Quantifying Gender Bias in a Educational Natural Language Corpus

Vivek Raman
Student, MS in Computer Science
San Jose State University
vivek.raman@sjsu.edu

***Abstract*—This paper presents a quantitative analysis of gender bias in the English Wikibooks corpus, a large-scale, crowdsourced repository of educational literature. Employing various linguistics models like lexical frequency analysis and co-occurrence analysis, the corpus is evaluated for bias prevailing between gendered terms and semantically relevant word classes. All metrics are computed using GenBit, a purpose-built toolkit for gender bias evaluation, on an October 2021 snapshot of the English-language subset of Wikibooks.org.**

I. Introduction

Gender bias is the inherent bias based on gender that humanity has come to identify in our history. It encompasses the implicit association of certain neutral words like societal, civil, or domestic roles to certain genders. These biases, internalized in society over the years, has inadvertently made its way into all types of literature. Understanding how gender is represented in educational and encyclopedic text is an important step toward identifying biases and gaps in knowledge.

As software systems are increasingly powered by artificial intelligence (AI) and natural language processing (NLP) models that are capable of picking up and internalizing these biases [1], it is important to recognize and mitigate bias in the data we feed these models.

A reputable publicly available source of crowdsourced educational and encyclopedic natural language corpus is Wikibooks.org [2]. This research seeks to apply linguistic models on a Wikibooks corpus in an attempt to quantify the gender bias prevalent in it.

II. Related Work

A recent study scanned over 1200 primary school textbooks to analyze the portrayal of different genders [3]. The authors found that there is measurable under-representation of female-leaning genders, and they are primarily associated with passive and domestic roles, whereas male-leaning genders are associated with achievement and progress.

A similar analysis conducted on children's books [4] elaborates the use of word embeddings to discover the qualities associated with different genders. The authors deduce a similar observation; that females are associated with appearance rather than competence, and males follow the reverse. The authors highlight the adverse effects that the prevalence of gender bias in media catered to children might have on their worldview.

Gender bias is also observed in Wikipedia, the most popular crowdsourced encyclopedic corpus [5]. Authors perform a multi-dimensional analysis of the corpus, observing similar patterns of gender bias to the previously discussed studies. Additionally, they observed that male biographies carried more "abstract positive language" whereas female biographies contained "negative or neutral terms".

III. Methodology

To study the Wikibooks corpus, various linguistic models and techniques may be applied. The models discussed here have shown accurate and detailed evaluations of various corpora.

A. Dataset

The Wikibooks corpus used here is a snapshot of the Wikibooks collection collected in October 2021 across 12 languages (100 million tokens) [6]. The contents are available in new-line-delimited plaintext as well as in HTML form. The subset of English books will be used for this study.

B. Lexical frequency analysis

This involves scanning the database for the occurrence of *gendered terms* and tallying their frequencies. Gendered terms include pronouns (he / she), nouns with distinct gendered forms (host / hostess), and kinship nouns (uncle / aunt). Lexical frequencies portray an idea of the distribution of gendered terms and identifies if there is disproportionate narrative involving one gender.

C. Co-occurrence analysis

Co-occurrence statistics measure the strength of word associations. In each sentence, the gendered terms are extracted along with the subject of the sentence, be it related to occupation or adjectives. In an unbiased setting, there is no significant co-occurence of gendered terms with different occupation types or adjectives.

IV. Results and conclusion

To collect these metrics, a promising tool purpose-built for gender bias analysis called GenBit [7] is used. Gathering various metrics by applying GenBit on the Wikibooks dataset will reveal insights on gender bias prevalent within the corpus.

REFERENCES

[1] R. Schwartz, R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, *Towards a standard for identifying and managing bias in artificial intelligence*, vol. 3. US Department of Commerce, National Institute of Standards and Technology . . . , 2022.

[2] Wikibooks.org, "Wikibooks:what is wikibooks," 1 2025.

[3] L. Crawfurd, C. Saintis-Miller, and R. Todd, "Sexist textbooks: Automated analysis of gender bias in 1,255 books from 34 countries," *PLOS ONE*, vol. 19, pp. 1–27, 10 2024.

[4] A. Adukia, P. Chiril, C. Christ, A. Das, A. Eble, E. Harrison, and H. B. Runesha, "Tales and tropes: Gender roles from word embeddings in a century of children's books," in *Proceedings of the 29th International Conference on Computational Linguistics* (N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, eds.), (Gyeongju, Republic of Korea), pp. 3086–3097, International Committee on Computational Linguistics, Oct. 2022.

[5] C. Wagner, E. Graells-Garrido, D. Garcia, and F. Menczer, "Women through the glass ceiling: gender asymmetries in wikipedia," *EPJ Data Science*, vol. 5, p. 5, Mar. 2016.

[6] D. Dave, "Wikibooks dataset - kaggle," 2021.

[7] K. Sengupta, R. Maher, D. Groves, and C. Olieman, "Genbit: measure and mitigate gender bias in language datasets," *Microsoft Journal of Applied Research*, vol. 16, pp. 63–71, 2021.