Gendered Language in Resumes

Anonymous ACL submission

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, removing gender sub-space 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 little correlation with gender direction of embeddings – meaning that, what is predictive of gender is not necessarily "gendered" in the masculine/feminine sense. We discuss the implications of these findings in the algorithmic hiring context.

1 Introduction

Advances in language models have fundamentally changed the nature of many natural language tasks. In job-resume matching, for example, simple keyword-based matching has been replaced by more sophisticated NLP models, promising higher quality matches (e.g., Maheshwary and Misra (2018); Lin et al. (2016); Luo et al. (2019)). At the same time, the black-box nature of these models has raised concerns about the potential for bias in downstream applications. For example, in 2018, Amazon came under fire for its resume screening tool that was reportedly biased against women (Dastin, 2018). 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".

The source of bias in this example is due to both the data (i.e. biased hiring data), and the model (i.e. ability for the model to discriminate between genders), both of which are necessary conditions for the overall system to be biased. A common thread of concern in model-based bias is that more sophisticated models can more easily learn gendered information from resumes. Within the context of resumes, such concerns, while theoretically plausible, are largely empirically unfounded, or at the least, anecdotal like in the Amazon case. The extent to which models can learn gendered information from resumes depends on the amount of gendered information in resumes, yet empirically little is known about the extent to which resumes are gendered. Moreover, it is unclear whether resumes contain simple lexical or structural gendered information that can easily be obfuscated. If it is the case that there is very little gendered information in resumes, or that this information can be easily obfuscated, much of the concerns surrounding the potential for model-based gender bias in downstream NLP tasks involving resumes would be less warranted.

To address this gap in the literature, we investigate the extent to which resumes are gendered using a predictive modeling approach. Using a corpus of 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. In addition, we investigate whether it is possible to obfuscate gender from resumes by conducting a series of experiments aimed to remove gendered information while preserving the main content. This includes 1. removing gender identifiers such as names, emails, 2. removing gender indicating words such as "male", "female", "salesman", "waitress", etc. 3. removing hobbies, 4. removing gender sub-space in embeddings. To the best of our knowledge, this is the first study within the field of computational linguistics to directly assess this concern using state-of-the-art models.

2 Related Work

Digitization of hiring has spawned an emerging line of research at the intersection of hiring discrimination and algorithmic bias. This research investigates how algorithmic hiring tools (e.g. resume screening) can be biased, and how that can lead to discrimination in hiring (Raghavan et al., 2020; Chen et al., 2018; Sühr et al., 2021; Peng et al., 2019). As more sophisticated NLP models get deployed in such tools, a cause for concern is that the gender representations learned in these models can propagate bias in downstream tasks. For example, research has shown that gendered wordings exist in job postings (Gaucher et al., 2011; Böhm et al., 2020). If a resume screening tool matches resumes to job descriptions using embeddings of the documents, male resumes may be more likely to be matched to job descriptions with masculine language. One approach to address this issue is to debias the NLP models to minimize potential bias in downstream tasks (Sun et al., 2019b). These general-purpose techniques, though not specific to resumes, include obfuscating/swapping gender (Zhao et al., 2018a), removing gender subspace in embedding models (Bolukbasi et al., 2016) (See also Gonen and Goldberg (2019) for the shortcomings of this approach), and training gender-neutral embeddings (Zhao et al., 2018b).

Another cause for concern is that language models can learn gendered information from resumes, propagating any bias in the training data downstream like in the Amazon example. These concerns, while theoretically plausible, are largely empirically unfounded within the resume context. The extent to which models can use and learn gendered information from resumes depends on the amount of gendered information and the nature of gendered information in resumes, including gender differences in language use.

