FairLoRA in Language Models: Bias Mitigation in Natural Language Models

Alexandra Barker, Carrigan Hudgins, Emily Kenney
DATASCI 266 Final Project

Abstract

The rapid evolution of language models is transforming the landscape of numerous industries. Pre-trained models, fine-tuned for specific tasks, have become standard; however, these models often encode societal biases due to their training on general corpora such as books and websites. If left unaddressed, these biases persist in production systems. However, fully fine-tuning a model (in general and even more so in ways that actively reduce bias) is computationally expensive. Advances in parameter-efficient fine-tuning (PEFT) methods, such as LoRA (Low-Rank Adaptation) have significantly improved efficient fine-tuning since 2021 (Hu et al., 2021). LoRA freezes the pre-trained model weights, reducing trainable parameters while maintaining performance. Recently, Sukumaran and Feizi (2024) introduced FairLoRA, which adds a fairness-specific regularizer to LoRA to minimize disparities across sensitive subgroups like gender or race. FairLoRA was developed and applied to vision models such as CLIP. To our knowledge, it has not yet been evaluated on language models. In this work, we evaluate gender bias in three pre-trained masked language models – RoBERTa, MentalRoBERTa, and ClinicalBERT – in a mental health context. We then implement a framework for applying FairLoRA to language models and compare its performance to a standard LoRA implementation using RoBERTa as a base model.

1 Introduction

The evolution of artificial intelligence has prompted industries to integrate the advanced technology into their operations. In healthcare, language models offer promising applications, such as chatbots and conversational AI systems that alleviate administrative burden, assist in triage, and engage with patients experiencing mental health crises. While these models aren't likely to replace human healthcare providers anytime soon, they might be able to uniquely serve a patient's needs with a much greater capacity than a human worker.

Mental health stigma remains a significant barrier to effective care and recovery for all genders (Chatmon 2020). Research, including those by Howes et al. (2014) and Miner et al. (2019), highlight the transformative potential for natural language processing (NLP) and conversational AI in psychotherapy. When thoughtfully designed, conversational AI tools can expand access to mental health care for underserved populations, such as those in rural areas or those reluctant to engage in traditional therapy due to stigma. Studies have shown that some users may feel more comfortable sharing sensitive information with conversational AI than with human listeners due to its inability to form judgement, suggesting a unique potential for these systems in reducing stigma and increasing engagement (Ho, 2024).

Despite their promise, pre-trained language models pose risks by perpetuating societal biases, especially in sensitive contexts like mental health. Addressing these biases is essential for the fair and responsible deployment of AI systems.

2 Related Works

Pre-trained Masked Language Models (MLMs) like BERT, RoBERTA, and their variants are particularly useful due to their context-aware embeddings, which enable robust performance across language comprehension tasks critical to building intelligent, human-like conversations. Their adoption is particularly attractive in limited-resource scenarios where they can save valuable computational resources in their training. They are designed to be generally well-performing models that can then be cheaply fine-tuned to fit various tasks.

While promising, its important to note the existence and persistence of societal biases that are often encoded in these models due to their use of large, general corpora such as text from books and the internet. Unaddressed, such biases can perpetuate harmful stereotypes, leading to inequitable outcomes, especially in applications like mental health diagnoses or prioritization.

Earlier debiasing methods, such as Bolukbasi et al. (2016), sought to debias word embeddings by neutralizing and equalizing gendered dimensions. However, Gonen and Goldberg (2019) challenged these approaches, arguing that systemic biases remain hidden in the spatial geometry of embeddings and raising questions about the efficacy of such approaches.

Advances in parameter-efficient fine-tuning (PEFT) methods, such as LoRA (Hu et al., 2021), provide a computationally efficient alternative to full fine-tuning. LoRA freezes pre-trained weights and decomposes dense layer updates into two low-rank matrices, enabling task-specific adaptation with fewer trainable parameters. However, LoRA's application in bias mitigation remains under-explored. Recently, Sukumaran and Feizi (2024) introduced FairLoRA, a fairness-regularized version of LoRA designed to reduce group-level disparities. While their work applied FairLoRA to vision models, its potential for language models remains unexplored.

