Investigating Gender Bias in Teaching Evaluation via Large Language Models

Introduction: Student evaluation of teaching plays a pivotal role in shaping higher education. Research has consistently demonstrated biases against female instructors in these assessments (Centra & Gaubatz, 2000; Mitchell & Martin, 2018; Khokhlova et al., 2023). Compared to their male counterparts, female instructors frequently face a challenging ”double-bind,” where gender expectations (e.g., warmth and approachability) conflict with the professional expectations of higher education (e.g., authority and deep knowledge), leading to evaluation scores that do not truly reflect their competence (MacNell et al., 2015). Recognizing and understanding such biases in teaching evaluations can help improve the situation for female instructors, promoting fairness within the educational and social system.

Traditionally, social scientists have relied on manual annotation and frequency-based techniques to analyze gender bias in textual datasets, but these approaches have limitations, especially when dealing with vast amounts of data. With the advent of large language models (LLMs) like GPT-4, there is a promising avenue for analyzing big data. However, limited research exists on harnessing these models to explore gender bias in educational contexts. This study seeks to bridge this gap by utilizing LLMs to automatically detect and analyze gender bias within expansive teaching evaluation data.

Research Plan: The present study is guided by three research objectives: (A) to develop an aspect-based gender bias detection approach utilizing LLMs, (B) to conduct a comprehensive analysis of gender bias over a pre-collected, large-scale dataset on student teaching evaluation, shedding light on the findings and their broader implications, and (C) to create a user-friendly tool, designed to assist social science researchers in efficiently examining potential biases within their respective datasets.

To accomplish our objectives, we divide our research into four phases. Experiments in all phases will be conducted based on our pre-collected dataset with over eleven million teaching evaluation entries, along with instructors’ demographic information. The dataset was gathered from RateMyProfessors.com, a widely recognized online platform where students can provide feedback and ratings for professors. Ratings are based on a five-point Likert scale, with students having the option to add comments, which gives insights into the instructor's performance from multiple aspects perceived by students.

In the initial phase, we will ask LLMs to directly process individual teaching evaluations, identifying the key aspects discussed, summarizing up to three keywords for each aspect, and assigning scores both for individual aspects and an overall evaluation score. Then we conduct the comparative analysis of keywords and scores between male and female instructors from each aspect, aiming to uncover potential biases in the evaluations. Additionally, this stage will involve experimenting with various models such as GPTs and LLaMAs to determine their effectiveness and comparative efficiencies in analyzing the evaluations.

The second phase is dedicated to the enhancement of the selected open-sourced LLM. The approach is multifaceted, from strategies such as in-context learning, i.e., learning from a few examples in the context (Dong et al., 2022), to instruction-tuning (Peng et al., 2023), a recently-developed effective technique to align LLMs with human intents. Specifically, we will adopt self-instruct tuning (Wang et al., 2022), where larger, more sophisticated models (e.g., GPT-4) are used to improve the performance of smaller open-source models like LLaMA. This method is not only cost-effective but also aims to elevate the smaller models to the performance level of their more advanced counterparts. The fine-tuned LLM will later be available as an API for social science researchers to apply to their own datasets.

The third phase of our methodology focuses on the human evaluation of the LLMs. This assessment involves recruiting 5 annotators who will each be given teaching evaluations alongside the LLM-generated aspects. For each evaluation regarding each aspect, these annotators will then give a score on a scale of 1-5 or indicate if the aspect is not covered in this evaluation. By aligning the model's output with human perception, we aim to validate the reliability and effectiveness of the LLM in identifying and understanding biases in teaching evaluations.

The final phase of our methodology centers on an empirical analysis of our large-scale teaching evaluation dataset. Based on the model proposed above, we aim to examine the underlying hypothesis that students hold instructors to gendered behavioral standards, demonstrating less tolerance for female instructors who deviate from these norms compared to their male counterparts. Through a nuanced aspect-based analysis, we intend to identify both the gender-specific and professional expectations embedded in student evaluations. By comparing the ratings of female instructors to those of male instructors, we aim to determine the degree to which these expectations influence evaluative outcomes.

Intellectual Contributions: In educational literature, the detection of gender bias within textual data has traditionally depended on either human annotation or frequency-based techniques. Human annotation involves researchers meticulously categorizing and interpreting interview transcripts or other textual data according to a predefined set of themes or codes, commonly known as a codebook (Saldaña, 2021). Despite its strength in capturing nuanced and context-rich insights, this approach is considerably time-consuming and can only be applied to relatively small data sets due to human resource constraints (Braun & Clarke, 2006). Conversely, frequency-based methods can process larger datasets but are limited by their reliance on specific term frequencies, often overlooking the deeper semantics and context enveloping the words, and thus failing to discern subtle biases (Rudman & Kilianski, 2000).

With the advent of natural language processing (NLP), researchers have begun exploring word-embedding-based methods to help detect and understand gender biases in textual data (Bolukbasi et al., 2016; Caliskan et al., 2017; Gonen & Goldberg, 2019). These techniques, representing words in high-dimensional spaces, allow for the analysis of semantic relationships between words based on their usage patterns. However, while these methods capture more nuanced biases than frequency-based approaches, they still lack the ability of broader contextual understanding to capture bias at the sentence or document Level (Mikolov et al., 2013). Also, since these models pick up ”the entire spectrum of human biases reflected in language" (Caliskan et al., 2017), their interpretation requires certain expertise, limiting their accessibility to researchers and educators.

The emergence of LLMs, such as OpenAI's GPT-series, has opened new frontiers for analyzing natural language data (Brown et al., 2020). These models have demonstrated a remarkable capacity to comprehend, generate, and provide explanations for large-scale natural language datasets (Wang et al., 2018). Their ability to parse vast datasets and generate coherent, context-aware responses holds marked potential for identifying and explaining gender bias in teaching evaluation texts. Moreover, these LLMs, with their deep learning capabilities, can unveil implicit biases embedded in language that are often invisible to human annotators and beyond the reach of frequency- or word-embedding-based methods (Webster et al., 2018).

In summary, while existing literature establishes a foundation for understanding and detecting gender bias in educational feedback, there is a clear need for methods that can automatically capture and analyze the substantial bias in textual data. This research proposes to leverage the capabilities of LLMs to address this gap and advance the detection and understanding of gender bias in student evaluation of teaching.

Expected Benefits: The significance of this work is two-fold. Firstly, recognizing and understanding gender bias in teaching evaluations can help improve the situation for female educators, promoting fairness within the educational system. This insight further aids in understanding broader gender biases and the underlying socio-cultural psychology, opening discussions on how to mitigate and reduce such biases.

Additionally, this work embodies an integration of disciplines, including computer science, psychology, and education. Utilizing LLMs for gender bias detection represents a methodological innovation that can analyze extensive textual data more efficiently and accurately than traditional human-centric methods. This approach not only promises improvements in the immediate context of teaching evaluations but also has the potential for wider application, providing a useful tool for ongoing assessment and improvement in various social science contexts.

References

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77–101.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.

Centra, J. A. & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The journal of higher education*, 71(1), 17–33.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., & Sui, Z. (2022). A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Gonen, H. & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Khokhlova, O., Lamba, N., & Kishore, S. (2023). Evaluating student evaluations: Evidence of gender bias against women in higher education based on perceived learning and instructor personality. *Front. Educ*, 8, 1158132.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40, 291–303.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mitchell, K. M. & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, 51(3), 648–652.

Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Rudman, L. A. & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and social psychology bulletin*, 26(11), 1315–1328.

Saldaña, J. (2021). *The coding manual for qualitative researchers*. sage.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6, 605–617.