The effects of gender bias in word embeddings on patient phenotyping in the mental health domain

Gizem Sogancioglu, Heysem
Kaya and Albert Ali Salah
Utrecht University
Utrecht, The Netherlands

***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 for a range of affective computing tasks in the mental health domain including the detection of psychiatric disorders such as depression, and alcohol/substance abuse. 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 a bias towards different gender groups depending on the type of embeddings and the dataset that are trained on. Moreover, we demonstrate that these undesired associations are transferred to the downstream tasks and can even be intensified during supervised training for patient phenotyping. We find that data augmentation by simply swapping gender words mitigates the bias significantly in the downstream tasks.**

***Index Terms*—- fairness, bias mitigation, gender bias, bias in mental health, fairness in machine learning**

I. Introduction

Biases related to gender and demographics in healthcare systems are potentially harmful to the society. Such biases can arise from gender differences in clinical trials and research [1], from differential treatment towards minorities [2] or from diagnosis criteria based on analysis of symptoms of a majority group [3]. Mental disorders are one of the healthcare categories that are heavily affected by societal and cultural norms. While many studies report gender inequalities [4] in the diagnosis of depression/anxiety, researchers also found that women take significantly more prescribed psychotropic drugs compared to men [5]. When designing affective computing applications focusing on mental healthcare, societal or statistical biases can creep into machine learning (ML) models, which may cause unfair treatment of groups based on gender or race.

While ML approaches in the mental health domain use different modalities [6], we focus on texts and Natural Language Processing (NLP) based processing in this paper. These approaches include, for example, depression diagnosis [7], [8], suicide risk prediction [9], and alcohol misuse detection [10]. Bias can be exhibited in multiple parts of NLP models in such applications, including the algorithms themselves [11]. For biases related to data, an important consideration is the potential bias in pre-trained word embeddings, which are the core of many state-of-the-art NLP models [12]–[14]. Our goal in this study is to shed light on biases in NLP models for a set of downstream mental health tasks, focusing on binary patient phenotyping problems for common psychiatric disorders, and using clinical notes as the main data source.

Specifically, we aim to understand whether word embeddings and trained models' biases reflect the prevalence rates in the mental health literature, and how these biases affect the models' behaviors qualitatively. Our definition of fairness for patient phenotyping is that a fair ML model should behave the same, given the same clinical notes that differ only in gender pronouns. This definition is in line with counterfactual token fairness in the literature [15], [16]. To this extent, we experimentally analyze the fairness of four phenotype classifiers, which are among the most prevalent psychiatric disorders [17], namely, depression, alcohol abuse, substance abuse, and other psychiatric disorders excluding depression.

For each phenotype classification, we conduct experiments using four different pre-trained embeddings. For the problem of patient phenotyping from clinical notes, we propose to neutralize the training data from gender terms to observe how the bias in word embeddings is translated into downstream classifiers. Since gender information is explicitly given by gender pronouns in clinical notes, it is easier to neutralize data for such a problem. Furthermore, as a bias mitigation approach, we use data augmentation, which was shown to be successful for co-reference resolution problems [18].

The contributions of our study are outlined as follows:

- We comprehensively examine the downstream effects of bias in embeddings using a set of phenotype recognition tasks, namely, *depression, alcohol abuse, substance abuse, and psychiatric disorders*. We show that these bias directions are not in line with prevalence rates reported in mental health literature for some types of embeddings.
- We demonstrate that augmentation of training data by swapping gender words is a simple yet effective method to mitigate the bias in such downstream tasks.
- We qualitatively analyze the model behaviors and show that even in the case of correct classification, some models seem to be unreliable due to existence of words that highly impact the decision.

This is the uncorrected author proof. Please cite as: Sogancioglu, G., H. Kaya and A.A. Salah, The effects of gender bias in word embeddings on patient phenotyping in the mental health domain, Int. Conference on Affective Computing & Intelligent Interaction (ACII), 2023.

II. Related Work and Background

In this section, we first summarize the recent works on fairness in NLP problems and specifically in the mental health domain. Then, we give a general background on fairness measures.

A. Fairness in NLP models

Fairness studies in NLP can be mainly grouped into two classes, namely, fairness in downstream models and fairness in word embeddings. Fairness in downstream models was extensively studied for a wide range of NLP problems recently, including hate speech detection [19], co-reference resolution [20], and machine translation [21]. Fairness was also studied for clinical NLP problems such as mortality risk prediction models [22], [23], anxiety prediction [24], and depression research using social media [25].

Following the findings of Bolukbası et al. about gender bias in word embeddings for stereotypical occupations (e. g., female vectors are closer to nurse, while the male vectors are to doctor) [26], many studies focused on various sub-problems of fairness such as quantifying bias in embeddings [27], [28], fairness analysis in contextual embeddings [28]–[30], methods for de-biasing embeddings [31] etc. Similar fairness analyses are also applied to clinical domain-specific embeddings [23], [32].

The closest work to ours [23] evaluated the fairness of BERT embeddings for the mental health category. It was shown that contrary to general domain BERT embeddings, clinical-BERT [33] was biased towards females. However, mental health is a very broad domain and we expect that bias direction will change from one disorder to another, as prevalence rates of gender groups are not the same for some mental disorders (e. g., depression is more prevalent for females, while alcohol abuse is more common for males [17]). Differently from [23], we perform in this paper a more in-depth analysis for the mental health domain and show the association between biases in word embeddings and the downstream tasks.

B. Fairness measures

A recent review paper lists ten definitions of fairness and its measures [34]. However, in practice, it is not possible to satisfy multiple fairness measures at the same time [35]. The appropriate fairness measure for a model depends on its use case and the definition of fairness for such an application [36].

The fairness notions can be mainly categorized into “group” and “individual” fairness, based on the definition of bias. The two most common group fairness measures are *demographic parity* and *equal opportunity*, respectively. *Demographic parity*, which is also known as statistical parity, is satisfied if the likelihood of a positive outcome is the same regardless of group and ground-truth value. On the other hand, the *equal opportunity* definition states that each group of given sensitive attribute should have equal true positive rates, while *equalized odds* expects the same equal rates additionally for false positives. On the other hand, *counterfactual fairness* [15] aims to satisfy individual fairness, and a decision is considered counterfactually fair towards an individual if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group. In this study, we measure *counterfactual token fairness* [16], which is based on the idea of simply creating perturbations by substituting tokens associated with identity groups.

Apart from studies on quantitative fairness measures for NLP models, a set of fairness measures were developed for contextual embeddings [23], [37] and non-contextual [26], [27] embeddings. In this study, we use the *direct bias* measure proposed by [26], as it is one of the popular measures and can easily be applied to both contextual and non-contextual embeddings:

$$DirectBias = \frac{1}{N}\sum_{w \epsilon N} |\cos(\overrightarrow{w}, g)| \qquad (1)$$

Direct bias, whose formula is given in Eq. 1, is computed by averaging the cosine similarity scores between the gender vector (g) and the words ($\overrightarrow{w}$) belonging to the target category. To compute the gender vector, a gender pairs list[1] (e. g., her-him, female-male) is used. Then, the embedding difference vectors of ten gender pairs are fed into a principal component analysis (PCA). The first eigenvector, which explains the majority of variance, represents a gender direction and this vector is referred to as the *gender vector*. The average absolute cosine similarity score between each word ($\overrightarrow{w}$) in the target category list and the gender vector (g) gives a Direct Bias score for the target category. If a synonym in the target category list does not carry any gender information, then we expect it to be orthogonal to the gender vector.

III. Experimental Setting

In this section, we first explain the dataset used for experiments. Next, we describe the features and experimental choices made to train phenotyping models. Finally, we introduce our experimental design, with which we aim to quantify the effect of bias introduced by word embeddings on the downstream models and to evaluate the bias mitigation methods thereof.

A. Dataset: MIMIC-III

The Multiparameter Intelligence Monitoring in Intensive Care (MIMIC-III) [38] Clinical Database consists of clinical note events that describe the diagnosis and treatment of more than 40.000 adult patients at the Intensive Care Unit of Beth Israel Deaconess Medical Center between 2001 and 2012. The clinical notes are of varying types, such as discharge summaries, nursing notes, and radiology reports. Moreover, patient demographic information and ICD-9 codes are stored in electronic health records. ICD-9 codes are assigned by the billing department and although they represent the conditions or treatments that patients have, much of the information is only present in the clinical notes. For this reason, a total of 844 examples were annotated by two groups consisting of both

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

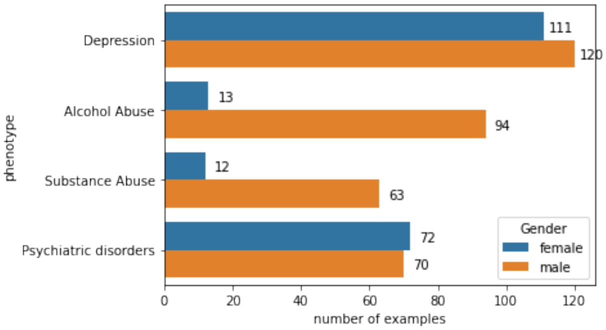


Fig. 1. The number of annotated phenotypes in the labeled set.

physicians and clinical researchers [39] for 13 clinical patient binary phenotypes including mental disorders and physical diseases, such as heart disease. If a phenotype occurs in the clinical note, this was annotated as 1, otherwise 0.

We used the most voted class among different annotations as a ground-truth value for the corresponding clinical note. In case of a tie, the positive class was used, i. e., we broke the tie in favor of the minority class. Since our focus is mental health in this study, we included four phenotypes among 13 for binary phenotype recognition problems, namely, ‘depression’, ‘alcohol abuse’, ‘substance abuse’, and ‘psychiatric disorders’ other than depression. The description of each phenotype is given in Table I [39]. The number of annotated examples for each phenotype is shown in Fig. 1. The examples which were annotated as ‘None’, meaning no indication or cue was apparent to the annotator, were used as negative examples.

The number of examples used for training and test sets for each phenotype dataset are shown in Table II. In order to minimize bias toward a class and gender group, we randomly split the subset (consisting of both positive and ‘None’ labeled examples for a given class) into training and test sets by preserving a similar number of examples for positive, negative, female, and male examples for the training set.

The clinical notes were written by the nurse or the practitioner, thus gender pronouns to refer to the patient are explicitly used. In a fair phenotype recognition model, we expect the model to behave the same given two clinical notes that differ only in gender pronouns. To measure the fairness from this angle, the reported test examples were doubled by swapping gender pronouns with the opposite group’s pronouns (e. g., he→she). The full list of gender terms used in the study is given in Table III. As an example; if the original clinical note has the term ‘he’, this was replaced by ‘she’.

B. Downstream tasks: mental health phenotype recognition

Following the baseline study [40], we train separate binary classification models for the recognition of each phenotype. We apply simple pre-processing techniques to clean the dataset before feature extraction; removal of digits, special symbols, punctuations, and one-character terms. As features, we extract four commonly used pre-trained word embeddings with different properties for comparability purposes, which are summarized below:

- **W2VecNews** [41] are non-contextual embeddings that are trained on a part of the Google News dataset and contain 300-dimensional vectors for 3 million words and phrases.
- **BioWordVec** [42] are non-contextual FastText [43] embeddings trained on PubMed corpus[2]. Each word is represented as a 200-dimensional vector.
- **Clinical-BERT** [33] 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 [38]. The extracted contextual word vectors are 768-dimensional.
- **W-BERT** [44] is word-level contextual BERT embeddings that are trained on Wikipedia and book corpus with masked language modeling objectives. We used the ‘bert-base-cased’ model, which is available in the HuggingFace interface[3]. The word vectors are 768-dimensional.

To extract features for a clinical note, we first obtain each word vector from the pre-trained word embeddings, then the average vector representation is used as a feature representation of the given text. The vectors are z-normalized using the parameters estimated from the respective training set and four different Support Vector Machine (SVM) models are trained for each phenotype classification. Hyperparameter tuning was done by 3-fold cross-validation on the training set for every model. We only tuned the C parameter (in [0.01, 100] range with exponential steps) and kernel (in {rbf, sigmoid, linear}). These trained models are referred to as *original* models in our experiments since they are trained on the *original* data.

1) Performance and fairness measures for phenotyping models: As a performance measure and to tune all models, we use macro-averaged F1 score since it is commonly used in the literature in case of an imbalanced dataset.

To quantify the fairness of trained ML models in terms of counterfactual token fairness, we use mismatch ratio, namely the number of pairs with mismatched predictions divided by the total number of pairs. Moreover, we computed True Positive Rate Ratio (TPRR) and False Positive Rate Ratio (FPRR), whose formulas are given in Equation 4.

$$TPR_i = TP_i/(FN_i + TP_i) \tag{2}$$

$$FPR_i = FP_i/(FP_i + TN_i) \tag{3}$$

$$(T|F)PRR = (T|F)PR_{disadvantaged}/(T|F)PR_{advantaged} \tag{4}$$

The disadvantaged group’s performance was divided over the advantaged group’s performance to make sure the score is in [0, 1] range. The higher the ratios, the fairer the model’s predictions. These measures are commonly used as group fairness measures as explained in Section II-B. However, we compute them on the augmented test dataset which consists

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

[3] https://huggingface.co/bert-base-cased

TABLE I
The descriptions of phenotypes that are used for identifying and annotation [39].

Phenotype	**Definition**
Depression	Diagnosis of depression; prescription of anti-depressant medication; or any description of intentional drug overdose, suicide, or self-harm attempts
Alcohol Abuse	Current/recent alcohol abuse history; still an active problem at the time of admission (may or may not be the cause of it).
Substance Abuse	Include any intravenous drug abuse (IVDU), accidental overdose of psychoactive or narcotic medications (prescribed or not). Admitting to marijuana use in history is not sufficient.
Psychiatric disorders	All psychiatric disorders in DSM-5 classification, including schizophrenia, bipolar, and anxiety disorders, other than depression.

TABLE II
The number of training and test examples per gender for each phenotype class. M: male, F: female.

	Training				Test			
	Positive		Negative		Positive		Negative	
	F	M	F	M	F	M	F	M
Depression	90	90	90	90	20	31	39	121
Alcohol abuse	10	50	50	50	3	44	47	116
Substance abuse	10	50	50	50	2	13	47	116
Psychiatric disorders	55	55	55	55	17	15	41	112

TABLE III
The list of female, male, and neutralized terms that are used for gender swapping and gender neutralization. Swapping: female terms ⟷ male terms, Neutralization: (female or male) terms → neutralized terms.

Male terms	Female terms	Neutralized terms
he	she	patient
man	woman	patient
his	her	its
him	her	its
male	female	SPACE CHAR
Sex: M	Sex: F	Sex: SPACE CHAR

of counterfactual pairs, thus alongside the mismatch ratio measure, we can obtain the performance of advantaged and disadvantaged groups in case of mismatch.

2) Measuring fairness in pre-trained word embeddings: To quantify bias in embeddings, we used the Direct Bias (DB) measure [26], which is also explained in detail in Section II-B. We used the same gender pairs as in the study [26] to obtain the gender vector. To compute the DB, we also need a synonym list for the target concept. Since we focus on mental health, it is important that extracted synonyms are verified by domain experts. Thanks to a recent study that made a depression synonym list publicly available [45], we could measure DB for the depression task. The list consists of symptom-related words such as 'depressed' and 'anxiety'.

The word vectors for gender and synonyms were extracted using the aforementioned pre-trained embeddings. However, since contextual embeddings require context to obtain vectors for the given word, we created template sentences 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 the terms: "*X is a synonym of depression.*". Then, as explained in Section II-B, the average cosine similarity between the words in the synonym list and the gender vector is used as the DB score.

C. Experimental design

To analyze the model behavior and bias introduced by embeddings or training data, we substituted the gender terms mentioned in the notes of training data and trained different classifiers. However, all these models are evaluated on the same augmented test set, which consists of counterfactual pairs. We summarize each model below.

1) *original*: train a binary classifier on the original training data as was explained in Section III-B.
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 (see Table III).
3) *neutralized*: neutralize the data by either removing or replacing all gender terms with gender-neutral counterparts and train the classifier on this neutralized set (see Table III).
4) *augmented*: train a binary classifier on the union of the original and swapped datasets. This approach is evaluated as a bias mitigation method as the model is taught not to make any differences explicitly for counterfactual pairs.

IV. Experimental Results

In this section, first, we present the results for fairness in word embeddings for the depression domain. Next, we show the results of the set of experiments conducted for five phenotype classifications to observe the effect of bias on downstream tasks. Finally, we provide qualitative analysis which we perform using LIME [46] to analyze the model behavior for an exemplar clinical note.

A. Fairness in word embeddings

Direct Bias scores of each embedding method for the depression domain are shown in Fig 2. While we observe gender bias in all pre-trained embeddings, the magnitude and direction of these biases vary. Although we cannot directly compare

TABLE IV
THE PERFORMANCE IN TERMS OF UNWEIGHTED AVERAGE RECALL (UAR) OF BINARY GENDER CLASSIFIERS TRAINED ON ORIGINAL DATA VS NEUTRALIZED DATASET FOR DEPRESSION SUBSET

Method	Trained on	Gender UAR
Baseline	-	0.50
W2VecNews	Original	0.98
	Neutralized	0.37
BioWordVec	Original	0.91
	Neutralized	0.32
W-BERT	Original	0.96
	Neutralized	0.36
Clinical-BERT	Original	0.97
	Neutralized	0.33

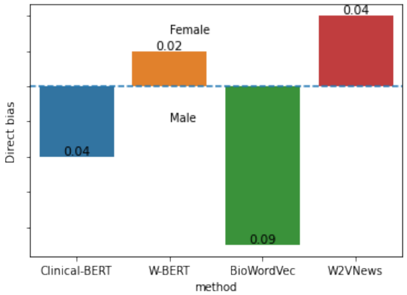


Fig. 2. *Direct bias* scores of word embeddings for *Depression* domain.

the fairness in 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 toward gender groups compared to domain-independent W-BERT. Furthermore, we observe different bias directions for embeddings trained on clinical datasets (closer to male) and domain-independent sets (closer to female). Although depression is more prevalent in females based on medical literature [4], clinical embeddings trained on PubMed articles or MIMIC-III datasets show bias towards the male group.

B. Fairness in phenotype recognition tasks

The performance and fairness measures of trained SVM models are given in Table V. We observe that while a few models (e.g. W-BERT model for Depression) trained on original data are fair in terms of our fairness measures, most of them have surprisingly low TPRR/FPRR and high mismatch ratio scores. As a striking example; the W-BERT model trained on the original set has a mismatch ratio of 14% for alcohol abuse phenotyping. Example-wise, it gives different predictions for 30 clinical note pairs out of 210 pairs when we substitute only the gender pronouns. This result motivates us to further analyze the source of bias in these models.

To understand the models' behavior and the source of biases, we need to analyze the differences in results of the designed four experiments, namely *original, swapped, neutralized, and augmented* that are introduced in Section III-C. Let's assume that none of the pre-trained embeddings are biased. In this case, (Assumption 1.) we expect to see very similar fairness scores for *original* and *swapped* experiments with a change in the bias direction. Moreover, (Assumption 2.) for the *neutralized* experiment, we expect to see a very low mismatch ratio and very high TPRR/FPRR scores. Because, we assume that training data is free of gender information and consequently, ML models will not learn any undesired associations between gender pronouns and the target domain and make the same predictions for counterfactual pairs.

To validate whether our neutralization algorithm was efficient and whether neutralized training data is free of gender information, we trained binary gender classifiers using the SVM classifier with original vs neutralized training features and evaluated the performance on the original test dataset of the depression subset. The results are given in Table IV. As a baseline method, we used the model that returns constant labels for all examples, and the Unweighted Average Recall (UAR) equals 0.50 (1 over the number of classes). The gender classifiers trained on original data obtained very high scores, reaching above 0.90 UAR for all embeddings. On the other hand, models trained on the neutralized dataset showed much worse performance than even the baseline classifier. These results show that although there might be still gender-prevalent words (e. g., diseases only prevalent in one gender group) after data neutralization by removing gender pronouns, the retained linguistic information is insufficient to infer any gender-related associations by the downstream model.