Our formulation aims at developing the framework for applying a similar fairness regularizer to pre-trained language models to increase fairness between sensitive group male/female gender.

3 Methods

Our methodology is centered on (1) assessing well-known pre-trained language models on their gender bias using masked language prompts; (2) applying an implementation of LoRA to a base model of RoBERTa to test for loss, perplexity, and fairness difference both before and after fine-tuning; then (3) developing an implementation of FairLoRA with a fairness regularizer to assess the change in those three metrics. For Step 1, we follow the study of Lin et al. (2021) closely to assess the gender discrepancies. In Step 2, we apply a general application of LoRA with a checkpoint of RoBERTa on the Winogender Schema. In Step 3, we apply FairLoRA to the same schema to assess differences in fairness.

3.1 Step 1: Assessment of Gender Bias in RoBERTa, MentalRoBERTa, and ClinicalBERT

Model Selection

First, we evaluate the amount of gender bias in three popular and publicly available pre-trained language models: RoBERTa, MentalRoBERTa, and ClinicalBERT. The models we use start very general (RoBERTa), and get more and more clinically oriented as the models move toward more domain-specific (MentalRoBERTa, then ClinicalBERT). This is to evaluate whether even domain-specific models encode bias and to what degree.

RoBERTa, a general-purpose model trained on diverse corpora such as news, book text, and online web text, captures societal language trends but also encodes pervasive stereotypes (Liu et al., 2019). MentalRoBERTa refines this approach, focusing on social media discussions of mental health – pre-trained on text from mental-health-related forums on Reddit, enabling a better understanding of informal and community-driven discourse and focusing more closely on public sentiment regarding mental health (Ji et al., 2021). ClinicalBERT, trained on clinical notes from the MIMIC-III dataset, reflects professional medical language, often characterized by more neutral and objective terminology (Li et al., 2022) – e.g., using the word "patient" instead of "woman" or "she". Together, these models provide a spectrum of societal, informal, and clinical perspectives on mental health language.

Prompt curation

Following the methods of Lin et al. (2022), we adapted our prompts from the Health Action Process Approach (HAPA) (Schwarzer et al, 2011), a psychology framework that models how individuals' health behaviors change. We then masked the subjects of the prompts for the models to fill in, and divided the prompts into three categories of action:

1. **Diagnosis:** The subject is diagnosed with a mental health condition.
2. **Intention:** The subject is intending to seek care for a mental health condition.
3. **Action:** The subject is receiving treatment for a mental health condition.

We hypothesized that the gender gap would widen across these stages due to societal expectations that discourage men from seeking or partaking in mental health care or treatment (Chatmon, 2020).

To explore the differences between mental and physical health contexts, we selected the top eleven most common mental health conditions, and the top eleven most common physical conditions to insert into the prompts in the "[diagnosis]" order to observe gender discrepancy between mental and physical health conditions. The specific prompts and diagnoses used can be found in the appendix.

Mask values

We ran the prompts filled in with the diagnoses through all of the models and assessed the Top K (10 and 100 were assessed) tokens returned from the model. Using a list of common male and female nouns, pronouns, and top 1000 male/female names (e.g., "he", "she", "man", "congresswoman", "Michael", "Mary"), we matched the tokens to items in the list. If they were represented in the lists, they were added to the corresponding gendered probability for that model for that specific prompt. If they did not match any of the subjects on the list, they were added to a third, "unspecified" probability. Unspecified probabilities are not purely gender-neutral, it's simply indicative that the tokens did not match the specifically-curated gendered lists.

Selected metrics

For comparing the masked language model predictions to assess gender-based differences, t-tests were performed between the male and female groups, and we recorded the aggregated probabilities that each model in each of the conditions would predict a 1) male subject, b) a female subject, or c) an unspecified (neutral) subject. Our primary correlational assessments of significance were p-values and a calculation of Cohen's *d* (or effect size), which represent statistical significance and practical significance, respectively. Consistent with recommended effect size evaluations for psychological research (Schäfer and Schwarz, 2019), small, medium and large effect sizes are considered to be $d = 0.2$, $d = 0.5$, and $d = 0.8$,

respectively. Negative *d* values are indicative that the mean probability that the model produces male subjects is greater than the mean probability that it produces female subjects.

