Job Applicant Resumes Are Effectively Impossible to De-Gender, AI Researchers Find

Originally published at https://www.unite.ai/job-applicant-resumes-are-effectively-impossible-to-de-gender-ai-researchers-find/

Researchers from New York University have found that even very simple Natural Language Processing (NLP) models are quite capable of determining the gender of a job applicant from a 'gender-stripped' résumé – even in cases where machine learning methods have been used to remove all gender indicators from the document.

Following a study that involved the processing of 348,000 well-matched male/female résumés, the researchers conclude:

> *'[There] is a significant amount of gendered information in resumes. Even after significant attempts to obfuscate gender from resumes, a simple Tf-Idf model can learn to discriminate between [genders]. This empirically validates the concerns about models learning to discriminate gender and propagate bias in the training data downstream.'*

The finding has significance not because it's realistically possible to hide gender during the screening and interview process (which it clearly is not), but rather because just getting to that stage may involve an AI-based critique of the résumé with no humans-in-the-loop – and HR AI has obtained a besmirched reputation for gender bias in recent years.

Results from the researchers' study demonstrate how resilient gender is to attempts at obfuscation:

	Matched Sample?	Obfuscation	Model	AUROC	With-in Job AUROC
1	No	None	Tf-Idf + Logistic	0.88	0.84
2	Yes	None	Tf-Idf + Logistic	0.85	0.83
3	Yes	Names/IDs removed	Tf-Idf + Logistic	0.78	0.76
4	Yes	Names/IDs, gender IW removed	Tf-Idf + Logistic	0.75	0.73
5	Yes	Names/IDs, gender IW, hobbies removed	Tf-Idf + Logistic	0.75	0.72
6	Yes	Names/IDs, gender IW, hobbies removed	Longformer	0.80	0.79
7	Yes	Names/IDs, gender IW, hobbies removed	Word2Vec + Logistic	0.68	0.65
8	Yes	Names/IDs, gender IW, hobbies removed, debiased Word2Vec	Word2Vec + Logistic	0.67	0.64

Results from the NYU paper. Source: https://arxiv.org/pdf/2112.08910.pdf

The findings above use a 0-1 *Area Under the Receiver Operating Characteristic* (AUROC) metric, where '1' represents a 100% certainty of gender identification. The table covers a range of eight experiments.

Even in the worst-performing results (experiments #7 and #8), where a résumé has been so severely stripped of gender-identifying information as to be non-usable, a simple NLP model such as Word2Vec is still capable of an accurate gender identification approaching 70%.

The researchers comment:

> *'Within the algorithmic hiring context, these results imply that unless the training data is perfectly unbiased, even simple NLP models will learn to discriminate gender from resumes, and propagate bias downstream.'*

The authors imply that there is no legitimate AI-based solution for 'de-gendering' resumes in a practicable hiring pipeline, and that machine learning techniques that actively enforce fair treatment are a better approach to the problem of gender bias in the work marketplace.

In AI terms, this is equivalent to 'positive discrimination', where gender-revealing résumés are accepted as inevitable, but re-ranking is actively applied as an egalitarian measure. Approaches of this nature have been proposed by LinkedIn in 2019, and researchers from German, Italy and Spain in 2018.

The paper is titled *Gendered Language in Resumes and its Implications for Algorithmic Bias in Hiring*, and is written by Prasanna Parasurama, from the Technology, Operations and Statistics department at NYU Stern Business School, and João Sedoc, Assistant Professor of Technology, Operations and Statistics at Stern.

Gender Bias in Hiring

The authors emphasize the scale at which gender bias in hiring procedures is becoming literally systematized, with HR managers using advanced algorithmic and machine learning-driven 'screening' processes that amount to AI-enabled rejection based on gender.

The authors cite the case of a hiring algorithm at Amazon that was revealed in 2018 to have rejected female candidates in a rote manner because it had learned that historically, men were more likely to be hired

> *'The model had learned through historical hiring data that men were more likely to be hired, and therefore rated male resumes higher than female resumes.*
>
> *'Although candidate gender was not explicitly included in the model, it learned to discriminate between male and female resumes based on the gendered information in resumes – for example, men were more likely to use words such as "executed" and "captured".'*

Additionally, research from 2011 found that job ads which implicitly seek men explicitly attract them, and likewise discourage women from applying for the post. Digitization and big data schemas promise to further enshrine these practices into automated systems, if the syndrome is not actively redressed.

Data

The NYU researchers trained a series of models to classify gender using predictive modeling. They additionally sought to establish how well the models' ability to predict gender could survive the removal of greater and greater amounts of potentially gender-revealing information, while attempting to preserve content relevant to the application.

The dataset was drawn from a body of applicant résumés from eight US-based IT companies, with each résumé accompanied by details of name, gender, years of experience, field of expertize or study, and the target job posting for which the résumé was sent.

To extract deeper contextual information from this data in the form of a vector representation, the authors trained a Word2Vec model. This was then parsed into tokens and filtered, finally resolving into one embedded representation for each résumé.

Male and female samples were matched 1-1, and a subset obtained by pairing up the best objectively job-appropriate male and female candidates, with a margin-of-error of 2 years, in terms of experience in their field. Thus the dataset consists of 174,000 male and 174,000 female résumés.

Architecture and Libraries

The three models used for the classification task were Term Frequency-Inverse Document Frequency (TF-IDF) + Logistic, Word Embeddings + Logistic, and Longformer.

The first model offers a bag-of-words baseline that discriminates gender based on lexical differences. The second approach was employed both with an off-the-shelf word embeddings system, and with gender-debiased word embeddings.

Data was split 80/10/10 between training, evaluation and testing,

As seen in the results displayed above, the transformer-based Longformer library, notably more sophisticated than the earlier approaches, was almost able to equal a completely 'unprotected' résumé in terms of its ability to detect gender from documents that had been actively stripped of known gender identifiers.

The experiments conducted included data-ablation studies, where an increasing amount of gender-revealing information was removed from the résumés, and the models tested against these more taciturn documents.

Information removed included hobbies (a criteria derived from Wikipedia's definition of 'hobbies'), LinkedIn IDs, and URLs that might reveal gender. Additionally, terms such as 'fraternity', 'waitress', and 'salesman' were stripped out in these sparser versions.

Additional Results

In addition to the results discussed above, the NYU researchers found that debiased word embeddings did not lower the capability of the models to predict gender. In the paper, the authors hint at the extent to which gender permeates written language, noting that these mechanisms and signifiers are not yet well-understood.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai