Gendered Language in Resumes and its Implications for Algorithmic Bias in Hiring

Prasanna Parasurama and **João Sedoc**
New York University
pparasurama@stern.nyu.edu

Abstract

Despite growing concerns around gender bias in NLP models used in algorithmic hiring, there is little empirical work studying the extent and nature of gendered language in resumes. Using a corpus of 709k resumes from IT firms, we train a series of models to classify the gender of the applicant, thereby measuring the extent of gendered information encoded in resumes. We also investigate whether it is possible to obfuscate gender from resumes by removing gender identifiers, hobbies, gender subspace in embedding models, etc. We find that there is a significant amount of gendered information in resumes even after obfuscation. A simple Tf-Idf model can learn to classify gender with AUROC=0.75, and more sophisticated transformer-based models achieve AUROC=0.8. We further find that gender predictive values have low correlation with gender direction of embeddings – meaning that, what is predictive of gender is much more than what is "gendered" in the masculine/feminine sense. We discuss the algorithmic bias and fairness implications of these findings in the hiring context.

This paper has been accepted as a non-archival publication.

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 74 - 74