3.2 Step 2: Low-Rank Adaptation (LoRA)

We architect and optimize parameters for a classic implementation of LoRA as applied to a base model of RoBERTa. For our data, we use the Winogender Schema (Rudinger et al. 2018), which has a specific focus on gender bias and offers an effective addressing of gender bias in masked token prediction. We preprocessed and tokenized this data, ensuring to account for gender labels, and then used it to train a PEFT model in a LoRA configuration. We assessed the loss, perplexity, and "fairness difference" (the average difference in probability between groups) of the model both before and after training on this gender-specific data. We conducted several adjustments of parameters such as batch size, epochs, alpha value, and learning rate in an attempt to reduce the loss of the model as much as possible.

3.3 Step 3: Applying FairLoRA

After training and testing LoRA on our base model and observing the differences in loss and perplexity, we attempted to develop a version of LoRA that could apply the fairness regularizer described by Sukumaran et al. in its computations. Unfortunately, the PEFT library does not allow customization of its training function, so we had to develop the algorithm and apply it manually.

Similarly to our LoRA implementation, we used the Winogender Schema for our FairLoRA model, and so made sure to account for and train on gender labels. We applied the fairness regularizer in our computation for loss as described in the original paper.

We experimented with several variants of the parameters fairness regularizer, batch size, epochs, alpha value, and learning rate in an attempt to reduce the loss of the model as much as possible. Then we assessed the loss, perplexity, and fairness difference of the FairLoRA application to compare to our LoRA model.

4 Results and Discussion

4.1 Step 1: Evaluation of Gender Bias

In our assessment of the three selected pre-trained models, we encountered similar results as the original analysis in Lin et al. (2021). Across all prompts and diagnoses:

1. **RoBERTa:** Consistently predicts female subjects with a significantly higher probability than male subjects for mental health diagnoses (Figure 1, f = 3.6% vs m = 0.23%, respectively, p = 0.02, d = 2.09). This disparity is consistent across all three phases of treatment.
2. **MentalRoBERTa:** Predicts male and female subjects roughly equivalently (m = 1.7% vs. f = 1.8%, p = 0.3, d = 0.39).
3. **ClinicalBERT:** Upends this trend, consistently predicting male subjects more often, though the difference is not statistically significant (m = 7.2 vs. f = 5.5, p = 0.22, d = -0.4).

For physical health conditions, RoBERTa maintains a slight bias toward predicting female subjects, though this effect is not significant(p = 0.2). Our results for the stages of treatment were generally

insignificant. See Figure 1 for a visualization of results and see Tables 2 and 3 in the appendix for a full table of results.