2.1 Gender Differences in Language Use

The earliest study of gender differences in language use investigates lexical differences between genders using verbal samples and finds that women use more words related to feeling, emotion, and motivation, whereas men use more words related to time, space, and quantity (Gleser et al., 1959). Scholars have since studied linguistic gender differences in a wide range of contexts including articles, social media posts, emails, grants (Newman et al., 2008; Argamon et al., 2003; Streib et al., 2019; Urquhart-Cronish and Otto, 2019; Colley and Todd, 2002). According to a meta-review involving 70 separate studies, women use more words related to psychological and social processes, whereas men use more words related to object properties and impersonal topics (Newman et al., 2008).

Although this literature has documented gender differences in writing across different contexts, studies have also shown that these differences become small or non-existent in formal and structured writing (Sterkel, 1988; Smeltzer and Werbel, 1986). Similarly, in a related study involving job applications, researchers study gender differences in self-presentation styles in applications for Teach for America jobs (Streib et al., 2019). They use hand-coded features such as “leader”, “self-growth”, “passionate” in candidates’ job applications, and find minimal gender differences in the coded features. Given that resumes are structured documents, it is unclear the extent to which resumes will be gendered. This is ultimately an empirical question, which we address in this paper.

3 Data and Methods

The primary dataset is a corpus of applicant resumes from 8 IT firms based in the U.S. These IT firms are clients of an HR analytics firm, who provided us the aggregated data as part of a research partnership. Along with the resume text, we have the applicant’s name, gender, years of experience, degree[1], field of study[2], and the job posting to which they applied.

3.1 Vector Representations of Resumes

In addition to applicant attributes mentioned above, we also extract the skills, competencies, job titles, and job-relevant keywords from the resume using a skills and job titles dictionary[3], and create a dense vector representation of each resume based on these keywords. To do so, we first train a Word2Vec model on resumes to learn a vector representation for all tokens (Mikolov et al., 2013). We then parse

[1] Associate, Bachelors, Masters, Doctorate

[2] Technical, Science, Business, Law, Other

[3] This dictionary was created by aggregating all the skills and job titles using a secondary LinkedIn dataset.

the body of the resume into tokens and filter for skills, competencies, job titles, and job-relevant keywords using the dictionary. Finally, we take the average vector representations of the filtered keywords to get one representation for each resume document – the resulting vector is an embedding of the resume in a skills vector space (See Appendix A for an illustration).

3.2 Matched Sample of Resumes

Occupations vary by gender, so occupational characteristics (i.e. skills, past job experiences, education, etc), are an obvious source of gendered information that a classifier can easily learn. However, such information is less relevant in the context of resume screening applications since applicants applying to the same job are likely to have similar education, skills, and experience. So, to ensure that the classifier learns gendered features beyond occupational characteristics, we match our samples so that male and female resumes are on average similar in observable occupational characteristics. Specifically, we perform 1-1 matching without replacement such that for each male resume, we find a female resume that is within 2 years of experience, has the same degree, field of study, and has a resume similarity score (i.e. cosine similarity of resume vector representations) of at least 0.7. If multiple female resumes match these criteria, we take the resume with the highest similarity score. This matching procedure yields 348k resumes (174k male, 174k female resumes).

3.3 Measuring Gendered Information in Resumes

We define gendered information as anything that is predictive of the applicant's gender. To measure the extent of gendered information, we take a predictive modeling approach, where we train a series of models on resumes to classify the gender of the applicant and measure the model's predictive power on a holdout test set.

We use three different sets of models for the classification task: (1) Tf-Idf+Logistic, (2) Word Embeddings+Logistic, (3) Longformer (See Appendix B for model specifications and hyperparameter tuning). We begin with a simple bag-of-words baseline using a Tf-Idf+Logistic model. This model is expected to discriminate between genders based on lexical differences. Next, we use Word Embedding+Logistic model using both off-the-shelf[4] word embeddings and gender-debiased word embeddings (Bolukbasi et al., 2016). This model is expected to discriminate between genders based on differences in document representations. Finally, to learn to discriminate based on more sophisticated features (e.g. contextual representation, structure of the document, etc.), we use LongFormer, a transformer-based model, optimized for long documents (Beltagy et al., 2020). For all of the above models, we use an 80/10/10 train/evaluation/test split.

