Farah Nadeem and Mari Ostendorf. 2018. Estimating Linguistic Complexity for Science Texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55.

Alex Olieman. 2017. pyspotlight 0.7.2.

Robert J. Rydell, Michael T. Rydell, and Kathryn L. Boucher. 2010. The effect of negative performance stereotypes on learning. *Journal of personality and social psychology*, 99 6:883–96.

Siyavula. 2014. Open Textbooks | Siyavula.

Scientific United Nations Educational and Cultural Organization. 2020. *Global Education Monitoring Report 2020*. United Nations.

Ming Te Wang and Jessica L. Degol. 2017. Gender gap in science, technology, engineering, and mathematics (stem): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, 29(1):119–140. Publisher Copyright: © 2016, Springer Science+Business Media New York.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix A

- Original hand-crafted lists in Nadeem (2020):

```
fem_p = ['she', 'her', 'hers', '
    herself']
male_p = ['he', 'him', 'his', '
    himself']
personal_p = ['i', 'me', 'we', '
    us', 'myself', 'ourself', '
    ourselves']
other_p = ['they', 'them', '
    their', 'you', 'themself', '
    themselves']
people = ['adult','adults', '
    person','people','child','
    children']
people_f = ['girl', 'girls','
    woman','women', 'mrs', 'ms',
     'mother', 'mothers']
people_m = ['boy','boys','man','
    men','mr','father','fathers
    ']
```

- Expanded hand-crated lists in our work:

```
fem_p = ['male', 'she', 'her', '
    hers', 'herself']
male_p = ['female', 'he', 'him',
     'his', 'himself']
personal_p = ['i', 'me', 'we', '
    our', 'us', 'myself', '
    ourself', 'ourselves']
other_p = ['they', 'them', '
    their', 'you', 'your', '
    themself', 'themselves']
people = ['people', 'adult','
    adults', 'person','people','
    child','children']

person_f_singular = ['girl','
    woman','mrs','ms','mother','
    mom','aunt','niece','sister
    ','wife','daughter','
    grandmother','grandma','
    grandmom','granddaughter','
    bride','girlfriend','gal','
    madam','lady']
person_m_singular = ['boy','man
    ','mr','father','dad','uncle
    ','nephew','brother','
    husband','son','grandfather
    ','grandpa','granddad','
    grandson','groom','boyfriend
    ','guy','gentleman','
    bachelor']
people_f_plural = ['females', '
    girls','women','mothers','
    moms','aunts','nieces','
    sisters','wives','daughters
    ','grandmothers','grandmas
    ','granddaughters','brides
    ','girlfriends','gals','
    ladies']
people_m_plural = ['males', '
    boys','men','fathers','dads
    ','uncles','nephews','
    brothers','husbands','sons
    ','grandfathers','grandpas
    ','grandsons','grooms','
    boyfriends','guys','
    gentlemen','bachelors']
people_f = person_f_singular +
    people_f_plural
people_m = person_m_singular +
    people_m_plural
```

B Appendix B

Approach Metric	**Previous work** (baseline)	**Previous work + expanded list**	**Our pipeline**
Precision	73.33%	78.95%	**80.00%**
Recall	64.71%	88.24%	**94.12%**
F-score	68.75%	83.33%	**86.49%**

Table 5: Our pipeline outperforms previous work in detecting female mentions.

Approach Metric	**Previous work** (baseline)	**Previous work + expanded list**	**Our pipeline**
Precision	84.00%	84.62%	**85.71%**
Recall	65.62%	68.75%	**75.00%**
F-score	73.68%	75.86%	**80.00%**

Table 6: Our pipeline outperforms previous work in detecting male mentions.

C Appendix C

We used Fisher's exact test as a test of independence between two binary variables such as gender and type of textbook (science or non-science). Holding one variable constant, for instance, by focusing on the female mentions, for binary variables like type of textbook, we test against the null hypothesis is that two categories (e.g. science or non-science textbook) are equally likely to occur using two-sided binomial test. For comparing a continuous variable like mentions-to-entity ratio across two populations, we use T-test for independent samples. An alpha level of 0.05 was used for our statistical tests.

D Appendix D

D.1 Gendered mentions

Our pipeline detected a total of $11,094$ gendered person mentions in the 33,575 samples of textbook data. We analyzing the distribution of these person mentions and summarize the distribution in Figure 7. In these textbooks, the number of person mentions are relatively even across science (n = 5564) and non-science (n = 5530) books. Of all the these person mentions, 3174 are person names that can correspond to real-world people found in DBpedia. These distinctions will be used to enrich our later analysis.

We observe that the number of male mentions (n = 7932) is more than twice that of female mentions (n = 3162), suggesting that females are much underrepresented in textbooks. We then further examine whether how then number of female and males mentions may be affected by the type of textbook (science or non-science). Across both science and non-science textbooks, two-sided binomial tests show that there are significantly more ($p < 0.001$) male mentions compared to female mentions. This analysis on person mentions suggests that females are underrepresented in both science and non-science textbooks.