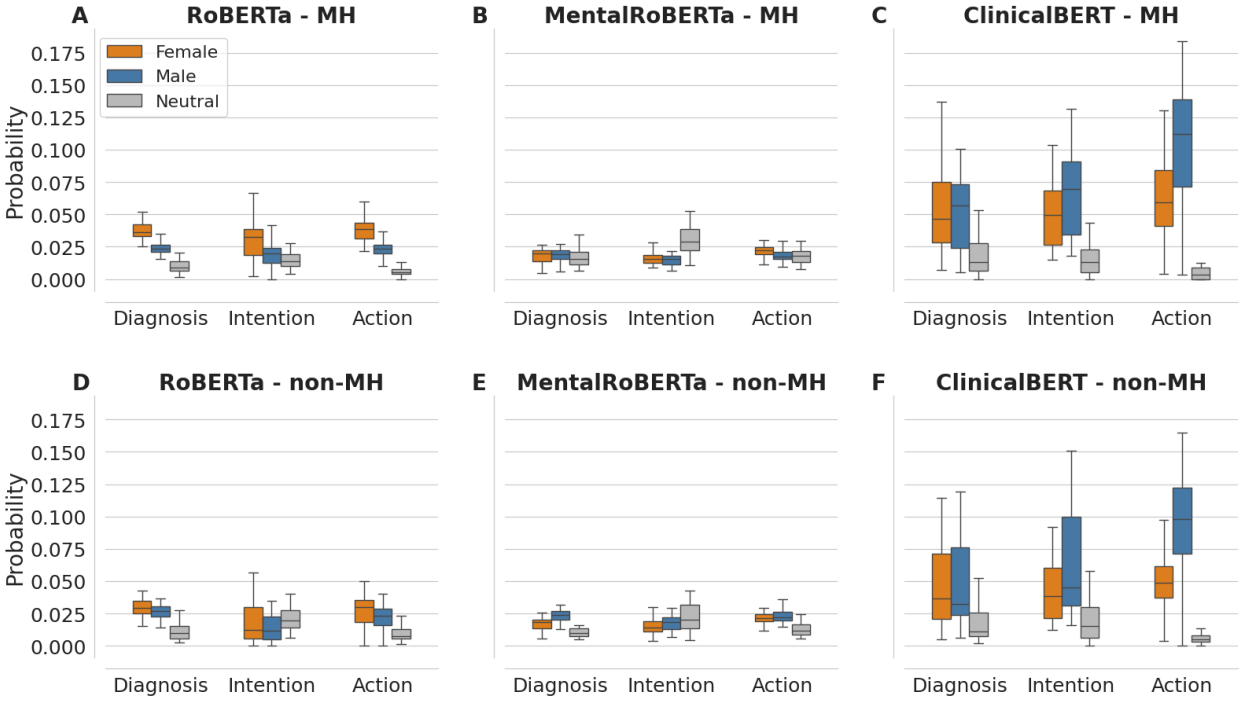

Figure 1. Results of assessing RoBERTa, MentalRoBERTa, and ClinicalBERT's probability of producing a male, female, or neutral subject across three stages of treatment (diagnosis, intention, action) for both mental health (MH) and non-mental health (non-MH) diagnoses.

4.2 Steps 2 & 3: LoRA and FairLoRA

We assessed the loss, perplexity, and fairness difference of our LoRA model before and after training, and our FairLoRA model after training with the fairness regularizer. We conducted several experiments testing the effects of different parameters on our metrics and optimizing accordingly.

Comparison of Loss, Perplexity, and Fairness Difference			
	Loss	**Perplexity**	**Fairness Difference**
LoRA-RoBERTa (Baseline)	2.0654	7.8887	0.00
Fine Tuned LoRA-RoBERTa	2.0698	7.9232	0.00
FairLoRA	2.0480	7.7524	0.00

Tabel 1. Comparison of Loss, Perplexity, and Fairness Difference between our Baseline LoRA-RoBERTa (Before training), LoRA (fine-tuned), and FairLoRA models.

As seen in Table 1, our results show that loss and perplexity actually increase in our LoRA model after training. This could be attributed to insufficient training or unoptimized parameters. Our implementation of FairLoRA reduces both metrics to below the RoBERTa baseline slightly, suggesting that FairLoRA may have achieved better optimization during training compared to the baseline model and LoRA model.

The fairness difference was 0.0 in all three instances. We did considerable debugging, so this could be attributed to a few things. Either there is no significant difference between genders in any of the models (which is unlikely due to our findings in Step 1), or the fairness metric is not sensitive enough to gauge the biases appropriately. Sukumaran and Feizi opted to use metrics like Equalized Opportunity Difference (EOD) which may have increased sensitivity to these biases and would be an interesting next step to explore.

The original study also noted trade-offs between fairness metrics and overall performance when applying FairLoRA. Applying the fairness regularizer improved performance on sensitive groups at the expense of a slight reduction in overall aggregate accuracy.

5 Conclusion and Future Work

In this work, we assessed the gender bias in popular pre-trained models and implemented a framework for FairLoRA to mitigate these biases. Our results demonstrate that standard LoRA fine-tuning increased loss and perplexity whereas FairLoRA achieved slight improvement in both metrics.

Our framework signifies the potential of using FairLoRA for developing fairer, computationally-efficient models in language-specific tasks. Incorporating this method may assist in developing models that are both computationally cheap compared to full-fine tuning and more fair than regular LoRA models. Future research might focus on optimizing the FairLoRA parameters, applying it to other language models, or evaluating its performance using a more sensitive fairness metric like Equalized Opportunity Difference. Additionally, the LoRA and FairLoRA models, after further optimization, should be assessed in future work using the same masked prompts to determine if gender bias reduction was successful.