Having the Direct Bias (DB) scores for the depression domain, we first take a closer look into the results of depression phenotyping models. Regarding the first point (Assumption 1), we observe that the gender bias direction does not change for BioWordVec and W2VecNews models. Regarding the second point (Assumption 2), we observe consistently lower TPR and FPR scores for the gender group that the embedding of the model is biased towards. These findings are in line with the DB measures of the embeddings. Based on these results, we can say that gender bias in embeddings is transferred to downstream tasks by favoring one group with higher positive predictions. On the other hand, based on these results, we do not see any correlations between the DB score and the magnitude of fairness measures in *neutralized* experiments. In other words, no correlation is observed between the bias score and its observed effect for the downstream task. However, it should be noted that intrinsic DB scores are computed using a pre-defined target synonym list, which might not generalize well to the downstream problem's dataset and the terms that are found important by the ML model.

For *alcohol abuse and substance abuse* phenotyping experiments, we observed that all models trained on original data are biased by making more positive predictions for males. This result is highly likely due to the higher number of positive examples in the training set, which is in line with prevalence rates in the literature [47]. Moreover, based on neutralized experiment results, we observed that unlike domain-independent W2VecNews and W-BERT, embeddings trained on clinical data, namely BioWordVec and Clinical-

TABLE V

PERFORMANCE MEASURES OF TRAINED SVM MODELS ALONGSIDE BIAS MEASURES OF EMBEDDINGS. C-BERT: CLINICAL-BERT, F1: MACRO-AVERAGED F1 SCORE. TPRR: TRUE POSITIVE RATE RATIO. $\neq$: MISMATCH RATIO. IF THE FEMALE GROUP'S TPR/FPR/POSITIVE CLASS PROBABILITY FOR THE MISMATCH IS GREATER THAN THE MALE GROUP, THIS IS DENOTED WITH $_F$, OTHERWISE WITH $_M$. THE HIGHER TPRR AND FPRR, AND THE LOWER THE MISMATCH RATIO, THE HIGHER FAIRNESS IS ACHIEVED. THE HIGHEST TPRR, FPRR, AND LOWEST MISMATCH AMONG EXPERIMENTS FOR EACH MODEL WERE SHOWN IN **BOLD**.

		Depression				Psychiatric disorders				Alcohol abuse				Subtance abuse			
		TPRR↑	FPRR↑	$\neq\downarrow$	F1↑	TPRR↑	FPRR↑	$\neq\downarrow$	F1↑	TPRR↑	FPRR↑	$\neq\downarrow$	F1↑	TPRR↑	FPRR↑	$\neq\downarrow$	F1↑
W2VecNews	Orig.	0.90$_F$	0.91$_F$	0.04$_F$	0.65	**1.00**	0.94$_F$	0.02$_F$	0.57	0.81$_M$	0.67$_M$	0.08$_M$	0.72	0.92$_M$	0.89$_M$	0.03$_M$	0.58
	Swap.	0.87$_F$	0.94$_F$	0.05$_F$	0.63	**1.00**	0.89$_F$	0.04$_F$	0.57	0.76$_F$	0.56$_F$	0.13$_F$	0.72	0.92$_F$	0.65$_F$	0.06$_F$	0.71
	Neutr.	0.70$_F$	0.51$_F$	0.25$_F$	0.64	0.92$_F$	0.90$_F$	0.04$_F$	0.57	0.94$_F$	0.94$_F$	0.02$_F$	0.75	0.92$_F$	0.65$_F$	0.06$_F$	0.71
	Aug.	**1.00**	**1.00**	**0.00**	0.69	**1.00**	**1.00**	**0.00**	0.74	**0.97**$_M$	**1.00**	**0.00**	0.73	**1.00**	**1.00**	**0.00**	0.69
BioWordVec	Orig.	0.97$_M$	**1.00**	0.01$_M$	0.69	0.92$_F$	0.89$_F$	0.04$_F$	0.73	0.91$_M$	0.71$_M$	0.07$_M$	0.75	**1.00**	0.85$_M$	0.02$_M$	0.76
	Swap.	0.97$_M$	0.96$_M$	0.02$_M$	0.69	0.96$_F$	0.94$_F$	0.01$_F$	0.74	**1.00**	0.92$_F$	0.01$_F$	0.71	**1.00**	0.93$_F$	0.01$_F$	0.73
	Neutr.	0.92$_M$	0.83$_M$	0.03$_M$	0.69	0.90$_F$	0.77$_F$	0.03$_F$	0.72	0.95$_M$	0.95$_M$	0.02$_M$	0.75	0.83$_M$	0.90$_M$	0.01$_M$	0.79
	Aug.	**1.00**	**1.00**	**0.00**	0.72	0.96$_M$	**1.00**	0.02$_M$	0.72	**1.00**	**1.00**	**0.00**	0.70	**1.00**	**1.00**	**0.00**	0.75
W-BERT	Orig.	**1.00**	**1.00**	**0.00**	0.69	0.94$_F$	0.97$_F$	0.01$_F$	0.60	0.76$_M$	0.38$_M$	0.14$_M$	0.72	0.75$_M$	0.62$_M$	0.10$_M$	0.64
	Swap.	0.94$_F$	0.97$_F$	0.02$_F$	0.66	0.94$_F$	0.90$_F$	0.03$_F$	0.61	0.78$_F$	0.48$_F$	0.11$_F$	0.71	0.75$_F$	0.57$_F$	0.11$_F$	0.64
	Neutr.	0.97$_F$	0.92$_F$	0.02$_F$	0.68	**1.00**	**1.00**	0.01$_F$	0.60	**1.00**	0.81$_F$	0.03$_F$	0.72	**1.00**	0.87$_F$	0.01$_F$	0.68
	Aug.	**1.00**	**1.00**	**0.00**	0.66	**1.00**	**1.00**	**0.00**	0.66	**1.00**	**0.95**$_M$	**0.01**$_M$	0.70	**1.00**	**1.00**	**0.00**	0.68
C-BERT	Orig.	0.97$_M$	**1.00**	0.01$_M$	0.67	0.86$_F$	0.86$_F$	0.05$_F$	0.63	0.86$_M$	0.65$_M$	0.07$_M$	0.76	0.92$_M$	0.64$_M$	0.04$_M$	0.77
	Swap.	**1.00**	0.96$_F$	0.01$_F$	0.69	0.90$_M$	0.83$_M$	0.06$_M$	0.63	0.92$_F$	0.72$_F$	0.06$_F$	0.75	**1.00**	**0.90**$_F$	0.03$_F$	0.77
	Neutr.	0.89$_M$	0.83$_M$	0.07$_M$	0.68	0.92$_F$	0.79$_F$	0.08$_F$	0.60	0.97$_M$	0.92$_M$	0.01$_M$	0.79	0.91$_M$	0.85$_M$	0.03$_M$	0.71
	Aug.	0.97$_M$	**1.00**	**0.01**$_M$	0.70	**1.00**	**1.00**	**0.01**$_M$	0.64	**1.00**	**1.00**	**0.00**	0.73	**1.00**	**0.90**$_M$	**0.01**$_M$	0.76

