The effects of gender bias in word embeddings on depression prediction

Firstname1 Lastname1*
Affiliation Name1
City1, Country1
firstname1.lastname1@domain.com

Firstname2 Lastname2
Affiliation Name2
City2, Country2
firstname2.lastname2@domain.com

Abstract

Word embeddings are extensively used in various NLP problems as a state-of-the-art semantic feature vector representation. Despite their success on various tasks and domains, they might exhibit an undesired bias for stereotypical categories due to statistical and societal biases that exist in the dataset they are trained on. In this study, we analyze the gender bias in four different pre-trained word embeddings specifically for the depression category in the mental disorder domain. We use contextual and non-contextual embeddings that are trained on domain-independent as well as clinical domain-specific data. We observe that embeddings carry bias for depression towards different gender groups depending on the type of embeddings. Moreover, we demonstrate that these undesired correlations are transferred to the downstream task for depression phenotype recognition. We find that data augmentation by simply swapping gender words mitigates the bias significantly in the downstream task.

1 Introduction

The gender differences and biases towards minorities in healthcare are heavily studied by researchers. The source of bias in healthcare can be due to multiple reasons such as gender differences in clinical trials and research Holdcroft (2007), professionals' unaware unfair treatment towards minorities DeAngelis (2019) or diagnosis based on symptoms of majority groups Arslanian-Engoren et al. (2006). Mental disorders are one of the health care categories that are heavily affected by societal and cultural norms. While many studies report gender inequalities Doering and Eastwood (2011) in the diagnosis of depression/anxiety, researchers also found that women take significantly more prescribed psychotropic drugs compared to men Bacigalupe and Martín (2021). These societal or statistical biases that exist in the real world and existing training resources are carried by ML models and consequently, this causes an unfair treatment of sensitive groups e.g., based on gender or race.

Bias can be exhibited in multiple parts of the natural language processing (NLP) models; from training data, and pre-trained word embeddings which are the core of state-of-the-art NLP models, resources, and algorithms themselves Chang et al. (2019). In addition, the final model's predictions can even amplify the biases present in the part of the pipeline Mehrabi et al. (2021). After the striking findings about gender bias in word embeddings for stereotypical occupations (e.g. female vectors are closer to nurse, while the male vectors are to doctor) by the study Bolukbasi et al. (2016), the research on the fairness of embeddings have accelerated. Many studies focused on various sub-problems of fairness such as quantifying bias in embeddings Caliskan et al. (2017); Kurita et al. (2019), fairness analysis in contextual Kurita et al. (2019); Basta et al. (2019); Zhao et al. (2019) or non-contextual embeddings Bolukbasi et al. (2016), methods for de-biasing embeddings Kaneko and Bollegala

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

(2019), and many more. Similar fairness analyses are also applied to clinical domain-specific embeddings Zhang et al. (2020); Agmon et al. (2022) and researchers studied bias in downstream clinical tasks such as in-hospital mortality prediction Zhang et al. (2020) and depression research using social media Aguirre et al. (2021).

Although there are many machine learning (ML) studies on depression diagnosis in the literature, there has been little attempt on analyzing these models in terms of fairness. A recent study analyzes the fairness of depression classifiers trained on social media, however, their focus is on the bias that exists in available depression datasets Aguirre et al. (2021). Therefore, our research question is whether word embeddings, which are commonly used in depression diagnosis Trotzek et al. (2018); Mallol-Ragolta et al. (2019), are gender biased for the depression category and if so how these biases are translated to the depression-related downstream tasks.

We summarize our contributions to this study as follows;

- For the first time, we analyze fairness in word embeddings for the depression category.
- We show that gender bias direction changes based on the dataset used to train embeddings.
- We perform a set of experiments repeated with different algorithms to examine the downstream effects of bias in embeddings using the depression phenotype recognition task.

2 Experimental Validation

In this section, we first explain the publicly available pre-trained embeddings used in our experiments and measures to quantify bias in embeddings. Then, we give details about the depression phenotype recognition task and the designed experiments to measure fairness in ML models.

2.1 Fairness Analysis in Embeddings

We used four different pre-trained embeddings in our experiments which are summarized below.