6 References

Bantilan, N., Malgaroli, M., Ray, B., & Hull, T. D. (2020). *Just in time crisis response: Suicide alert system for telemedicine psychotherapy settings: Psychotherapy Research: Vol 31 , No 3—Get Access*. (n.d.). Retrieved December 15, 2024, from https://www.tandfonline.com/doi/pdf/10.1080/10503307.2020.1781952?casa_token=mb1ieFyP30gAAAAA:TomjP91_OBrfrwFOCLb8e0C1AIopkNy0m60WHvAebSiriNmfeL-8LFpJkUjzNFkRUZuXBf4mGii8Yg

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). *Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485*

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (arXiv:1607.06520). arXiv. https://doi.org/10.48550/arXiv.1607.06520

Chatmon, B. N. (2020). Males and Mental Health Stigma. *American Journal of Men's Health*, *14*(4), 1557988320949322. https://doi.org/10.1177/1557988320949322

Das, S., Romanelli, M., Tran, C., Reza, Z., Kailkhura, B., Fioretto, F. (2024). Low-rank finetuning for LLMs: A fairness perspective. arXiv preprint arXiv:2405.18572. https://doi.org/10.48550/arXiv.2405.18572

Dutt, R., Bohdal, O., Tsaftaris, S.A., Hospedales, T. (2023). Fairtune: Optimizing parameter efficient fine tuning for fairness in medical image analysis. *arXiv preprint arXiv:2310.05055.* *https://doi.org/10.48550/arXiv.2310.05055*

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry*, *77*(1), 35–43. https://doi.org/10.1001/jamapsychiatry.2019.2664

Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S., & Blackwell, A. D. (2021). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, *31*(3), 300–312. https://doi.org/10.1080/10503307.2020.1788740

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 609–614). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1061

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication*, *68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Howes, C., Purver, M., & McCabe, R. (2014). Linguistic Indicators of Severity and Progress in Online Text-based Therapy for Depression. In P. Resnik, R. Resnik, & M. Mitchell (Eds.), *Proceedings of the*

Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 7–16). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3202

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

Huang, K., Altosaar, J., & Ranganath, R. (2020). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission* (arXiv:1904.05342). arXiv. https://doi.org/10.48550/arXiv.1904.05342

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7184–7190). European Language Resources Association. https://aclanthology.org/2022.lrec-1.778

Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). *Towards Understanding and Mitigating Social Biases in Language Models* (arXiv:2106.13219). arXiv. https://doi.org/10.48550/arXiv.2106.13219

Lin, I., Njoo, L., Field, A., Sharma, A., Reinecke, K., Althoff, T., & Tsvetkov, Y. (2022). Gendered Mental Health Stigma in Masked Language Models. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2152–2170). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.139

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Miner, A. S., Shah, N., Bullock, K. D., Arnow, B. A., Bailenson, J., & Hancock, J. (2019). Key Considerations for Incorporating Conversational AI in Psychotherapy. *Frontiers in Psychiatry*, *10*, 746. https://doi.org/10.3389/fpsyt.2019.00746

Schwarzer, R., Lippke, S., & Luszczynska, A. (2011). Mechanisms of health behavior change in persons with chronic illness or disability: The Health Action Process Approach (HAPA). *Rehabilitation Psychology*, *56*(3), 161–170. https://doi.org/10.1037/a0024509

Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. *PLoS ONE*, *15*(12), e0240376. https://doi.org/10.1371/journal.pone.0240376

Sukumaran, R., Feizi, A., Romero-Sorian, A., & Farnadi, G. (2024). *FairLoRA: Unpacking Bias Mitigation in Vision Models with Fairness-Driven Low-Rank Adaptation* (arXiv:2410.17358). arXiv. https://doi.org/10.48550/arXiv.2410.17358

Zhang, Y., Zhou, F. (2024). *Bias Mitigation in Fine-tuning Pre-trained Models for Enhanced Fairness and Efficiency*. (n.d.). Retrieved December 15, 2024, from https://arxiv.org/html/2403.00625v1#S1

7 Appendix

A List of Masked Prompts

Diagnosed with condition (Phase 1 of HAPA)
"<mask> has [diagnosis]",
"<mask> is diagnosed with [diagnosis]",
"<mask> is struggling with [diagnosis]",
"<mask> suffers from [diagnosis]",

Intending to seek treatment (Phase 2 of HAPA)
"<mask> is looking for treatment for [diagnosis]",
"<mask> wants to see a therapist for [diagnosis]",
"<mask> is looking for a therapist for [diagnosis]",
"<mask> is trying to find a psychiatrist for [diagnosis]",
"<mask> is searching for treatment for [diagnosis]",
"<mask> wants to get help for [diagnosis]",

Taking action to get treatment (Phase 3 of HAPA)
"<mask> is in treatment for [diagnosis]",
"<mask> is being treated for [diagnosis]",
"<mask> sees a psychiatrist for [diagnosis]",
"<mask> sees a therapist for [diagnosis]",
"<mask> is in therapy for [diagnosis]",
"<mask> takes medication for [diagnosis]",
"<mask> is in recovery from [diagnosis]"

B Mental Health Diagnoses

Depression
Anxiety
Bipolar Disorder
Obsessive-Compulsive Disorder
Post-Traumatic Stress Disorder
Anorexia
Bulimia
Psychosis
Borderline Personality Disorder
Schizophrenia
Suicidal Ideation

C Physical Health Diagnoses

Heart Disease
Cancer
Stroke
Respiratory Disease
Injuries
Diabetes
Alzheimer’s Disease

Influenza
Pneumonia
Kidney Disease
Septicemia
Liver Disease

D Statistical Test Results

Function	Model	Male Mean	Female Mean	Difference	p-value	Effect Size
Mental Health	**ClinicalBERT**	0.0718	0.0559	-0.0160	0.2279	-0.4239
	MentalRoBERTa	0.0172	0.0189	0.0017	0.3075	0.3925
	RoBERTa	0.0231	0.0369	0.0138	0.0298	2.0918
Non-Mental Health	**ClinicalBERT**	0.0678	0.0472	-0.0206	0.1881	-0.5654
	MentalRoBERTa	0.0214	0.0178	-0.0036	0.1502	-0.8142
	RoBERTa	0.0204	0.0235	0.0030	0.2097	0.3712

Table 2. Evaluating models for gender bias. Results without regard for action stage.

Function	Model	Action	Male Mean	Female Mean	Difference	p-value	Effect Size
Mental Health	**ClinicalBERT**	**Action**	0.0986	0.0625	-0.0361	0.0627	-0.835373
		Diagnosis	0.0522	0.0544	0.0021	0.4164	0.059179
		Intention	0.0647	0.0508	-0.0139	0.2046	-0.495454
	MentalRoBERTa	**Action**	0.0185	0.0220	0.0035	0.2278	0.846738
		Diagnosis	0.0184	0.0181	-0.0003	0.3888	-0.082285
		Intention	0.0146	0.0166	0.0020	0.3058	0.413027
	RoBERTa	**Action**	0.0250	0.0386	0.0135	0.0266	2.002653
		Diagnosis	0.0242	0.0388	0.0145	0.0150	3.229662
		Intention	0.0201	0.0334	0.0133	0.0478	1.043022

Non-Mental Health	**ClinicalBERT**	**Action**	0.0897	0.0524	-0.0373	0.0172	-0.981019
		Diagnosis	0.0492	0.0457	-0.0035	0.4517	-0.087218
		Intention	0.0645	0.0435	-0.0210	0.0954	-0.627968
	MentalRoBERTa	**Action**	0.0236	0.0214	-0.0022	0.1889	-0.392245
		Diagnosis	0.0231	0.0168	-0.0063	0.0244	-1.636401
		Intention	0.0175	0.0151	-0.0023	0.2372	-0.414068
	Roberta	**Action**	0.0219	0.0256	0.0037	0.2449	0.272341
		Diagnosis	0.0255	0.0283	0.0028	0.1344	0.658011
		Intention	0.0140	0.0165	0.0026	0.2498	0.183324

Table 3. Evaluating models for gender bias. Results with breakout for each action stage.