Thesis and Grant Proposal

Abstract

The integrity of medical residency selection is paramount in cultivating competent physicians. With recent changes to the selection process, the weight of the letter of recommendation has increased, bringing potential gender and ethnic biases to the forefront. This study proposes a novel computational approach in applicant assessment research, to quantify, interpret, and visualize the root cause of implicit gender and racial biases. Using advanced natural language processing techniques, including word embeddings through word2Vec and other custom transformer-based models, and a suite of bias-detection algorithms, this study not only seeks to influence policy, fostering a fairer and equitable selection process but also hopes to lay the groundwork for a broader application of bias examination in various qualitative assessment records across multiple languages.

Introduction

The selection of qualified applicants for medical residency programs in any specialty is a very important step in developing successful physicians. Recent changes like the transition of the USMLE Step 1 to a pass/fail scoring system and the restriction of sub-internship opportunities for medical students have introduced new challenges and considerations in evaluating candidates' competencies (Reghunathan et al., 2021). These recent adjustments to the evaluation criteria have streamlined applicant credentials, highlighting the Letter of Recommendation (LoR) as a crucial, and uniquely qualitative, part of the application amidst predominantly quantitative evaluation metrics. This shift underscores a pressing issue: the need to critically examine and address implicit gender and ethnic biases/stereotypes that may pervade LoRs. Uncovering and quantifying biases can lead to fairer evaluation processes, promote diversity, and reduce discrimination, making the selection process more equitable and inclusive.

Current methodologies for evaluating LoRs (Filippou et al., 2019; French et al., 2019) for bias involve utilizing textual analysis tools like Linguistic Inquiry and Word Count (LIWC). We conducted a preliminary analysis using LIWC and a 400-word dictionary on a 5-year dataset comprising 5,679 LoRs of applicants for the UW-Madison plastic surgery residency program. The analysis showed significant bias in multiple categories. For example, see figures 1 and 2, which show a significant difference in gendered terms used for men, and a significant lack of sensitive/nervous adjectives used for African American females. The results warranted a further analysis into the root causes of these biases. However, the LIWC methodology posed a few fundamental limitations. LIWC operates with a predefined set of dictionaries and assigns words to these categories. This can limit its ability to capture the full semantic richness of the LoRs and increase the potential for misidentification of bias. Furthermore, LIWC analyzes text based on the presence of words corresponding to its categories, without considering the context in which a word is used, which can lead to overgeneralization of linguistic

context. Consequently, we can only seek surface-level insights about the existence of bias through LIWC. This

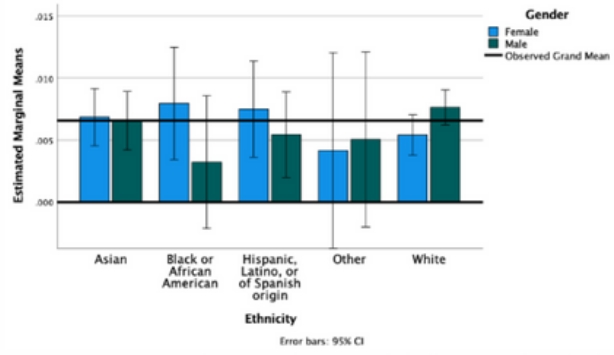

Figure 1: Estimated marginal means of challenging/detached

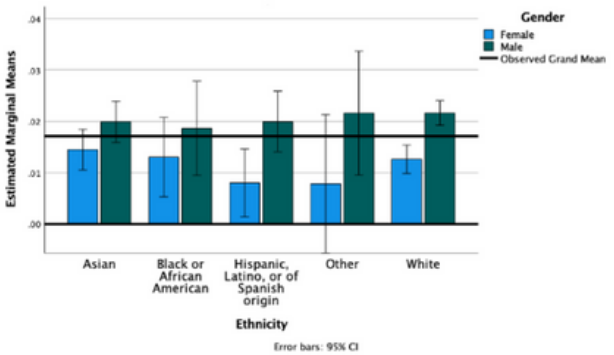

Figure 2: Estimated marginal means of gendered terms

makes it difficult to understand and trust the findings, limiting their applicability in addressing the root causes of biases.

Recognizing the issue of transparency in textual analysis with traditional tools like LIWC, an approach called *word embeddings* has become popular in numerous fields to extract nuanced semantic insights from large language datasets. Word embeddings are high-dimensional vector representations/mappings of words trained through a Neural Network. They encode semantic relationships in geometric/mathematical structures such that similar words are closer together to capture their meaning based on contexts. The unique geometric properties of word embeddings enable one to employ novel bias analysis techniques. They have been used in multiple studies to investigate cultural bias in psychological sciences (Charlesworth et al., 2021; Durrheim et., 2022) and another study used embeddings to illustrate bias in literary books and novels, finding multiple narratives for reinforcing traditional gender stereotypes (Xu et al., 2019).

The current study aims to utilize word embeddings with a novel approach inspired by Kozlowski et al. (2019), where word embeddings were used to analyze cultural and social meanings, related to class, as represented in language. By using the proximity of words within the vector space across multiple dimensions, we aim to detect, quantify, and visualize gender and ethnic biases in LoRs. We expect this approach to not only allow the illustration of these biases clearly but also interpret their significance within the context of medical residency selection, building on insights from traditional linguistic analysis tools. Additionally, we expect that word embeddings' unique properties will allow us to capture the dynamic and nuanced context of each word in a LoR, increasing the accuracy and reliability of bias detection. Our goal with this approach is to create a

transparent process that allows the identification of specific sentences, phrases, and paragraphs contributing to bias in LoRs, offering actionable insights that can help guide policy changes, inform training programs for letter writers, and contribute to developing a more equitable and inclusive selection process in medical residencies.