3.4 Obfuscating Gendered information

In addition to measuring the extent of gendered information, we also investigate whether it's possible to obfuscate gender from resumes while preserving its main content. Keeping in mind the application context (i.e. downstream NLP tasks involving resumes), there is a tradeoff between obfuscating gendered information and obfuscating useful task-relevant information. For example, in resume screening, removing all content from resumes except for a handful of job-specific skills and keywords certainly removes gendered information, but at the cost of also removing useful information in the body of the resume. On the other hand, removing names from resumes obfuscates gender without much effect on task-relevant information. First, we remove names (both by string-matching applicant names, and via named entity recognition), emails, LinkedIn IDs, and other URLs from the resume, and replace the tokens with `[DEL]`. Second, we remove gender indicating words such as "he", "she", "salesman", "waitress", etc. (See Appendix C for the full list). Third, we remove hobbies from the resume using the Wikipedia dictionary of hobbies[5]. Finally, for models based on word embeddings, we compare gender debiased word embeddings to off-the-shelf word embeddings to understand whether gender-debiasing helps to obfuscate gendered information in resumes.

4 Results

Table 1 reports the out-of-sample gender classification performance using Area Under the Receiver Operating Characteristic (AUROC) scores for a series of obfuscation experiments. As noted earlier, these scores are a measure of the amount of

[4] https://code.google.com/archive/p/word2vec/

[5] https://en.wikipedia.org/wiki/List_of_hobbies. Accessed 9/28/2021

	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

Table 1: Out-of-Sample Gender Classification Performance

gendered information in resumes. As a more conservative measure, we also calculate the within-job AUROC score, which measures the performance on the subset of applicants that applied to the same job posting. We calculate this score for each job posting separately and aggregate the scores across jobs by taking a weighted average based on the number of applicants in each job.

Classification performance decreases as we increasingly remove gendered information. Experiment (1) is the baseline with no matching and no obfuscation, which achieves an AUROC of 0.88. Experiment (2) uses the matched sample, which reduces the AUROC to 0.85. Across (3)-(5), removing names, gender indicating words (IW), and hobbies further reduces AUROC to 0.75. Since experiments (1)-(5) use a bag-of-words Tf-Idf model, the discriminatory features are based on lexical differences between genders. In (6), we replace the Tf-Idf model with a transformer-based Longformer model, which can learn to discriminate on features beyond lexical differences including style and structure of writing. Indeed, AUROC increases from 0.75 to 0.8. Finally, to test whether general-purpose embedding debiasing methods help to obfuscate gender, we train two classifiers using Word2Vec embeddings as features. In (7), we use an off-the-shelf Word2Vec model which achieves an AUROC of 0.68. In (8), we use a gender-debiased version of the same embedding model used in (7), however, this does little to obfuscate gender from the classifier (AUROC=0.67). To understand why this is the case, we analyze how predictive gender features are related to the gender direction in word embeddings. Specifically, we regress the average SHAP value of tokens (a measure of predictive value) on the gender direction in the word embedding model (Lundberg and Lee, 2017) (See Appendix C). We find that there is little correlation between a token's gender predictive value and its corresponding gender direction in word embeddings($R^2 = 0.038$), thus using debiased word embeddings did not decrease the gender predictive performance.

5 Conclusion

There are two important takeaways from these results. First, 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 (AUROC=0.75). This empirically validates the concerns about models learning to discriminate gender and propagate bias in the training data downstream. Second, there is little correlation between gender predictive values and gender direction in word embeddings. Therefore, general-purpose gender debiasing methods for NLP models such as removing gender subspace from embeddings are not effective in obfuscating gender. 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. This calls for active consideration of fairness in downstream tasks such as employing "fairness through awareness" techniques (Dwork et al., 2012; Geyik et al., 2019; Zehlike et al., 2017) that explicitly take into account the protected class to achieve individual or group-level fairness.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*.