BERT, are biased towards males. Although gender prevalence rates are different for *depression and alcohol/substance abuse*, the bias direction of embeddings is the same for these domains. We think that pronouns and symptoms might co-occur more for the male group in these clinical datasets (Pubmed, MIMIC-III), which causes the embeddings to learn spurious associations. However, we should note that although the neutralized experiment gives a good idea about embeddings bias, the domain is restricted to the synonyms/words that the model learned.

On the other hand, as shown in Table I, *psychiatric disorders* contains a group of mental disorders excluding depression. For this reason, this category is more difficult to compare with mental health literature. However, interestingly, despite the balanced training dataset, models trained on the original, or neutralized dataset with different embeddings, make consistently a higher number of positive predictions for females.

Moreover, we expect to see improved fairness measures 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 [16], [18], a simple augmentation approach (shown in the *augmented* experiment) consistently improves the fairness of the models with TPRR and FPRR being mostly (close to) 1.0. We repeated the same set of experiments with two more learners, namely, Random Forest (RF), and Multilayer Perceptron (MLP) to validate whether the findings are algorithm-specific and observed consistent results across learners.

C. Qualitative analysis

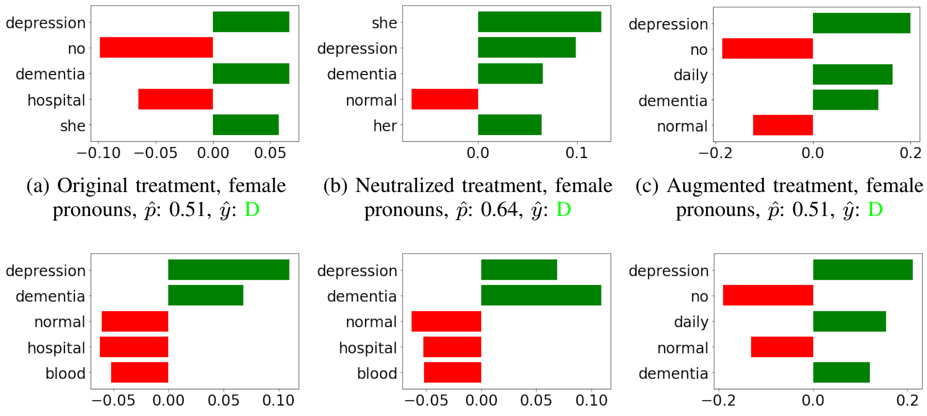
To understand the models' reasoning in case of unfair predictions, feature importances for a few examples were computed by LIME [46]. Fig. 3 shows an example that was predicted differently by the models trained on original or neutralized datasets for the same notes with opposite gender pronouns. Although original and neutralized models´ predictions are correct for a test example with female pronouns, explanations show that decisions were not only based on phenotype-related synonyms but also on undesired associations with gender pronouns. On the other hand, the model trained on the augmented dataset gives decisions based on the same set of (domain-relevant) top terms for the counterfactual pair and gives similar importances to gender pronouns.

V. DISCUSSION

In this study, we evaluated the gender bias in four popular pre-trained embeddings and showed their implications for a set of patient phenotyping tasks namely depression, alcohol/substance abuse, and other psychiatric disorders. To analyze the effect of bias in embeddings on the phenotyping tasks, we proposed a set of experiments that substituted gender information and helped us to observe the effect of different components easily.

