Gender Biases in Language Model and Debiasing Techniques

Abstract

This article proposes a solution to reduce bias from model embedding and model predictions, consequently promoting social inclusion and gender-based diversity. The Use of AI in sensitive areas has sparked the concept of fairness in AI. Models trained on biased data or data reflecting societal or historical inequalities can lead to further increase of this gap.

This covers both methods of contextual and non-contextual word embedding debiasing techniques, and also aims to compare bias in differing models. This article includes differing examples from BERT to GPT3 on the OpenAI playground, depicting gender-based stereotyped occupations and Shapely values for visualization of bias.

We have covered mitigating techniques by fine-tuning BERT on the WinoBias Dataset and have also developed a novel method for removing bias by using GANs principles, to shift the prediction of female gender prediction towards male gender prediction by discriminator strategy for GANs. In addition, this generated further distribution, which was closer to the prediction of male pronouns, and Wasserstein distances[1] have been decreased from 0.8 to 0.4. This was used further for the generator in Electra, for mitigating bias in word embedding.

1 Introduction

Why is it important to study Societal? Increasing wage gaps, decreasing labor participation by under-represented groups of society are problems which are further intensifying due to pandemic. This is proven by studies conducted by reputed organizations like World Economic Forums who state that "Globally, women earn on average just 68% of what men are paid for the same work, and just 40% on average in countries with the least gender parity"

Examples of biases in machine translation models,Google Translate has been accused of stereotyping based on gender, such as its translations presupposing that all doctors are male and all nurses are female. Figure 1 shows a translation from gender neutral Finnish to English, where leader and journalist is being associated with male where as child care and headache is associated more with female.

GPT-3 playground on OpenAI Example, When we use gender neutral name and use this structure in OpenAI playground, GPT-3 model completes sentences as shown in Figure 2 and Figure 3. In most cases the model predicts and associates Doctor with He/His/Him and Teacher with She/Her.

Sources of Human Biases in Machine Learning, the major source of biases are imbalance of data, under representation of certain groups or total absence of the group.
In the methods proposed, We have demonstrated, how these models exhibit unfairness towards under-privileged groups and issues that it can create if deployed in the real world. The methods are based on non-contextualized, contextualized embedding and prediction. Method for non-contextualized word-embedding: Removing Biases with projection(**?**). For contextualized word-embedding: Fine-tuning on unbiased data set (**?**) and Letter-level embedding like Flair (**?**) . Debiasing based on prediction of Gender/Group, we have Used GAN to shift the prediction probability (**?**)

Evaluations are done on the basis of measuring distance between the male/female pronoun embedding with respect to profession embedding in

[1]Details on wassterstein distances and debiasing are present in appendix

non contextual, whereas it is difference between prediction of male pronoun and female pronoun for Fine-tunned BERT (**?**) and Flair embedding (**?**), In the GAN based shift of predictions, it's based on wassterstein distance moving from 0.8 to 0.4.

2 Related Work

One of the most noteworthy works done on removing biases from embedding is a paper published in NeurIPS 2016 titled Man is to Computer Programmer as Woman is to Homemaker?. It highlights techniques on study of Gender Stereotypes in non-contextual word embedding. It also deals with debiasing with soft bias correction. It provides a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words "receptionist" and "female" via the projection method, while maintaining desired associations such as between the words queen and female.

But drawbacks of this method were brought up in Lipstick on a Pig which is a series of experiments that show this method of projection can only hide the biases but not remove it.

One of the prominent work that represents biases in clinical tasks to measure biases due to marginalized population embedding encoding differently is a research paper titled Quantifying Biases in Clinical Contextual Word

The way to measure stereotypical biases in pretrained language was proposed in StereoSet: Measuring stereotypical bias in pretrained language models StereoSet, a large-scale natural dataset in English to measure stereotypical biases in four domains: gender, profession, race, and religion.