W2VecNews Mikolov et al. (2013) embeddings are trained on a part of the Google News dataset and contain 300-dimensional vectors for 3 million words and phrases.
BioWordVec Chen et al. (2019) are FastText Bojanowski et al. (2017) embeddings trained on PubMed corpus [2].
Clinical-BERT (Alsentzer et al., 2019) embeddings were trained on all available clinical notes of the MIMIC-III Clinical Database which consists of the medical notes describing the diagnosis and treatment of 46.520 patients at the Intensive Care Unit (Johnson et al., 2016).
W-BERT Devlin et al. (2018) is word-level contextualized BERT embeddings that are trained on Wikipedia and book corpus with masked language modeling objective. We used the 'bert-base-cased' model.

2.1.1 Direct Bias

To quantify bias in embeddings, we used the Direct Bias (DB) measure Bolukbasi et al. (2016), which is also used in many previous studies for both non-contextual and contextual embeddings Basta et al. (2019). *Direct bias* is computed by averaging the cosine similarity scores between the gender vector and the words belonging to the target category. We used the depression synonym list that was extracted by a recent study Moseley et al. (2020) as our target category list. The list consists of symptom-related words such as 'depressed', 'anxiety', and 'falling asleep'.

To compute the gender vector, we used gender pairs list[3] (e.g her-him, she-he). The difference vectors of the ten gender pairs are fed into the principal component analysis (PCA). The first eigenvector, which explains the majority of variance, represents a gender direction. The average absolute cosine similarity score between each word in the depression synonym list and the gender vector gives a DB score for the depression category.

The word vectors were extracted by using the aforementioned pre-trained embeddings. However, since contextual embeddings require context to obtain vectors for the given word, we created sentences

[2] available from https://github.com/ncbi-nlp/BioWordVec

[3] `https://github.com/tolga-b/debiaswe/blob/master/data/definitional_pairs.json`

Model	DB	original		swapped		neutralized		augmented	
		FNRR	F1	FNRR	F1	FNRR	F1	FNRR	F1
Clinical-BERT	0.04_M	0.81_F	0.64	0.61_M	0.66	0.62_M	0.64	0.97_F	0.63
W-BERT	0.02_F	0.95_F	0.58	0.93_M	0.58	0.76_F	0.59	0.97_F	0.59
BioWordVec	0.09_M	0.90_M	0.63	0.96_M	0.62	0.72_M	0.61	0.97_M	0.67
W2VNews	0.04_F	0.90_F	0.61	0.95_F	0.61	0.36_F	0.60	0.97_M	0.63

Table 1: Performance measures of trained SVM models alongside bias measures of embeddings. DB: Direct bias score of given word embedding, if embedding is biased towards the female, denoted with subscript $_F$, otherwise with $_M$, F1: macro-averaged F1 score. FNRR: false negative rate ratio. If the female group's FNRR is lower than the male group, this is denoted with $_F$, otherwise with $_M$.

containing gender or depression words. For gender pairs, we constructed simple sentences by swapping given gender pairs (e.g. he is a man, she is a woman). For depression words, we used a template that does not contain any gender pronouns yet can be used as a simple explanation of terms: "*X is a synonym of depression.*".

2.2 Downstream Task: Depression Phenotype Recognition

Evaluating the effect of bias in embeddings on depression-related problems is a difficult task as the bias can be transferred from multiple parts of the NLP model. To simplify our analysis, we needed a dataset in which gender information is explicitly given by gender pronouns but possibly does not exist in other words.

MIMIC-III-subset We used the subset of MIMIC-III Johnson et al. (2016) clinical notes events, which were annotated by two human experts (Moseley et al., 2020) for 15 clinical patient binary phenotypes. The clinical notes are written by a 3rd person (e.g. nurse or practitioner), thus gender pronouns to refer to the patient are extensively used. Since our focus in this study is depression, we only included the notes that were either labeled with 'depression' phenotype or 'none' phenotype. 'None' label means that no indication or cue was apparent to the annotator. This resulted in a total of 672 labeled clinical notes. For the training set, we used 90 'depression' and 90 'none' labeled notes for each gender group, in order to minimize bias towards a class and gender-group. In the depression phenotype recognition model, we expect the model to behave the same to the same notes with different gender pronouns. To measure the fairness from this angle, the remaining 312 notes were doubled by swapping gender pronouns with the opposite group's pronouns (e.g. he->she, him->her).