We draw a few conclusions from our experiments. First, we observe that biases or even accurate prevalence rate differences present in datasets used for training might harm the downstream model by causing different predictions for the clinical notes that differ only in gender terms. Second, we find that the bias direction in embeddings changes based on the dataset they are trained on and some of them do not reflect the prevalence rates in the real world. For example; for the depression, alcohol, and substance abuse categories, embeddings trained on the clinical datasets show bias towards the male group, while domain-independent embeddings are biased towards the female group. Finally, we found that the models make more positive predictions for the gender group that their embeddings are biased towards (e.g. mostly similar bias directions for original/swapped/neutralized experiments exist for depression and psychiatric disorders).

The study [48] finds no correlation between intrinsic metrics and extrinsic measures and suggests focusing on extrinsic metrics for the task of bias mitigation. As mentioned earlier, embeddings may not be the only source of bias that affects the downstream model's behavior, and consequently, its effect



(a) Original treatment, female pronouns, $\hat{p}$: 0.51, $\hat{y}$: D

(b) Neutralized treatment, female pronouns, $\hat{p}$: 0.64, $\hat{y}$: D

(c) Augmented treatment, female pronouns, $\hat{p}$: 0.51, $\hat{y}$: D

(d) Original treatment, male pronouns, $\hat{p}$: 0.48, $\hat{y}$: ND

(e) Neutralized treatment, male pronouns, $\hat{p}$: 0.48, $\hat{y}$: ND

(f) Augmented treatment, male pronouns, $\hat{p}$: 0.51, $\hat{y}$: D

Fig. 3. Top 5 important words computed by LIME based on a given prediction of a test example by binary depression phenotype classifier (W2VecNews + SVM). The ground-truth label is D(epressed). The positive values colored in green denote a contribution to the positive class while words with negative values shown in red contribute to the negative class. D: Depressed, ND: Not Depressed, $\hat{p}$: probability of the positive (depressed) class, $\hat{y}$: predicted class.

can be eliminated by other bias sources (e.g. imbalance in training data dominates the downstream model's bias direction as shown in alcohol/substance abuse). However, we think that understanding these implications is crucial to understand the trained ML model better. Moreover, it helps us to understand the biases of the source dataset on which the embeddings are based.

As a bias mitigation method, inspired by [18], we simply augmented the training dataset by gender swapping and showed that it improves the fairness of the models by increasing fairness measures markedly without deteriorating the predictive performance. This is a much simpler approach compared to other debiasing methods in the literature [49] and is well-suited for the problems such as phenotype classification, and coreference resolution. On the other hand, applying this method to settings where a given text is written by a single person and thus gender information is spread to the entire text implicitly, is quite challenging (e.g. depression recognition from social media data).

VI. Ethical Impact Statement

Disentangling the sources of gender bias in NLP-based mental healthcare models is the main goal of this work. Therefore, the discussion in the former section not only covers our findings but also our ethical concerns on the uses of the mentioned methods for such critical downstream applications.

We would like to warn the reader about the limitations of our study. Due to dataset characteristics, we used a binary definition of gender, as gender pronouns are likely chosen based on the patient's sex. Moreover, we could only measure the direct bias score in word embeddings for the depression task, since the depression synonym list was crafted earlier by clinical experts and observed the same bias directions on the downstream models when trained on the neutralized dataset. The creation of these synonym lists, which clinicians use to make decisions, will not only help to improve the model's decisions, but also contribute to fairness and explainability literature largely.

There are also limitations of the bias analysis approach based on neutralizing or swapping the gender pronouns. Such a study can be conducted in many Western languages such as English or German, however, it is not directly applicable e. g., in Arabic, which has gendered adjectives and verbs.

Access to the MIMIC-III dataset is possible after completing a recognized course in protecting human research participants and signing a data use agreement, which prohibits publishing exemplar sentences even though they are de-identified. We thus conducted a word-level explainability analysis to show how the model's behavior changes in the counterfactual scenario and avoided sharing contextual information from the sentences.

References

[1] A. Holdcroft, "Gender bias in research: how does it affect evidence based medicine?" pp. 2–3, 2007.

[2] T. DeAngelis, "How does implicit bias by physicians affect patients' healthcare," *Monit. Psychol*, vol. 50, no. 3, p. 22, 2019.

[3] C. Arslanian-Engoren, A. Patel, J. Fang, D. Armstrong, E. Kline-Rogers, C. S. Duvernoy, and K. A. Eagle, "Symptoms of men and women presenting with acute coronary syndromes," *The American journal of cardiology*, vol. 98, no. 9, pp. 1177–1181, 2006.

[4] L. V. Doering and J.-A. Eastwood, "A literature review of depression, anxiety, and cardiovascular disease in women," *Journal of Obstetric, Gynecologic & Neonatal Nursing*, vol. 40, no. 3, pp. 348–361, 2011.

[5] A. Bacigalupe and U. Martín, "Gender inequalities in depression/anxiety and the consumption of psychotropic drugs: are we medicalising women's mental health?" *Scandinavian journal of public health*, vol. 49, no. 3, pp. 317–324, 2021.

