A Comparative Study of Gender Bias Associated with Professions in Benchmark Language Models

Ananya Malik
amalik88@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Raj Kothari
rkothari@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Shlok Shah
shlokshah@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

Gender equality is aimed at promoting an equal representation for all genders, especially in the workplace.Language Models have gained much traction in the last few years in the NLP community. These models are used in a host of applications in the real world such as text-generation, summarization, entity matching etc.Our proposed research aims at identifying and comparing the gender bias that exists in these models, especially focusing on bias with respect to different professions. We aim to conclude a qualitative search on which model propagates a greater bias and hope to lead a dialogue into debiasing these models.

Keywords: language models, neural networks, gender bias, natural language processing

ACM Reference Format:
Ananya Malik, Raj Kothari, and Shlok Shah. 2021. A Comparative Study of Gender Bias Associated with Professions in Benchmark Language Models.

1 Introduction

Prior work has shown that machine learning systems can inadvertently capture human stereotypes, including gender bias. This can lead to amplified gender-bias problems in models used for hiring, campus admissions and many such applications. Further, studies have been focused on finding gender bias in one model at a time such as BERT or GPT-2. However, through this project, we plan to compare gender bias, associated with professions, present in various benchmark language models used in Natural Language Processing (NLP). Our research focuses on models such as BERT, GPT-2, XLNet, RoBERTa and T5. We plan to conduct comparative research on gender bias associated with professions for two class of NLP models: Masked and Unmasked. For a list of professions, we will analyze the output of these models as to whether it is biased towards a gender. Further, a model which could classify gender from given text into three categories: male, female and neutral would be used to identify the gender from the text generated by the unmasked model. Finally, we will conduct a study amongst all models and compare the results with ground-truth data manually labeled by us.

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

2 Current Approaches

In this section, we explore the various aspects and domains that can be utilised to determine the gender bias in language models with respect to the professions. We classify the researched papers into the following categories:

2.1 Word Embeddings

Several research has been conducted for finding biases in various word embeddings. One of the main reasons for this extensive research is that most of the NLP models use these word embeddings in their first layers and biases in these word embeddings have an amplified effect on the downstream model. Hence, it becomes necessary to have sufficient knowledge about biases present in word embeddings.[2] studied biases present in w2vNEWS (Word-to-Vec embeddings for google news) and GLoVE embeddings. They asked crowd workers to label occupations and divide them into three categories: male-stereotypical, female-stereotypical, and neutral. Each occupation is labeled by 10 workers and hence rated on a scale of 1-10. The authors then project these embeddings into the vector space and they found that the stereotypes found in word embeddings were same as those annotated by crowd workers.

Further, they also analyzed various analogies such as man - woman = programmer − x. Surprisingly, it was found that the best answer for x was “homemaker”. This experiment bolstered the presence of gender bias in word embeddings. [3] researched on biases present in Glove embeddings by performing Implicit Association Test (IAT) on human subjects and comparing the results with a probability-based method called Word Embeddings Association Test (WEAT). The IAT follows a reaction time paradigm, which means subjects are encouraged to work as quickly as possible. There is an enormous difference in response times when subjects are asked to pair two concepts that they find similar, in contrast to two concepts that they find different. The authors found that

there indeed is a gender bias in both – the way human perceive relations (IAT test) and the word embeddings (WEAT methodology) and that they are similar. This lead to the conclusion that if NLP models learn our language well, they will also learn the culture and injustice/bias present in our language too.

The authors conducted research on various embeddings in [5], namely, Skip-gram embeddings trained on Google News, Twitter micro-posts, PubMed central open access subset and FastText embeddings trained on GAP corpus. Experiments were performed for 5 categories: career vs family, maths vs arts, science vs arts, intelligence vs appearance and strength vs weakness.

For gender bias in career vs family category, both google news and twitter embeddings showed extensive bias. PubMed embeddings demonstrated the least amount of bias which might be due to scientific nature of the set. GAP embedding obtained from Wikipedia was expected to show high amount of bias since Wikipedia is developed by collaboration of huge number of users. However, limited bias was observed. This relatively low bias measurement could be due in part to the fact that GAP's vocabulary lacks many of the attribute and target word lists used in the tests.

