FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

OF HIGHER EDUCATION

ITMO UNIVERSITY

Faculty of Digital Transformation

Education program: Big Data and Machine Learning

Subject area (major): Applied Mathematics and Informatics

REPORT

Towards Understanding and Mitigating Social Biases in Language Models

Student: Korneev A.I., group J42332c

Supervisor: Gladilin P.E.

Date: 09.01.2022

St. Petersburg

2022

Contents

1 Introduction

I chose the paper which is called “Towards Understanding and Mitigating Social Biases in Language Models” from ICML 2021 by Paul Liang, Chiyu Wu, Louis-Philippe Morency and Ruslan Salakhutdinov.

Machine learning is developing rapidly in recent years, permeating various types of human activity. Machine learning tools are beginning to be used for decision making in the judiciary, banking and many other areas. The social impact of these uses has come under scrutiny, with potential discrimination based on protected attributes being a key feature. The need to study this issue, formalize what harmful discrimination is and ways to avoid it, motivate more and more scientific articles on this topic.

Authors of considering paper aim to provide a more formal understanding of social biases in language models (LMs), in particular they focus on representational biases. Representational biases are harmful biases resulting from stereotyping that propagate negative generalizations about particular social groups, as well as differences in system performance for different social groups, text that misrepresents the distribution of different social groups in the population, or language that is denigrating to particular social groups.

The authors have made two main contributions in the paper – disentangling two sources of bias, which were not distinguished in prior works, and proposing the new method called Autoregressive INLP. First one is about different types of bias: fine-grained, which is considered as local biases at a particular time step that reflect undesirable associations with the context (one can determine this type by analysis of probability distribution for new token conditioned on context) and high-level global, which reflects representational biases in multiple phrases. The new method is proposed to mitigate these types of biases in large pretrained LMs during text generation.

2 Model

The first contribution is dedicated to theoretical formalization of biases, which is later used in results estimating. The main conclusion about local bias we can obtain is that it can be evaluated by measuring the difference between biased and debiased LM using KL divergence and the Hellinger distance methods. To evaluate global bias authors used human evaluation and analyzed the difference in sentiment and regard of the resulting sentence using a pretrained classifier. Author have discovered strict dependence between performance of LM and bias mitigating, in order to evaluate performance they have used the same metrics as for the local bias between prediction and ground truth, taking into consideration context association.

Approach for mitigating biases in LMs provided by authors contains two steps. First step is dedicated to identifying sources of local and global biases by finding it through sensitive tokens. Authors propose new learning-based approach that can detect new bias sensitive words to ensure fair generation. Authors identify the bias subspace by starting with several definitional bias pairs such as “he” and “she”, “father” and “mother” for gender, and “jew”, “christian”, “muslim” for religion. They embed each bias-defining word using GloVe and took the SVD of differences between each pair of vectors to obtain a low-dimensional bias subspace. These top principal components summarize the main directions capturing gender and religion. Then they project all possible candidate generation tokens onto our bias subspace and the tokens with high projection values are regarded as bias sensitive tokens.

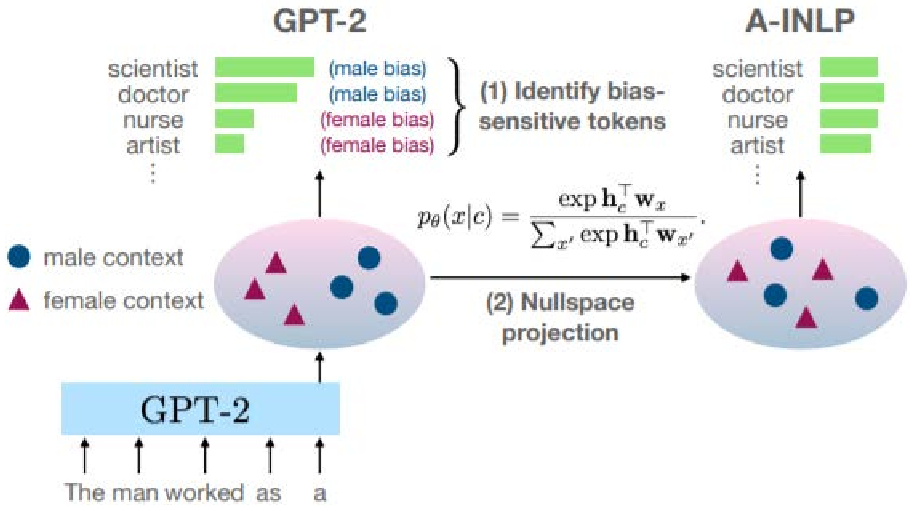

Figure 1 – Approach to mitigate bias

The second step in bias mitigating task is using a special method in order to obtain a more uniform distribution over possibly sensitive tokens. Their method is based on iterative nullspace projection method (INLP). The main idea is to find a linear guarding function that removes linear dependence between word embeddings set and set of corresponding protected attributes. To achieve this, INLP used a linear classifier to predict a protected attribute values from word embeddings and then project word embedding onto the nullspace of classifier parameter, which serves the purpose of removing all information used by classifier to predict protected attribute. The guarding function can be represented as a matrix which allows us to obtain linear transformation and as a result to obtain new embedding without dependence on protected attribute.