Stephan Böhm, Olena Linnyk, Jens Kohl, Tim Weber, Ingolf Teetz, Katarzyna Bandurka, and Martin Kersting. 2020. Analysing Gender Bias in IT Job Postings: A Pre-Study Based on Samples from the German Job Market. In *Proceedings of the 2020 on Computers and People Research Conference*, SIGMIS-CPR'20, pages 72–80, New York, NY, USA. Association for Computing Machinery.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? debiasing Word Embeddings. page 9.

Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14, Montreal QC, Canada. ACM Press.

Ann Colley and Zazie Todd. 2002. Gender-Linked Differences in the Style and Content of E-Mails to Friends. *Journal of Language and Social Psychology*, 21(4):380–392.

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA. Association for Computing Machinery.

Danielle Gaucher, Justin Friesen, and Aaron C. Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1):109.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. *arXiv:1905.01989 [cs]*.

Goldine C. Gleser, L. A. Gottschalk, and W. John. 1959. The relationship of sex and intelligence to choice of words: A normative study of verbal behavior. *Journal of Clinical Psychology*, 15:182–191.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv:1903.03862 [cs]*.

Yiou Lin, Hang Lei, Prince Clement Addo, and Xiaoyu Li. 2016. Machine Learned Resume-Job Matching Solution. *arXiv:1607.07657 [cs]*.

Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*.

Yong Luo, Huaizheng Zhang, Yonggang Wen, and Xinwen Zhang. 2019. ResumeGAN: An Optimized Deep Representation Learning Framework for Talent-Job Fit via Adversarial Learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 1101–1110, New York, NY, USA. Association for Computing Machinery.

Saket Maheshwary and Hemant Misra. 2018. Matching Resumes to Jobs via Deep Siamese Network. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 87–88, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3):211–236.

Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? the Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):125–134.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 469–481, New York, NY, USA. Association for Computing Machinery.

Larry R. Smeltzer and James D. Werbel. 1986. Gender Differences in Managerial Communication: Fact or Folk-linguistics? *The Journal of Business Communication (1973)*, 23(2):41–50.

Karen S. Sterkel. 1988. The Relationship Between Gender and Writing Style in Business Communications. *The Journal of Business Communication (1973)*, 25(4):17–38.

Jessi Streib, Jane Rochmes, Felicia Arriaga, Carlos Tavares, and Emi Weed. 2019. Presenting Their Gendered Selves? how Women and Men Describe Who They Are, What They Have Done, and Why

They Want the Job in Their Written Applications. *Sex Roles*, 81(9):610–626.

Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. Does Fair Ranking Improve Minority Outcomes? understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 989–999, New York, NY, USA. Association for Computing Machinery.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019a. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*, Lecture Notes in Computer Science, pages 194–206, Cham. Springer International Publishing.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019b. Mitigating Gender Bias in Natural Language Processing: Literature Review. *arXiv:1906.08976 [cs]*.

Mackenzie Urquhart-Cronish and Sarah P. Otto. 2019. Gender and language use in scientific grant writing. *FACETS*, pages 442–458.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, pages 1569–1578.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *arXiv:1804.06876 [cs]*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. *arXiv:1809.01496 [cs, stat]*.