The relationship between gender of detected mention and type of textbook is summarized in Figure 8. We performed Fisher's exact test, which shows that there is a significant association ($p < 0.001$) between gender and type of textbook. Although across both science and non-science textbooks, two-sided binomial tests show that there are significantly more ($p < 0.001$) male mentions compared to female mentions, we observe interesting trends as to within each gender which type of textbook the mentions tend to occur in. Within, female mentions, a two-sided binomial test shows that female mentions are significantly more likely in a science textbook compared to non-science textbook ($p < 0.001$). In contrast, males are more likely in non-science textbooks than non-science textbooks ($p < 0.001$).

D.2 Gendered entities

Distribution of entities: Our pipeline detected a total of 5924 gendered person mentions in the textbook data. We summarize the distribution of these gendered in Figure 9. In these textbooks, there is slightly more gendered mentions in science (n = 3148) textbooks compared to non-science (n = 2776) textbooks. Of all the these entities, 1285 correspond to real-world people found in the DBpedia database. We observe that the number of male entities (n = 3912) is close to twice that of female entities (n = 2012), suggesting that females are much underrepresented in textbooks.

The relationship between gender of entities and type of textbook is summarized in Figure 4b. We performed Fisher's exact test, which shows that there is a significant association ($p < 0.001$) between gender of the entities and type of textbook it appears in. As reported in the main text, two-sided binomial tests show that there are significantly more ($p < 0.001$) male entities across both types of textbooks. We observe interesting trends as to within each gender which type of textbook the mentions tend to occur in. Within, female mentions, a two-sided binomial test shows that female mentions are significantly more likely in a science textbook compared to non-science textbook ($p < 0.001$). In contrast, males are more likely in non-science textbooks than non-science textbooks ($p < 0.05$).

Mentions per entity: Interestingly, the mentions-to-entity ratio in Figure 4a for males is significantly larger ($p < 0.05$) for in non-science textbooks compared to science textbooks. This indicates that each unique male entity is mentioned more times when it is in a non-science textbook. However, mentions-to-entity ratio is not significantly different across the different types of textbooks for females ($p = 0.411$).

D.3 Real-world vs. hypothetical entities

For the purposes of this work, through differentiating real-world and hypothetical entities, our main goal is to focus part of our analysis on hypothetical entities and how authors of textbooks make their choices on these. However, there are also some interesting trends if one compares real-world and hypothetical entities in terms the mentions-to-entity ratio (Figure 10) and distribution of the entities (Figure 11). We summarize some of these below.

Mentions-to-entity ratio - Science vs. Non-science: Mentions-to-entity ratio for real-world female characters is larger for non-science textbook than science textbook. Whereas, each hypothetical female character is mentioned more when it appears in science textbooks. These differences are statistically significant ($p < 0.001$). No significant difference is found for male characters (both real-world and hypothetical) in their mentions-to-entity ratios across science and non-science textbooks.

Mentions-to-entity ratio - Male vs. Female: Generally, mentions-to-entity ratio is greater for males than females. However, interestingly mentions-to-entity ratio is greater for females (3.32) than males (2.44) for real-world entities in non-science textbook.

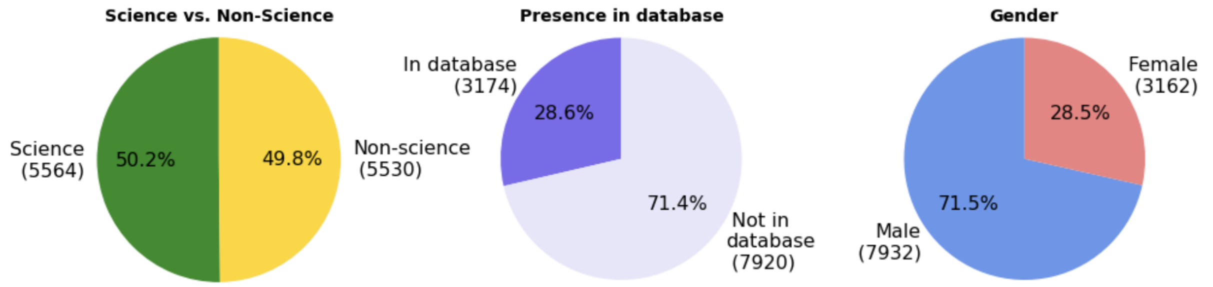


Figure 7: Distribution summary of detected gendered person mentions in textbook data. Female mentions are much underrepresented in textbooks.

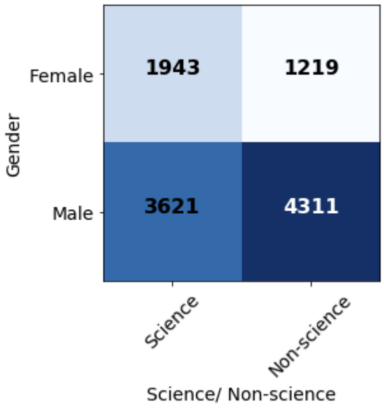


Figure 8: Analysis of gendered person mentions in science and non-science textbooks.

Mentions-to-entity ratio - Real-world vs. Hypothetical: Mentions-to-entity ratio is consistently larger for real-world entities compared to hypothetical entities.

Distribution of entities - Science vs. Non-science: For real-world entities, across both genders, the number of entities in non-science textbooks is significantly greater (more than twice) than that of science textbooks. Whereas for hypothetical entities, across both genders, the number of entities is greater in science textbooks. In all cases, $p < 0.001$.

Distribution of entities - Male vs. Female: It is consistent across all cases that there are more male entities than female entities (in all cases, $p < 0.001$). Interestingly, the gap closes slightly for hypothetical entities in science textbooks.

Distribution of entities - Real-world vs. Hypothetical: There are consistently more hypothetical entities compared to real-world entities in all cases.

D.4 A closer look within hypothetical entities

We present further analysis on mentions-to-entity ratios within hypothetical entities in Figure 12

Mentions-to-entity ratio - Science vs. Non-science: For named hypothetical entities, mentions-to-entity ratio for male entities is significantly larger for non-science compared to science ($p < 0.05$) textbooks. For generic hypothetical entities, mentions-to-entity ratio for female characters is larger for science textbook than non-science textbook ($p < 0.05$).

Mentions-to-entity ratio - Male vs. Female: Generally, mentions-to-entity ratio is greater for males than females. However, interestingly mentions-to-entity ratio is the same for for females and males (1.91) for named hypothetical entities in science textbooks.

Mentions-to-entity ratio - Named vs. Generic: Mentions-to-entity ratio is consistently larger for named hypothetical entities compared to generic ones.

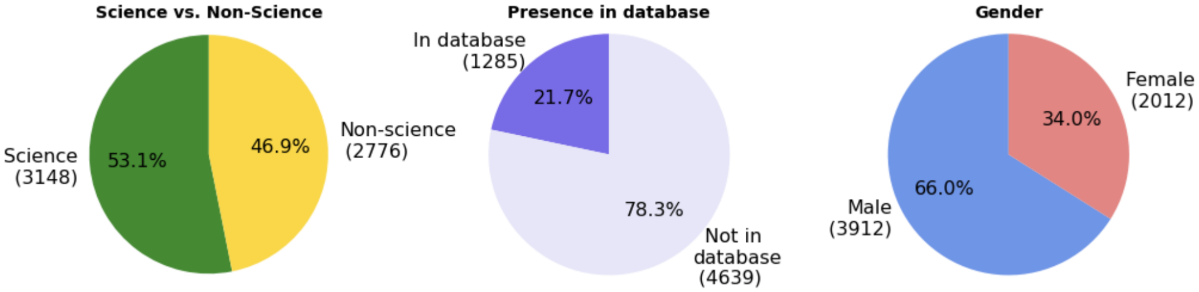


Figure 9: Distribution summary of detected gendered entities in textbook data. Female entities are much underrepresented in textbooks.

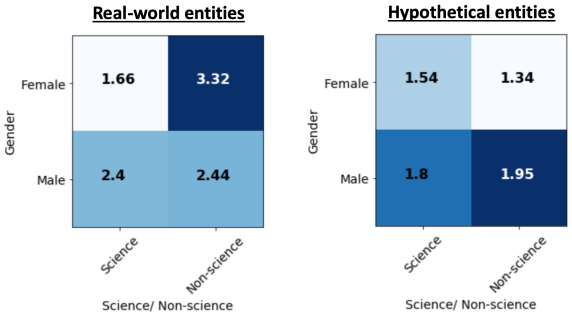


Figure 10: Comparing real-world vs. hypothetical entities: number of mentions per entity for female and male entities across science and non-science textbooks

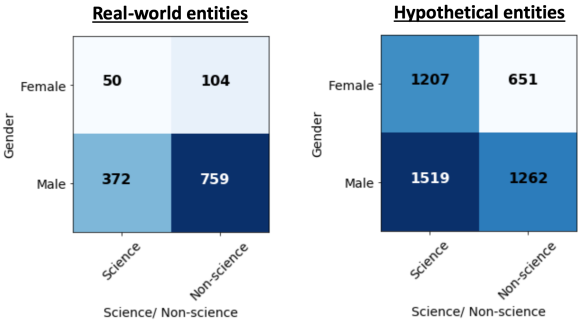


Figure 11: Comparing real-world vs. hypothetical entities: analysis of gendered entities across science and non-science textbooks.

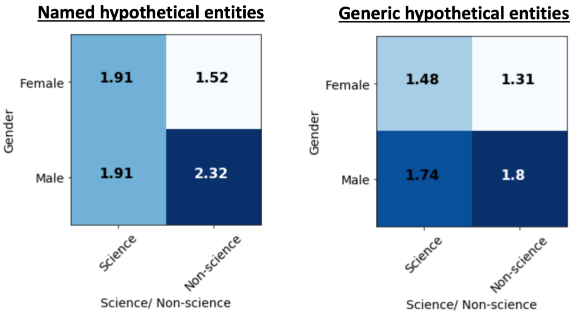


Figure 12: Comparing named vs. generic hypothetical entities: number of mentions per entity for female and male entities across science and non-science textbooks