Methods

The study uses a mixed-methods approach that involves using word embeddings trained through Word2Vec and other models on a comprehensive 11-year corpus of LoRs provided by candidates to the UW-Madison Plastic Surgery Residency Program in PDF form. The dataset will be converted to text using Python OCR (Optical Character Recognition) software, with conversion errors corrected by the SymSpell library, which efficiently corrects misspellings and standardizes text through fuzzy search. Extensive Python scripts will redact all identifying information, leveraging POS (Part-of-Speech) tagging and NER (Named Entity Recognition) software to remove personal identifiers. The data will be further processed by lowercasing, removing stop words, and then normalized through lemmatization to maintain semantic integrity, and tokenization to break down text into individual words for analysis.

Training word2Vec along with Other Models and Extracting Word Embeddings

We will use the processed corpus of text to train a word2Vec Neural Network using the gensim Python library. We will also use the skipgram model to capture semantic relations for relatively small datasets like ours (Mikolov et al., 2013). The Skip-gram architecture will be configured with optimal hyperparameters, including context window size, vector dimensionality, and minimum word count thresholds, to ensure that the word embeddings developed are semantically rich and of high quality. The extracted embeddings from this process will be used for the ensuing bias analysis. Embeddings generated from the hidden-layer representations of a transformer-based model will be similarly extracted for the following bias analysis.

Context-based Bias Analysis and Visualization

We will use Analogy Tests to uncover contextual results from the dataset. These tests uncover stereotypical information by comparing semantically similar word pairs. The relationship "a is to b as c is to what?" can be uncovered with simple vector arithmetic. For example, a biased scenario with the analogy "man: doctor :: woman: ?" would give extremely female-stereotyped occupations like nursing if bias exists.

We will use the Single-Category Word Embedding Association Test (SC-WEAT) to quantify bias by evaluating the strength of association of a single target category (e.g., terms typically associated with a certain

gender or ethnicity) with attribute sets (e.g., positive or negative words, or words associated with professional competence). A significant difference will indicate a bias toward or against the category in relation to the attribute set.

Natural Language Inference (NLI) Tests will allow us to probe the vector space of word embeddings with a premise hypothesis structure. By crafting hypotheses that target gender and race-associated fallacies, such as assumptions about leadership abilities or emotional intelligence, the extent to which such stereotypes affect LoRs can be systematically studied.

Concordance Analysis will help us investigate the co-occurrence and context of specific words or phrases within texts to uncover patterns of language use. By examining how frequently certain attributes (e.g., adjectives denoting competence or intelligence) appear near gender or ethnicity identifiers in letters of recommendation, this test will reveal biases in the description of different groups. A higher incidence of positive terms linked with certain genders or ethnicities, as opposed to others, can indicate an underlying bias. Concordance analysis not only identifies these disparities but also quantifies them by analyzing the distribution and context of word usage, offering a nuanced understanding of implicit biases in evaluative texts.

We will use t-distributed Stochastic Neighbor Embedding (t-SNE), a machine learning algorithm to visualize underlying patterns, clusters, and relationships between words, and create an interactive platform using tensorboard to explore word relationships.

Finally, we will use *Cosine Similarity Heatmaps* for mapping word similarities to color intensity in a matrix. the mapping between word vectors related to gender and ethnicity against those associated with professional attributes/competencies will reveal patterns of association that might indicate bias.

Conclusion/Future Directions

This study focuses on making the process of selecting medical residents fairer by examining the subtle biases in recommendation letters with advanced technology. This work will provide a starting point for the creation of an open-source codebase for replicating the methodology across different use cases, datasets, and languages. The approach proposed in this study holds the potential to make evaluation processes more equitable with broader applications, allowing for a wide-ranging examination of biases in different contexts. Nevertheless, this study will also have an adaptable framework for evaluating gender and ethnic stereotypes in various qualitative admission application components, extending its impact beyond letters of recommendation.

References

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (arXiv:1607.06520). arXiv. https://doi.org/10.48550/arXiv.1607.06520

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, *32*(2), 218–240. https://doi.org/10.1177/0956797620963619

Durrheim, K., Schuld, M., Mafunda, M., & Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, *62*(1), 617–629. https://doi.org/10.1111/bjso.12560

Filippou, P., Mahajan, S., Deal, A., Wallen, E. M., Tan, H.-J., Pruthi, R. S., & Smith, A. B. (2019). The Presence of Gender Bias in Letters of Recommendations Written for Urology Residency Applicants. *Urology*, *134*, 56–61. https://doi.org/10.1016/j.urology.2019.05.065

French, J. C., Zolin, S. J., Lampert, E., Aiello, A., Bencsath, K. P., Ritter, K. A., Strong, A. T., Lipman, J. M., Valente, M. A., & Prabhu, A. S. (2019). Gender and Letters of Recommendation: A Linguistic Comparison of the Impact of Gender on General Surgery Residency Applicants☆. *Journal of Surgical Education*, *76*(4), 899–905. https://doi.org/10.1016/j.jsurg.2018.12.007

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, *84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Reghunathan, M., Mehta, I., & Gosman, A. A. (2021). Improving the Standardized Letter of Recommendation in the Plastic Surgery Resident Selection Process. *Journal of Surgical Education*, 78(3), 801–812. https://doi.org/10.1016/j.jsurg.2020.09.005

Xu, H., Zhang, Z., Wu, L., & Wang, C.-J. (2019). The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. *PLOS ONE*, *14*(11), e0225385. https://doi.org/10.1371/journal.pone.022538