Autoregressive INLP (A-INLP) extends IMLP for autoregressive text generation process. At this step bias-sensitive tokens and nullspace matrix should be obtained. At every time step during text generation process, authors apply INLP to the current context embedding to ensure that generation of next tokens is invariant to gender in the context. They have also added hyperparameter to be able to adjust balance between biased and debiased LMs. The idea of building debiased GPT-2 model is presented at the Figure 1.

Despite the advantages of this approach such as decreasing the bias level in text generation models in comparison with prior works, the ability to use pretrained LM without any changes in its structure and as a result to provide a way to debias a large LM for users with low computational recourses, there are several disadvantages: firstly, the performance of model is decreasing, secondly, it takes additional time and recourses to build a classifier while searching for matrix P, lastly, this approach depends on the quality of determination of bias-sensitive tokens subspace.

3 Experimental results

For my experiments I have decided to use an approach from original paper to debias LM with respect to social groups of different ages. For simplicity I have considered only two categories: “young” and “old”. I have conducted the full experiment for GPT-2 model, however, I have also separately looked for bias-sensitive tokens and created a classifier for both GPT-2 and GloVe word embeddings. As a text corpus I have used Wikipedia content, reddit content and texts from news web-sources.

I have used different initial pairs of bias-sensitive words, because not all of the words had predefined embeddings, for GloVe I have used: ['young', 'old'], ['youthful', 'elderly'], ['adolescent', 'senescent'], ['teenage', 'aged'], ['child', 'adult'], for GPT-2 I have used ['young', 'old'], ['youthful', 'elderly'], ['children', 'adults'], ['teenage', 'aged'], ['child', 'adult']. By applying the proposed by authors approach for GloVe emebeddings I have obtained 47 biased tokens for first group (old) and 46 tokens for second one (young), results are presented in the Table 1.

First social group (threshold 0.22)	Second social group (threshold 0.32)
'metalware', 'refectory', 'firebrick', 'lysosomes', 'senescent', 'salvaged', 'lyse', 'ironsides', 'verdigris', 'terrazzo', 'depreciated', 'renumbering', 'gravestones', 'mahabalipuram', 'switchvox', 'Aged', 'infirm', 'refinish', 'unrestored', 'rotted', 'flagstones', 'obsolescent', 'Adult', 'souq', 'overprint', 'dsdt', 'spindles', 'adult', 'thomann', 'decommissioned', 'aged', 'Elderly', 'lysed', 'tombstones', 'unserviceable', 'elderly', 'patina', 'crotchety', 'creaky', 'retrive', 'old', 'Old', 'Senescent', 'glomeruli', 'friable', 'leaved', 'scavenged', 'pbxs'	'fantasies', 'rape', 'adolescence', 'youthful', 'mentoring', 'angst', 'bullying', 'Young', 'teenaged', 'bisexual', 'struggles', 'violent', 'feelings', 'Child', 'addiction', 'violence', 'emotional', 'sexuality', 'psychotherapy', 'motivation', 'Youthful', 'adolescents', 'lesbian', 'psychology', 'young', 'child', 'parenting', 'teenagers', 'Teenage', 'adolescent', 'kid', 'girls', 'teen', 'emotions', 'innocent', 'underage', 'teenager', 'girl', 'sex', 'youth', 'childhood', 'erotic', 'teenage', 'queer', 'sexual', 'Adolescent', 'teens'

Table 1 – Biased tokens for GloVe embeddings

Following the original code I have filtered neutral words for both social groups. Finally, I have obtained two lists of biased tokens that have been found in the text corpus (Table 2). As a result, 1714 sentences contained tokens for the first social group and 2187 sentences contained tokens for the second social group.

First social group	Second social group
'decommissioned', 'salvaged', 'aged', 'leaved', 'adult', 'old', 'obsolescent', 'depreciated', 'Adult', 'Old', 'creaky', 'rotted', 'elderly', 'crotchety', 'Aged'	'mentoring', 'childhood', 'youthful', 'teenaged', 'kid', 'teen', 'adolescent', 'teenager', 'adolescence', 'girl', 'child', 'youth', 'Teenage', 'teens', 'underage', 'Young', 'young', 'Child', 'adolescents', 'girls', 'innocent', 'parenting', 'addiction', 'teenagers', 'teenage'

Table 2 – Biased tokens for GloVe embeddings which were found in text corpus

For GPT-2 emebeddings I have obtained 15 biased tokens for first group and 54 tokens for second one, results are presented in the Table 3.