Experimental design To evaluate the effect of bias in embeddings to downstream tasks, we define 4 different experiments;

1. *original*: train a binary classifier on the original training data. 2. *swapped*: train a binary classifier on the training dataset in which gender pronouns in the original data were swapped with other gender group's pronouns. 3. *neutralized*: neutralize the data by either removing or replacing all gender pronouns with gender-neutral pronouns and train the classifier on this neutralized set. 4. *augmented*: train a binary classifier on the union of the original and swapped datasets.

We repeat the same set of experiments with different learners including Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP) to validate whether the findings are algorithm-specific. Hyperparameter tuning was done by 3-fold cross-validation on the training set for every model. For SVM, we only tuned C parameter (in [0.01, 100] range with exponential steps) and kernel (in {rbf, sigmoid, linear}). For RF, we tuned the maximum depth of the tree in the range of [1-50]. And, for MLP, we used 1 hidden layer having 100 neurons with ReLU activation and only tuned alpha parameters in the range of [0.1, 10].

3 Experimental Results

Direct Bias scores of each embedding method, alongside performance measures of trained SVM classifiers are given in Table 1 (for the results of MLP and RF models, see the Section 5). While we observe gender bias in all pre-trained embeddings, the magnitude of these biases varies. Although we cannot directly compare the contextual and non-contextual embeddings, we observe that BioWordVec

carries a higher bias than domain-independent W2VecNews. At the same time, Clinical-BERT is more biased towards gender groups compared to domain-independent W-BERT. Another interesting finding is that the bias direction is opposite between embeddings trained on the clinical dataset (closer to male) and domain-independent set (closer to female). Although depression is more prevalent in females based on medical literature, clinical embeddings trained on PubMed articles or MIMIC-III datasets show bias towards to male group.

Depression phenotype recognition models trained on original data have 10 to 30 mismatch predictions over 312 clinical note pairs with opposite gender pronouns. In other words, the model gives different predictions up to 30 notes when we change only gender pronouns in the test examples. This result motivates us to further analyze the bias in these models. To observe the effect of this bias direction exist in embeddings, we need to analyze the differences in 4 experiments namely *original, swapped, neutralized, and augmented*. Let's assume that none of the pre-trained embeddings are biased. In this case, 1. we expect to see a very similar score between *original* and *swapped* experiments with a change in the bias direction. Moreover, 2. for the *neutralized* experiment, we expect to see a very high false negative rate ratio (FNRR) between gender groups. Because, training data is free of gender information and consequently, ML models will not learn any undesired relations between gender pronouns and depression. Regarding the first point, we observe that the gender bias direction either does not change or the FNRR score changes largely for BioWordVec and W2VecNews embeddings for all the experiments with 3 different learners. Regarding the second point, we observe consistently lower false negative rate scores for the gender group that the embedding of the model is biased towards. These findings support the bias measures in the embeddings. Based on these results, we can say that gender bias in embeddings is transferred to downstream tasks by favoring one group with less false negative rates. Moreover, we expect to see improved FNRR with *augmented* experiment as the model is taught to make no difference based on gender pronouns by using identical notes with swapped gender. Similar to findings reported in Zhao et al. (2018), a simple augmentation approach (shown in the *augmented* experiment) improves the fairness of the models with an above 90% FNRR.

On the other hand, based on these results, we do not see any correlations between the Direct Bias score and the magnitude of change in the *swapped* or *neutralized* experiments, in other words, no strong correlation is found between the bias score and the observed effect of it in the downstream task. However, it should be noted that Direct Bias scores are generated by using pre-defined target dictionaries which might not generalize well to the downstream problem's dataset.