A Illustration of Resume Vector Representation

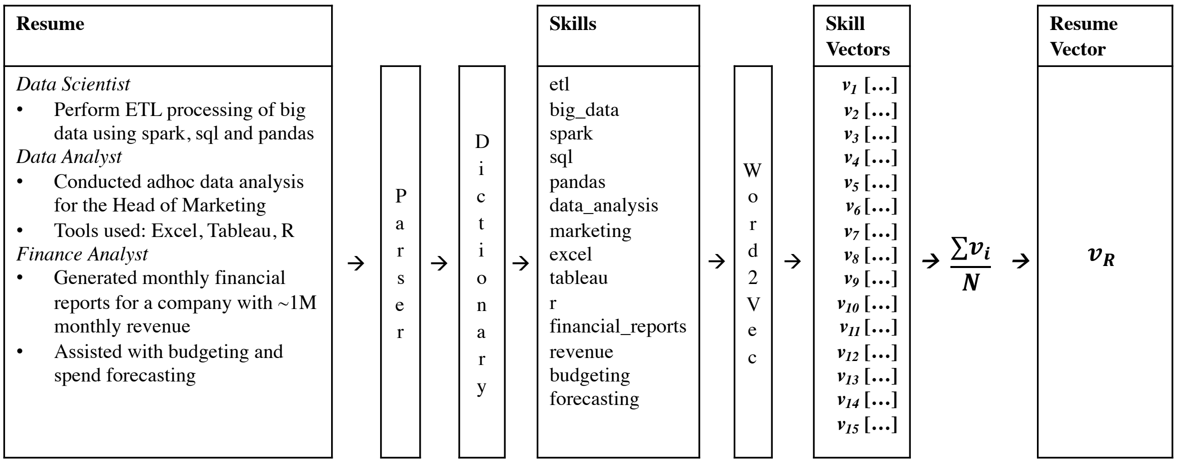

Figure 1: Illustration of Resume Vector Representation

B Model Specifications

For the Tf-Idf + Logistic model, we tried different classifiers including random forest, naive Bayes, SVM, and MLP, and picked the elastic-net logistic regression with mixing parameter `l1=0.5` based on 5-fold cross-validation.

For the word embedding model, we use Google's Word2Vec model[6] as the baseline, and Bolukbasi et al. (2016) for the debiased embeddings.

For the Longformer model, we follow Sun et al. (2019a) for hyperparameters and fine-tune by making small adjustments. The following parameters yielded the best results based on the area under ROC criteria: `Epochs=3, Batch Size=32, Learning Rate=2e-5, Weight Decay=2e-5`.

C List of Gender Indicating Words

"woman", "women", "womens", "she", "her", "girl", "girls", "sorority", "female", "hostess", "waitress", "mother", "saleswoman", "man", "men", "mens", "male", "boy", "boys", "guy", "he", "his", "him", "fraternity", "salesman", "father"

[6] https://code.google.com/archive/p/word2vec/

D SHAP Values vs. Word Embeddings

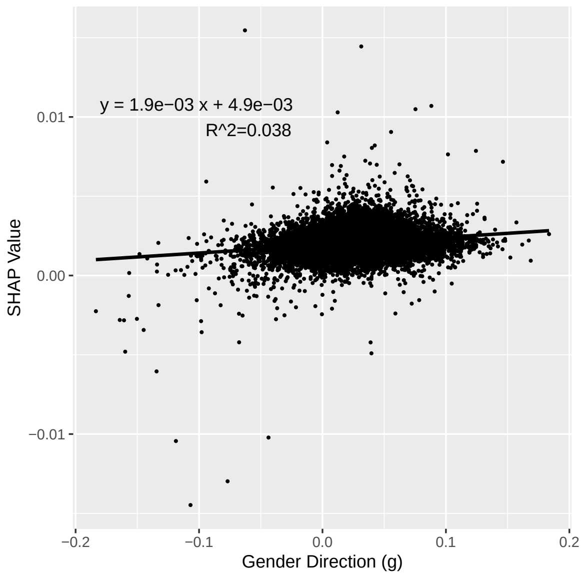

Figure 2: SHAP Value vs. Word Embedding Geneder Direction

We define the embedding gender-direction g of a token t with its vector representation $\vec{t}$ as follows:

$$g(t) = cos(\overrightarrow{man}, \overrightarrow{t}) - cos(\overrightarrow{woman}, \overrightarrow{t})$$

Gender predictive values (measured by SHAP) and gender direction of word embeddings have a slight positive relationship ($\beta = 1.9 \cdot 10^{-3}, p < 0.001$); more notably, however, there is very little correlation between the two measures ($R^2 = 0.038$).