[6] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems," *ACM Transactions on Computer-Human Interaction*, vol. 27, no. 5, pp. 1–53, 2020.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the international AAAI conference on web and social media*, vol. 7, no. 1, 2013, pp. 128–137.

[8] F. Van Steijn, G. Sogancioglu, and H. Kaya, "Text-based interpretable depression severity modeling via symptom predictions," in *Proceedings of the ICMI 2022*, 2022, pp. 139–147.

[9] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," *Biomedical informatics insights*, vol. 10, p. 1178222618792860, 2018.

[10] M. Afshar, A. Phillips, N. Karnik, J. Mueller, D. To, R. Gonzalez, R. Price, R. Cooper, C. Joyce, and D. Dligach, "Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 254–261, 2019.

[11] K.-W. Chang, V. Prabhakaran, and V. Ordonez, "Bias and fairness in natural language processing," in *Proceedings of the 2019 EMNLP-IJCNLP: Tutorial Abstracts*, 2019.

[12] G. Sogancioglu, F. Mijsters, A. van Uden, and J. Peperzak, "Gender bias in (non)-contextual clinical word embeddings for stereotypical medical categories," *arXiv preprint arXiv:2208.01341*, 2022.

[13] M. Trotzek, S. Koitka, and C. M. Friedrich, "Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia." in *CLEF (Working Notes)*, 2018.

[14] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," 2019.

[15] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *NEURIPS*, vol. 30, 2017.

[16] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 219–226.

[17] R. K. McHugh and R. D. Weiss, "Alcohol use disorder and depressive disorders," *Alcohol research: current reviews*, vol. 40, no. 1, 2019.

[18] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.

[19] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[20] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender bias in coreference resolution," *arXiv preprint arXiv:1804.09301*, 2018.

[21] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," *arXiv preprint arXiv:1906.00591*, 2019.

[22] H. Singh, R. Singh, V. Mhasawade, and R. Chunara, "Fairness violations and mitigation under covariate shift," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 3–13.

[23] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: quantifying biases in clinical contextual word embeddings," in *proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 110–120.

[24] K. Zanna, K. Sridhar, H. Yu, and A. Sano, "Bias reducing multitask learning on mental health prediction," in *Proc. ACII*. IEEE, 2022, pp. 1–8.

[25] C. Aguirre, K. Harrigian, and M. Dredze, "Gender and racial fairness in depression research using social media," *arXiv preprint arXiv:2103.10550*, 2021.

[26] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *NEURIPS*, vol. 29, pp. 4349–4357, 2016.

[27] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[28] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," *arXiv preprint arXiv:1906.07337*, 2019.

[29] C. Basta, M. R. Costa-Jussà, and N. Casas, "Evaluating the underlying gender bias in contextualized word embeddings," *arXiv preprint arXiv:1904.08783*, 2019.

[30] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Gender bias in contextualized word embeddings," *arXiv preprint arXiv:1904.03310*, 2019.

[31] M. Kaneko and D. Bollegala, "Gender-preserving debiasing for pre-trained word embeddings," *arXiv preprint arXiv:1906.00742*, 2019.

[32] S. Agmon, P. Gillis, E. Horvitz, and K. Radinsky, "Gender-sensitive word embeddings for healthcare," *Journal of the American Medical Informatics Association*, vol. 29, no. 3, pp. 415–423, 2022.

[33] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[34] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[35] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[36] P. Czarnowska, Y. Vyas, and K. Shah, "Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1249–1267, 2021.

[37] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency, "Towards debiasing sentence representations," *arXiv preprint arXiv:2007.08100*, 2020.

[38] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[39] E. Moseley, L. A. Celi, J. Wu, and F. Dernoncourt, "Phenotype annotations for patient notes in the mimic-iii database," *PhysioNet*, 2020.

[40] E. T. Moseley, J. T. Wu, J. Welt, J. Foote, P. D. Tyler, D. W. Grant, E. T. Carlson, S. Gehrmann, F. Dernoncourt, and L. A. Celi, "A corpus for detecting high-context medical conditions in intensive care patient notes focusing on frequently readmitted patients," *arXiv preprint arXiv:2003.03044*, 2020.

[41] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *NEURIPS*, vol. 26, 2013.

[42] Q. Chen, Y. Peng, and Z. Lu, "Biosentvec: creating sentence embeddings for biomedical texts," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2019, pp. 1–5.

[43] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[45] T. A. Koleck, N. P. Tatonetti, S. Bakken, S. Mitha, M. M. Henderson, M. George, C. Miaskowski, A. Smaldone, and M. Topaz, "Identifying symptom information in clinical notes using natural language processing," *Nursing research*, vol. 70, no. 3, p. 173, 2021.

[46] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.

[47] A. M. White, "Gender differences in the epidemiology of alcohol use and related harms in the united states," *Alcohol research: current reviews*, vol. 40, no. 2, 2020.

[48] S. Goldfarb-Tarrant, R. Marchant, R. M. Sánchez, M. Pandya, and A. Lopez, "Intrinsic bias metrics do not correlate with application bias," *arXiv preprint arXiv:2012.15859*, 2020.

[49] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.