4 Discussion

In this study, we evaluated the gender bias in 4 different pre-trained embedding alternatives and showed their implications for the downstream task with a set of experiments. We draw a few conclusions from our experiments. First, we find that depending on the dataset these embeddings are trained on, the bias direction might be different. For the target depression category, while W2VecNews embeddings are biased towards females, BioWordVec models trained on the PubMed dataset show bias towards the male group. Second, although it is not possible to directly compare non-contextual and contextual embeddings, similar to previous studies, we find that contextual embeddings have lower bias scores compared to non-contextual. Finally, we found that bias magnitude and direction in embeddings affect the false negative rates for gender groups in the downstream task.

To analyze the effect of bias in embeddings on downstream problem, we proposed a set of experiments that changes gender information and helps us to observe the effect of different components easily for depression phenotype recognition problem. Our experiments show that when gender pronouns are removed from the training data, the model's bias direction is in line with the embeddings and we observe lower false negative rates for gender group that embeddings are biased to.

As a bias mitigation method, we simply augmented the training dataset by gender swapping and showed that it improves the fairness of the models by increasing FNRR markedly. On the other hand, applying this method to other problems, in which text is written by the person itself and thus gender information is spread to all text implicitly, might be quite challenging (e.g. depression recognition from social media data).

References

Shunit Agmon, Plia Gillis, Eric Horvitz, and Kira Radinsky. 2022. Gender-sensitive word embeddings for healthcare. *Journal of the American Medical Informatics Association* 29, 3 (2022), 415–423.

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. *arXiv preprint arXiv:2103.10550* (2021).

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

Cynthia Arslanian-Engoren, Amisha Patel, Jianming Fang, David Armstrong, Eva Kline-Rogers, Claire S Duvernoy, and Kim A Eagle. 2006. Symptoms of men and women presenting with acute coronary syndromes. *The American journal of cardiology* 98, 9 (2006), 1177–1181.

Amaia Bacigalupe and Unai Martín. 2021. Gender inequalities in depression/anxiety and the consumption of psychotropic drugs: are we medicalising women's mental health? *Scandinavian journal of public health* 49, 3 (2021), 317–324.

Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783* (2019).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5 (2017), 135–146.

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. *Advances in neural information processing systems* 29 (2016), 4349–4357.

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

Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–5.

T DeAngelis. 2019. How does implicit bias by physicians affect patients' healthcare. *Monit. Psychol* 50, 3 (2019), 22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Lynn V Doering and Jo-Ann Eastwood. 2011. A literature review of depression, anxiety, and cardiovascular disease in women. *Journal of Obstetric, Gynecologic & Neonatal Nursing* 40, 3 (2011), 348–361.

Anita Holdcroft. 2007. Gender bias in research: how does it affect evidence based medicine? , 2–3 pages.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742* (2019).

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337* (2019).

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. (2019).

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

Edward Moseley, Leo Anthony Celi, Joy Wu, and Franck Dernoncourt. 2020. Phenotype annotations for patient notes in the MIMIC-III database. *PhysioNet* (2020).

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia.. In *CLEF (Working Notes)*.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*. 110–120.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310* (2019).

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).