First social group (threshold 0.15)	Second social group (threshold 0.15)
'Adults', 'age', 'old', 'Old', 'expired', 'Old', 'aging', 'elderly', 'Adult', 'older', 'adult', 'aged', 'adult', 'Aging', 'Elder', 'adults'	{'boy', 'Children', 'Girl', 'Kids', 'girls', 'kids', 'young', 'pupil', 'youngster', 'sons', 'Boys', 'sexual', 'parenting', 'daughters', 'baby', 'son', 'teenager', 'Children', 'daughter', 'youths', 'student', 'Boy', 'Girls', 'students', 'fiery', 'toddler', 'teenagers', 'child', 'Girl', 'Kids', 'Child', 'teens', 'kid', 'kidnapping', 'teenage', 'Youth', 'Students', 'father', 'youthful', 'femin', 'youth', 'school', 'boys', 'talent', 'girl', 'rebellious', 'Boy', 'children', 'child', 'Child', 'childhood', 'youngsters', 'Teen', 'Girls', 'teen'}

Table 3 – Biased tokens for GPT-2 embeddings

The results for GPT-2 embeddings seem to be better from the human perception. Finally for GPT-2 after filtering neutral tokens, tokens from the Table 4 have appeared in 1953 sentences for the first social group and 4770 for the second social group.

First social group	Second social group
'old', 'aged', 'Old', 'Aging', 'Elder', 'expired', 'Adult', 'adults', 'aging', 'elderly', 'adult', 'Adults', 'age', 'older'	'kids', 'teen', 'Kids', 'daughter', 'daughters', 'students', 'boy', 'son', 'girls', 'Children', 'girl', 'Girls', 'teens', 'young', 'Teen', 'youth', 'boys', 'youthful', 'sons', 'children', 'teenager', 'pupil', 'childhood', 'youngster', 'Girl', 'kid', 'baby', 'Students', 'toddler', 'Youth', 'school', 'Boys', 'student', 'teenage', 'Boy', 'child', 'Child', 'youths', 'parenting', 'teenagers', 'youngsters'

Table 4 – Biased tokens for GPT-2 embeddings which were found in text corpus

As in the original paper I have used 80 iterations to build a classfier. For GloVe embeddings I have obtained the accuracy 39%, which is higher than results for gender in the original paper. Thus, it means that classifier might contain information to distinguish classes. For GPT-2 embeddings I have obtained accuracy 34%, which is almost the same as the results for gender in the original paper (33%), here we can make the same conclusions: after the nullspace projection, the context embedding cannot be classified with respect to the bias attributes and thus does not contain distinguishable bias information.

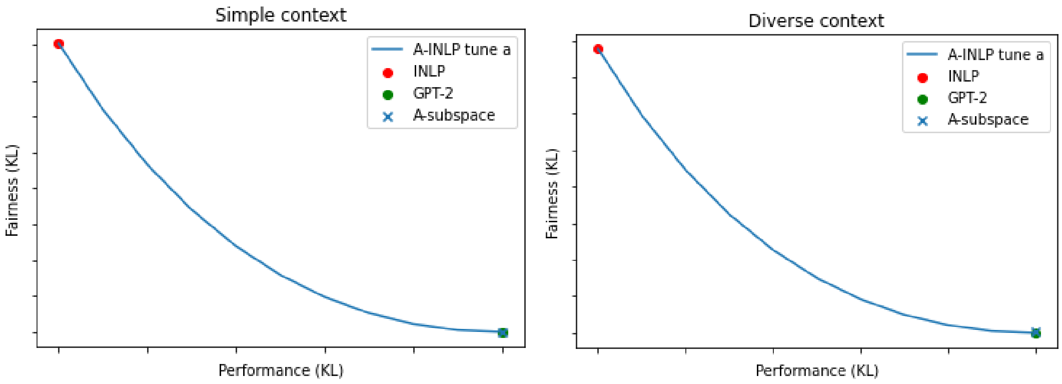

Figure 2 – Bias metrics on age both simple and diverse contexts

Local Bias measuring is provided for GPT-2 model. In the original paper simple context was created by starting the sentence with “The man” or “The woman” and concatenating one of those prefixes with phrases such as “was known for”, “was described as” and etc., then the LM finished the phrase. In my experiments I have saved the same contexts, but changed prefixes to “The young person” and “The old person”. For diverse context I have used the same data as it was in the original paper. In Figure 2, we can see trade-off plots of performance vs fairness as measure across local metrics with both simple and diverse age context. As in the original paper GPT-2 exhibits the best performance while being the most unfair with respect to different social groups. For a=1 INLP baseline is recovered and achieves the best fairness results. For our case we can also see that performance is significantly decreased when the method achieves better fairness value. Curves for both age contexts are steeper than curves for gender contexts from the original paper, which means that it is more difficult to achieve better fairness for age context without losing LM performance. It can be possibly explained by the fact that all diversity of ages have been divided only into two groups (“young” and “old”), however considering age as continuous variable can lead to better results, but the approach of authors give us an opportunity to work only with categorical variables as a protected attribute.

4 Conclusion

The authors assay fairness in machine learning within the context of text generation process. Their approach has its own limitations, but also creates an incentive to scrutinize the data closely. Authors proposed new method to mitigate social bias and disentangle different sources of bias. Moreover, their approach contains a new way to find bias-sensitive tokens, which can be used separately for other methods for bias mitigating. Limitations of the approach have been discussed in the “Model” chapter. Results of my experiments for social biases within the age context coincide with results as in the original paper for gender contexts. To sum up, the original paper opens a diversity of new directions at the technical, empirical and conceptual level and provides a background for future investigations.