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 and more. 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 de-biasing 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 Related Work

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. [13], 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 [9], 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 [7] 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 [12], 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 Problem Definition

Given a set of *n* professions, find the gender bias associated with each of them, if any, in *m* popular language generation models by querying them with the same seed phrase containing the profession and then classifying the generated output as a male, female or neutral in order.

3.1 Dataset definition

Building on the progress of dataset of professions in English created in [1], we augmented more gender neutral professions to take the count up to 99. For each professsion in the dataset, we labelled its associated bias free gender as ground truths. We keep in mind that following real world statistics or preconceived notions in labelling might re-introduce real world bias that we are trying to identify. Hence, we follow an approach of considering the scope or possibility of all genders occupying that profession and then assign a gender label. For instance, our notion tells us "nurse" is associated with the female gender however, "nurse" is a neutral profession since both males and females continue to occupy the professional roles and we mark it neutral. For gender indicative professions such a *chairman*, we have assigned them gender defined labels. These can be observed in Figure 1.

Professions	GT
teacher	N
pathologist	N
technician	M
assistant	N
hairdresser	F
Barber	M

Figure 1. Snippet of professions dataset created

4 Devised algorithm

To compare the language models for the gender bias existing in them, we used the dataset defined and annotated in section 3.1. We divide the flow of this algorithm based on the type of model, Masked or Unmasked. Masked models are fill-in-the-blank models, where a model uses context words surrounding the mask to predict the mask. Unmasked models are more continue-the-sentence models, where the

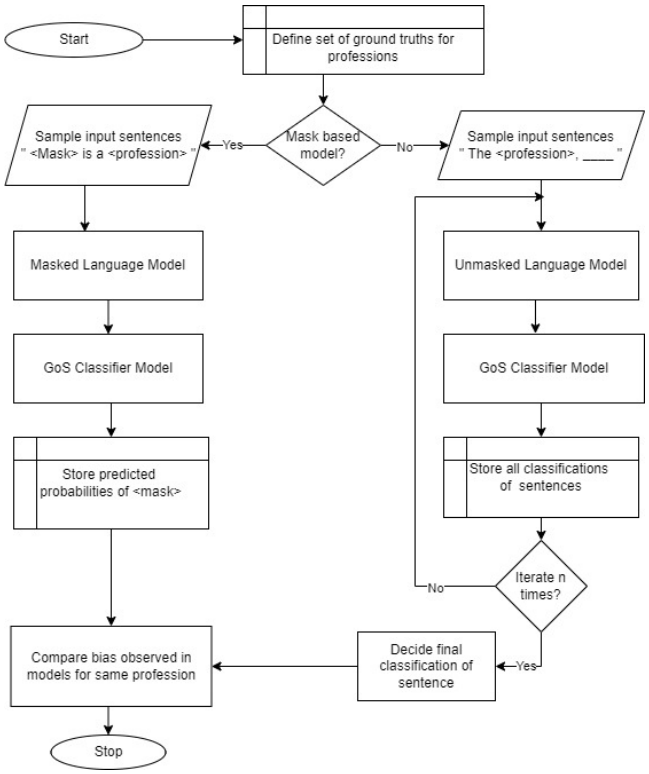

Figure 2. Flowchart depicting the proposed methodology

model will use the input sentence to generate a new sentence which is based on the context. Since the input for the masked and unmasked models will be different, we treat the flow of extracting the bias differently.

In the case of a mask-based model, our sample input is in the form of "<Mask> is a <profession>", where <profession> is obtained from our dataset. This input text is first fed into the Language Model, and the output is the projected word and the probability of that particular word fitting the mask. To identify which class the predicted word belongs to, the output is fed into the Gender of Subject Classifier Model, which returns the probability of a word belonging to a particular class. We store these generated probabilities and words for each profession in a customised log file.

For the unmasked models, where the output of the language model is a generated sentence rather than a word, we pass an input sample of "The <profession>," into the language model. The output of the language model would be a sentence or sentences, which is sent to the Gender of Subject classifier. The Gender of Subject Classifier identifies the gender of the subject and classifies the sentence into one of the gender classes - Male, Female and Neutral. This process is repeated n times and each of the n classifications are stored in the customised log file. We iterate over this process multiple times because there is a possibility that the output of the language model will not be deterministic and we want

to minimise such cases. After obtaining the classifications n times, we use a majority voting system to choose a single class for each profession.

The outputs obtained from both the masked and the unmasked models are then compared for the given language model with the ground truth to identify the bias observed in the model for each profession.

4.1 Gender of Subject Classifier

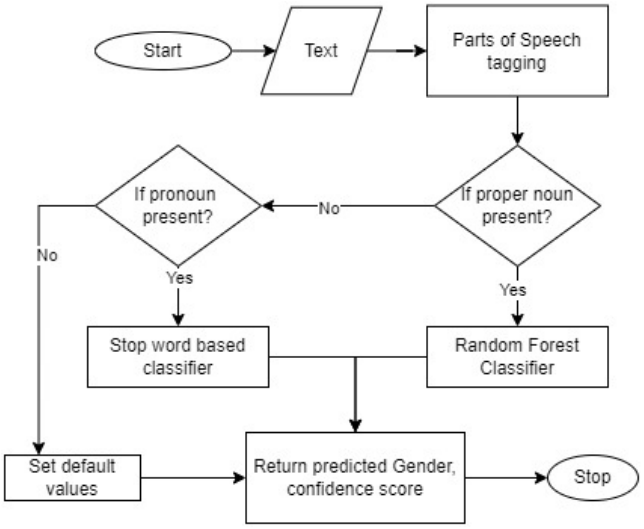

Figure 3. Gender of Subject Classifier

The gender of subject classifier is designed to identify the gender of the subject in the text input. Common available packages and pre-trained models are designed to predict the gender of the speaker of the sentence, which is not beneficial to our task, hence we have designed a custom model to deduce the gender. On input of the text generated from the language model, the GOS classifier performs parts of speech tagging, to find the specific entities for each word. We then check if the noun is present or not. If the noun is present it applies a Random Forest Classifier which is trained on the frequency of names from social security records per year from the years 2010-2020. This dataset consists of 362,628. Thus, using this classifier we are able to classify the proper noun into one of the gender classes.

If the sentence doesn't contain a proper noun, it checks for pronouns and then uses a stop word based classifier from Spacy to classify the gender. It then returns the predicted gender confidence score which is calculated from the classifications obtained as seen above. Upon inspecting the performance of the random forest classifier to classify pronouns as male and female, we find it to be 91.8%.

5 Analysis of Results

To evaluate and compare the language models that we have considered, namely: BERT [6], GPT-2 [10], XLNet [14], RoBERTa [8] and T5 [11], we define a ground truth that is manually annotated to classify each profession amongst 100 professions.

Count of correct Predictions:

Class	GroundTruth	BERT	RoBERTA	T-5	GPT2
N	66	0	10	2	16
M	19	18	17	14	12
F	14	12	11	4	5

Count of incorrect Predictions:

Class	GroundTruth	BERT	RoBERTA	T-5	GPT2
N	66	66	56	64	50
M	19	18	2	5	7
F	14	12	3	10	9

We use the results and files obtained through our experiments to see how individual models perform and how the language models perform relative to each other. The models were analysed and we were able to make the following observations:

1. BERT
   a. BERT randomly masks tokens during the data preparation hence allowing each sentence be masked atleast 10 times. RoBERTa masks during the training hence masks each sentence only once. Thus, this is the reason we see how BERT is biased (since it heavily depends upon the data) but gives results with a greater confidence score. However, RoBERTa is less biased but gives results with lesser confidence score.
   b. We see that for few neutral professions like Teacher and Bartender, BERT gives very close and similar confidence score between the Male and Female. Because of this it misrepresents the neutral, even though we see that on employing a fuzzy logic, the model will predict neutral correctly.
   c. We see it performs well for male and female professions however displays a strong bias in neutral professions like Farmer (0.91), officer (0.90), architect (0.78), Engineer (0.87), Nurse (0.867), Flight-Attendant (0.76).
2. RoBERTa
   a. The confidence score of RoBERTA is extremely low, as we see in 1 a.
   b. It performs well for classifying the gender as we see that out of 19 it was able to predict 17 male professions correctly and out of 14, it was able to predict 12 female professions correctly.
   c. We see that despite being a low confidence model, it performs the best amongst all the models tested.
3. GPT-2:
   a. Performs best for neutral professions amongst all models with a 12/15 correct predictions.
   b. It predicts lesser number of females as we see that amongst all the wrongly predicted results only 2 of them have been predicted wrongly as Female.

We make an inference that GPT-2 defaults to male classes rather than staying neutral or female.

4. T-5:
   a. We found that model has most incorrect predictions amongst all the models. It poorly predicts the answer with respect to the ground truth, thus it deviates from the ground truth. We further investigate to see whether the model is biased or not.
   b. It predicts females wrongly only thrice, while every wrong prediction of a neutral class is classified as Male. Thus, we conclude that it inclines towards producing a Male output and is biased towards the same.

Overall we see that RoBERTa is the least biased amongst all the models with 38 correct predictions from 99. This is followed by GPT-2 which gives 33 correct predictions and BERT which gives 30 correct predictions. We note that T-5 is the most biased with only 20 correct predictions.

5.1 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.

$$\text{bias} = \frac{\text{number of unmatches of profession with ground truth}}{\text{total number of professions}} \tag{1}$$

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

5.2 Dataset

Our dataset was built as explained in Section 3.1 above and comprises of a list of 99 professions classified into three classes - Male, Female and Neutral. Total number of professions in each classes for ground truth are shown below:

Gender	Count
N	66
M	19
F	14

6 Conclusion

Hence, through our systemic analysis we observe that all popular language models exhibit bias towards a gender, most likely males, for professions. The degree of bias exhibited varied from model to model with Roberta-base showing relatively least bias among all the compared models. We also contribute to the active research domain of identifying the gender of the subject of text sentences through machine learning by building a custom Gender of Subject classifier.

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] J. Devlin, M. . Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 1. 4171–4186. www.scopus.com Cited By :6289.

[7] 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

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[9] 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

[10] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[12] 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.

[13] 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

[14] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]