2.2 Survey of profession gender bias across models

Vig et al. [9], uses causal mediation analysis to detect gender bias by analysing which of the given model components contribute to the gender bias. The paper uses the defined ground truth to identify a stereotypical candidate and an anti-stereotypical for each of the given profession instances. The paper calculates the fraction of the given instance or profession giving a probability of a stereotypical candidate over the probability of the anti-stereotypical candidate given the instance or profession. The paper then investigates the gender bias in the given model by evaluating different language models for the total effect, which is calculated by assessing the change in the fraction calculated above. Thus, using causality, the paper can obtain which model component contributes to the gender bias and can set a benchmark for evaluating the gender bias in a given model.

Qian et al [7], focuses on investigating gender bias in language models and suggests methods to mitigate this existing bias. The paper proposes a new loss function, which via the introduction of a new loss term aims to equalize the predicted probabilities of gender pairs like he/she, man/woman. The paper evaluates the bias in the language models in terms of the causal bias, co-occurrence bias and word embedding bias.

2.3 Models to identify the gender of the subject

The Natural Language Tool Kit (NLTK) is an open-source project enabling functionalities in the Natural Language Processing Domain on Python. Leveraging the fact that patterns can be observed in the names of males and females, NLTK provides us with an inbuilt supervised classifier and list of names of males and females that can be used to train the classifier. However, this classifier can be prone to underfitting due to the lack of large and diverse amount of data as well as prone to overfitting since the feature-set chosen was poor and weak. Hence, using NLTK to classify names as male or female might give us a base to start our work on but cannot be independently used to receive accurate results.

This survey [6] of existing gender bias in existing machine learning model leads to the authors creating a taxonomy of the structural and contextual gender biases that can manifest themselves in the models. They identify structural bias can be recognized from grammatical construction which includes looking for syntactic patterns or keywords which induce a bias in a gender-neutral sentences such as looking for gender-exclusive pronouns (he, his, she, her, herself and so on) and explicitly marked indicators of sex (policemen, seamstress and so on). Contextual bias does not follow any rules and needs understanding of the gender-bias word in each context which can be seen through social stereotypes such as "senators needing their wives to support them" which assumes the gender of a senator to be male and the gender of the supporter to be female and through behavioral stereotypes such a "All boys are aggressive" which maps traits to gender. In the works of the paper, the authors design a filter using AllenNLP and NLTK to identify biased sentences through their syntax, as exampled above.

Since men and women are talked differently through headlines and the frequency of men being subjects of news headlines is much more than that of women, news headline are a great source of highly bias text content. Using a Convolutional Neural Network (CNN) built with the GloVe word embeddings, this paper [4] analyses headlines of news articles to identify whether that headline talks about a man, woman or neither/both. They achieve an accuracy of 86.7 % through this proposed architecture. As baselines, a Naïve Bayes and a Support Vector Machine (SVM) classifier are defined. For each headline, the F-measure is calculated which explores a text's relative contextually and is used to distinguish between the tones of male and female writing. Through experimentation they find that the CNN with Factor Analysis and GloVe word embedding gives the best accuracy as compared to other variations and the baseline models.

Through an investigate approach, the performance of LSTM and CNN architecture for the task of gender classification and the effect of varying word embeddings on them is observed. Through the paper [8], the authors employ these models on the names of entities in order to classify them as male or female. Since, this becomes a specific niche of the issue of gender classification, we mark this paper for further review under the scope of future work to expand the range of gender classification models.

