FairLoRA in Language Models

Bias Mitigation in Natural Language Models

Alexandra Barker, Carrigan Hudgins, Emily Kenney

DATASCI 266: Natural Language Processing

Why does fairness matter?

- In healthcare, companies are considering the use of these models to alleviate administrative burden and to allow providers to focus on delivering medical interventions to their patients.
- Up-and-coming applications include chatbots and conversational AI models that can assist in intake and triage, and can converse with patients experiencing mental health crises to diagnose and deliver interventions.
- Pre-trained Masked Language Models (MLMs) are useful due to their context-aware embeddings, and they form the backbone for many language comprehension tasks critical to building intelligent, human-like conversations.
- Debiasing keeps these harmful stereotypes from misdiagnosing, downgrading patient priority, or otherwise framing a medical situation incorrectly

Key Concepts

Pre-Trained Models

Models like RoBERTa and MentalRoBERTa are pre-trained on large datasets, which saves computational time but may encode societal biases.

Low-Rank Adaptation (LoRA)

Proposed by Hu et at. (2021), LoRA freezes most of the pre-trained model's parameters and focuses on injecting low-rank decomposition matrices. This reduces computational cost significantly.

FairLoRA

Recently introduced by Sukumaran and Feizi (October 2024), FairLoRA adds a fairness regularizer to LoRA, aiming to minimize disparities between sensitive groups, such as gender or race.

In this study, we adapt FairLoRA, originally used for vision models, to mitigate gender bias in language models.

Research Objectives

Evaluate Gender Bias

Assess the level of gender bias in pre-trained language models in the context of mental health.

Apply LoRA

Architect and parameterize a classic implementation of LoRA as applied to a base model of RoBERTa.

Implement FairLoRA

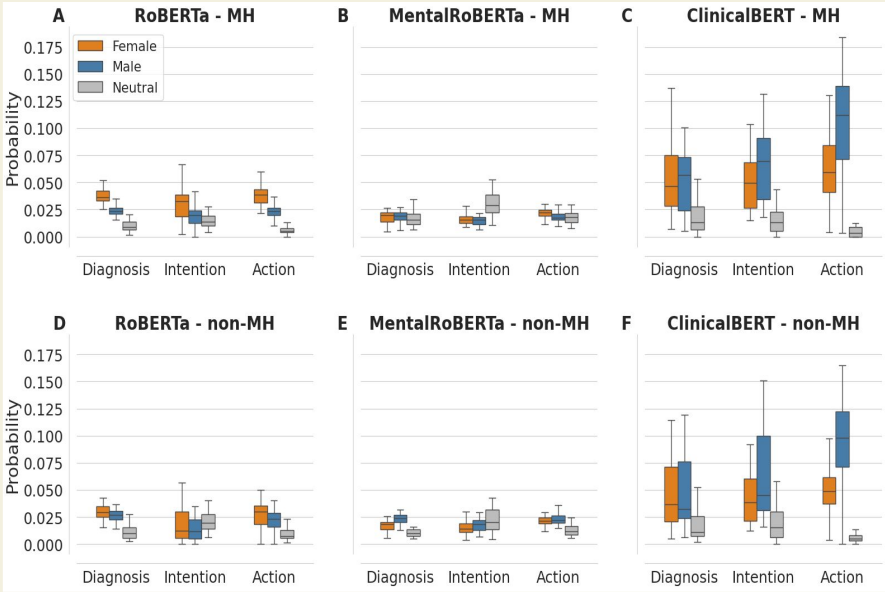
Develop and implement FairLoRA adapted for language models and compare performance using loss, perplexity, and fairness difference.

Assessing Gender Bias in Pre-Trained Models

- Model selection
  - RoBERTa
  - MentalRoBERTa
  - ClinicalBERT
- Prompt curation
  - Prompts adapted from the Health Action Process Approach (HAPA) (Schwarzer et al., 2011).
  - Masked gendered subjects to test each model's probability of predicting male, female, or neutral outputs.
- Stages of mental health action
  - Diagnosed
    `<mask> has [diagnosis]`
  - Intending to seek treatment
    `<mask> is looking for treatment for [diagnosis]`
  - Taking action
    `<mask> is in therapy for [diagnosis]`
- Diagnoses
  - Mental health (MH) diagnoses (e.g., depression, anxiety)
  - Non-mental health (non-MH) diagnoses (e.g., cancer, heart disease)

Results: Assessing Gender Bias

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



LoRA & FairLoRA

LoRA

- Using the Winogender Schema
  - Designed to test gender bias.

Output text: `{'text': 'The <occupation> told the customer that <gender_labels> could pay with cash.', 'occupation':'technician', 'gender_labels': 'he'}`

- Optimized parameters like batch size, learning rate, epochs, etc.
- Tested metrics before and after training

FairLoRA

- Also used Winogender Schema
- Developed framework for FairLoRA adapted for use in language models
- Addition of the fairness regularizer during training.
- Optimized same parameters as the LoRA model, in addition to adding a fairness regularizer.

Metrics

- **Loss** - how well the model fits the data
- **Perplexity** - measure of uncertainty in predictions
- **Fairness Difference** - average gender disparity

Results: LoRA & FairLoRA

Model	Loss	Perplexity	Fairness Difference
LoRA-RoBERTa (Baseline)	2.0654	7.8887	0.0
LoRA-RoBERTa (Fine-tuned)	2.0698	7.9232	0.0
FairLoRA	2.0480	7.7524	0.0

Takeaways & Future Work

- Takeaways
  - Evaluated gender bias in popular pre-trained models for mental health contexts.
  - Applied LoRA, achieving efficient fine-tuning but limited fairness improvements.
  - Implemented FairLoRA, showing promising reductions in key metrics, though fairness results require further evaluation.
- Future Work
  - Optimize parameters for FairLoRA to enhance performance.
  - Test alternative fairness metrics like Equalized Opportunity Difference.
  - Apply FairLoRA to other pre-trained models and sensitive groups, such as race or age.

Thank you!