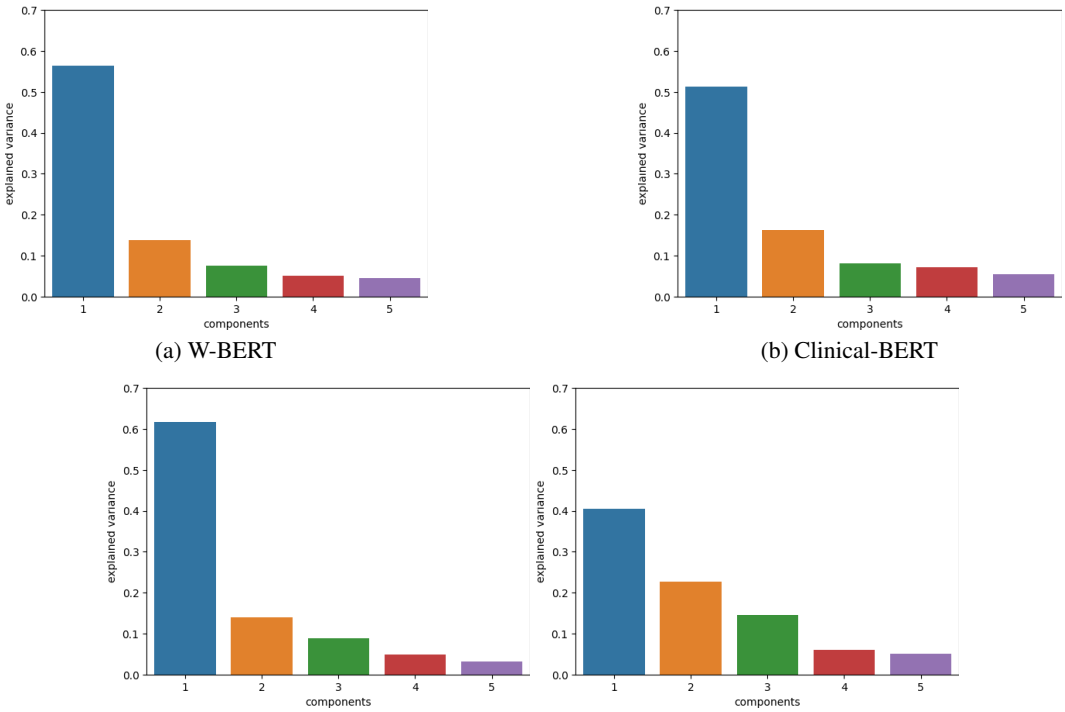
5 Supplementary Material

Similarly to the findings by Bolukbasi et al. (2016), Fig. 1 shows that the first eigenvalue of PCA is significantly larger than the other components and that there is a single direction describing the majority of variance in all these vectors. However, we observe that the difference between the percentage of variances is higher for domain-independent embeddings compared to their clinical-domain-specific versions.

Model	DB	original		swapped		neutralized		augmented	
		FNRR	F1	FNRR	F1	FNRR	F1	FNRR	F1
Clinical-BERT	0.04_M	0.93_F	0.60	0.91_M	0.63	0.92_M	0.62	0.94_M	0.64
W-BERT	0.02_F	0.97_M	0.60	0.94_F	0.58	1.00	0.58	0.97_F	0.58
BioWordVec	0.09_M	0.88_M	0.64	0.94_M	0.65	0.79_M	0.64	0.95_F	0.64
W2VNews	0.04_F	0.91_F	0.62	0.90_F	0.62	0.88_F	0.64	0.90_F	0.65

Table 2: Performance measures of trained RF models alongside bias measures of embeddings. DB: Direct bias score of given word embedding, if embedding is biased towards the female, denoted as $_F$ otherwise $_M$, F1: macro-averaged F1 score. FNRR: false negative rate ratio. If the female groups´s FNRR is lower than the male group, this is denoted as $_F$, otherwise, $_M$.

Figure 1: Sorted eigenvalues (explained variance) per embedding method for the dataset obtained with the difference of embedding vectors of the gender pair words.



(a) W-BERT (b) Clinical-BERT

(c) W2VecNews (d) BioWordVec

Model	DB	original		swapped		neutralized		augmented	
		FNRR	F1	FNRR	F1	FNRR	F1	FNRR	F1
Clinical-BERT	0.04_M	0.93_F	0.7	0.96_M	0.69	0.92_M	0.68	0.96_M	0.71
W-BERT	0.02_F	0.96_F	0.66	0.94_M	0.66	0.87_F	0.64	1.00	0.69
BioWordVec	0.09_M	1.00	0.69	0.88_F	0.69	1.00	0.60	0.92_F	0.70
W2VNews	0.04_F	0.61_F	0.69	0.97_M	0.69	0.37_F	0.64	0.93_F	0.68

Table 3: Performance measures of trained MLP models alongside bias measures of embeddings. DB: Direct bias score of given word embedding, if embedding is biased towards the female, denoted as $_F$ otherwise $_M$, F1: macro-averaged F1 score. FNRR: false negative rate ratio. If the female groups´s FNRR is lower than the male group, this is denoted as $_F$, otherwise, $_M$.