3 Proposed Plan

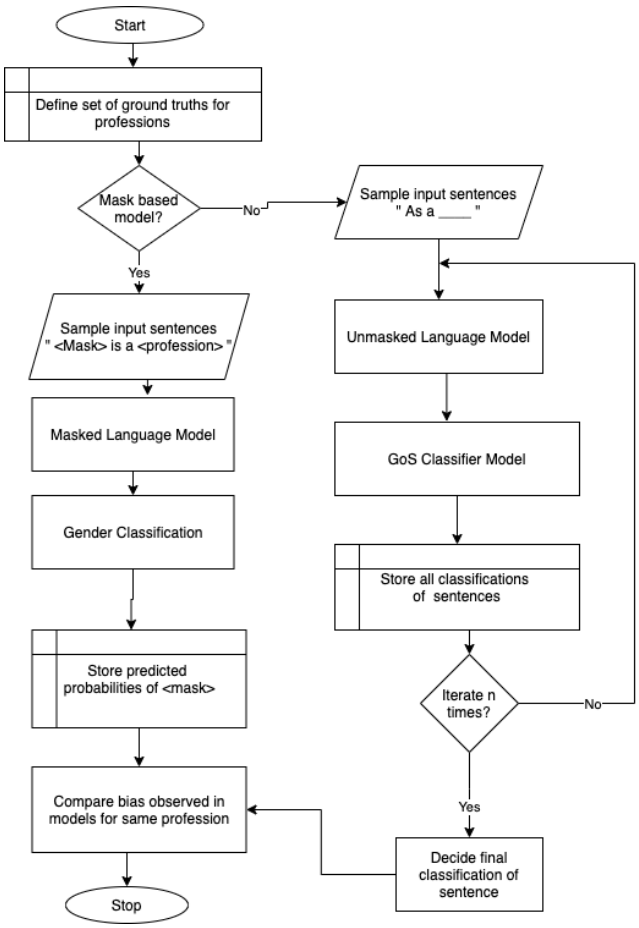

Figure 1. Flowchart depicting the proposed methodology

To compare the language models for the gender bias existing in them, we define a ground truth which is annotated to classify a given profession into a particular class based on a dataset of 100 professions. We divide the flow of this project depending upon the type of language model being used, such as either Masked or Unmasked. This is because the input and output for both of these categories are different. For the masked language models, the input to the model would be in the form of a single sentence focusing on the Mask label as the subject. Thus, the model of the masked type will return the probability of a particular word fitting the Masked Label. We then threshold these probabilities to classify the output for the given model belonging to the following categories: Male, Female or Neutral. We recognize that gender is non-binary, however, for this study, the paper focuses on gender being categorized into the above classes.

For unmasked language models, the input is in the form of an incomplete sentence or a seed phrase. The models generally results in generation of a sentence with the given input sentence as the starting point. Unlike masked language models, these models don't return the probability of the sentence belonging to a particular word, or in the given use case a gender class. Thus, the project aims to solve this by sampling the outputs of the given unmasked models multiple times for a fixed input by running a novel classifier called the 'GoS Classifier'. The GoS classifier then predicts the probability of the generated sentence into one of the chosen classes. The overall probability of a given class would be calculated by aggregating the probabilities obtained from the model in each iteration of the given sample. We then threshold these outputs to classify the model to the given categories: Male, Female or Neutral, which is then compared to the ground-truth to analyze the gender bias.

Thus, the paper achieves a standardized technique to compare language models of different types and also compare each model with the ground truth to determine the bias present.

4 Evaluation Plan

To evaluate and compare the language models that we have considered, namely: BERT, GPT-2, XLNet, RoBERTa and T5, we define a ground truth that is manually annotated to classify each profession amongst 100 professions.

4.1 Metrics

We calculate the probability for each of the profession as derived from each model, which is then classified based on a threshold value, into the classes defined above. We compare these classes with the defined ground truth. We are then able to define the bias as the fraction of unmatch with the ground truth to the total number of instances.

$$\text{bias} = \frac{\text{number of unmatches with ground truth}}{\text{total number of instances}} \quad (1)$$

A greater bias score will denote a greater bias in the model which then can be compared across all considered models

4.2 Dataset

Our dataset comprises of a list of 100 professions and is inspired from the dataset used in [1].We then manually annotate the professions present in this dataset into 3 classes: Male, Female and Neutral.This dataset is defined as our ground truth and is then used to find gender bias present in each model.

References

[1] Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. *CoRR* abs/2010.14534 (2020). arXiv:2010.14534 https://arxiv.org/abs/2010.14534

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *CoRR* abs/1607.06520 (2016). arXiv:1607.06520 http://arxiv.org/abs/1607.06520

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (14 April 2017), 183–186. https://doi.org/10.1126/science.aal4230

[4] Stephanie Campa, Maggie Davis, and Daniela Gonzalez. [n. d.]. Deep amp; Machine Learning Approaches to ... - web.stanford.edu. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15787612.pdf

[5] Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 25–32. https://doi.org/10.18653/v1/W19-3804

[6] Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 8–17. https://doi.org/10.18653/v1/W19-3802

[7] Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. *CoRR* abs/1905.12801 (2019). arXiv:1905.12801 http://arxiv.org/abs/1905.12801

[8] Ritesh and Chakravarthy Bhagvati. 2018. Word Representations For Gender Classification Using Deep Learning. *Procedia Computer Science* 132 (2018), 614–622. https://doi.org/10.1016/j.procs.2018.05.015 International Conference on Computational Intelligence and Data Science.

[9] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12388–12401. https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf