Evaluating Gender Bias in Pre-trained Indic Language Models

Neeraja Kirtane* and **V Manushree***
RBC-DSAI, Indian Institute of Technology, Madras and Manipal Institute of Technology, Manipal
{kirtane.neeraja, manushree635}@gmail.com

Aditya Kane*
Pune Institute of Computer Technology, Pune
adityakane1@gmail.com

Abstract

The gender bias present in the data on which language models are pre-trained gets reflected in the systems that use these models. The model's intrinsic gender bias shows an outdated and unequal view of women in our culture and encourages discrimination. Therefore, in order to establish more equitable systems and increase fairness, it is crucial to identify and mitigate the bias existing in these models. While there is a significant amount of work in this area in English, there is a dearth of research being done in other gendered and low resources languages, particularly the Indian languages. English is a non-gendered language, where it has genderless nouns. The methodologies for bias detection in English cannot be directly deployed in other gendered languages, where the syntax and semantics vary. In our paper, we measure gender bias associated with occupations in Hindi language models. Our major contributions in this paper are the construction of a novel corpus to evaluate occupational gender bias in Hindi and quantify this existing bias in these systems using a well-defined metric.

1 Introduction

Transformer-based language models like BERT (Devlin et al., 2018) have now become the new benchmark for all the tasks in NLProc, like machine translation, text classification, summarization, etc. While these models are very effective in capturing and understanding the given information and task, they have an inherent bias present in them. Wang et al. (2021) shows how there is a discrimination with respect to gender in job recommendation systems. This bias treats some people differently than others. Therefore it is essential to address bias in these models before using them for specific tasks to ensure that it is fairer for everyone.

Bolukbasi et al. (2016) was the first paper to quantify and mitigate bias in word embeddings. They used the WEAT test (Caliskan et al., 2017) to measure bias and mitigate it through a method called Hard-Debiasing. Finding out the gender bias in contextual embeddings was first done by Zhao et al. (2019). Bartl et al. (2020a) quantified and mitigated bias in BERT-like models.

Relatively lesser work is done in gendered and low-resource languages. Zhou et al. (2019) was one of the first papers to quantify bias in gendered languages like Spanish and French. To quantify the bias, they used a modified version of the WEAT test (Caliskan et al., 2017). Work done in Indic languages to reduce bias was first done by Pujari et al. (2019). They used an SVM classfier to quantify the bias. More recent work by Ramesh et al. (2021) quantified bias in English-Hindi machine translation. They used a TGBI metric to quantify the bias. Kirtane and Anand (2022) quantified and mitigated gender stereotypes in Hindi and Marathi word embeddings.

Our main contributions in this paper are creation of a novel Hindi dataset suitable to measure occupational gender bias in large language models and quantification of the bias in the model using the aforementioned dataset.

2 Dataset for analysis of gender bias

In this work, we create a dataset using professions and gendered nouns to study gender bias in Hindi. We measure gender bias using sentence templates as shown in Figure 1. The gendered nouns and professions are later filled into these templates to generate the corpus. We use pronouns in a way that remove any gender marking information from the templates.

Our contribution is the creation of the bias evaluation corpus with professions in Hindi that are gender-invariant (BEC-Pro-Hindi) (Bartl et al., 2020a). Our profession list is similar to the one in Kirtane and Anand. These professions are gender neutral, meaning the words themselves are gender-

*Equal Contribution

Original template	वह **[PERSON]** का काम **[PROFESSION]** का हैं वह **[PERSON]** एक **[PROFESSION]** हैं वह **[PERSON]** एक होनहार **[PROFESSION]** हैं	The **[PERSON]** works as a **[PROFESSION]** The **[PERSON]** is a **[PROFESSION]** The **[PERSON]** is a skilful **[PROFESSION]**
Person masked (PM)	वह **[MASK]** का काम **[PROFESSION]** का हैं वह **[MASK]** एक **[PROFESSION]** हैं वह **[MASK]** एक होनहार **[PROFESSION]** हैं	The **[MASK]** works as a **[PROFESSION]** The **[MASK]** is a **[PROFESSION]** The **[MASK]** is a skilful **[PROFESSION]**
Person + Profession masked (PPM)	वह **[MASK]** का काम **[MASK]** का हैं वह **[MASK]** एक **[MASK]** हैं वह **[MASK]** एक होनहार **[MASK]** हैं	The **[MASK]** works as a **[MASK]** The **[MASK]** is a **[MASK]** The **[MASK]** is a skilful **[MASK]**

Figure 1: Templates and masked templates.

agnostic, doctor for example. Additionally, we make a list of twelve nouns to represent masculine and feminine genders as well as a list of three gender-neutral nouns representing people for our bias comparison benchmark. Concretely, we start with a template that has a placeholder for the gendered noun and a placeholder for the profession. To quantify bias, we use a metric defined in Section 3. It requires two input sentences: one with the person noun masked and one with both the person and professions noun masked. The templates used to generate our dataset are shown in Figure 1. We obtain a total of 3330 sentences using 74 professions, 15 gendered and neutral nouns, and three templates.

3 Quantifying gender bias

We quantify the bias by evaluating the effect of the gendered nouns on the likelihood of the target, similar to the work of Bartl et al. (2020b). The model used for quantifying bias is the Muril model (Khanuja et al., 2021), a language model trained on Indian languages. For the evaluation of association bias, we first create the person masked sentence (PM) by masking the gendered noun in the original sentence and then a person + profession masked sentence (PPM) by masking both the token denoting the person and profession. We then calculate the probabilities P_{person}, P_{prior} of the mask being the person, given the person masked sentence and person + profession masked sentence as shown in the equations below.

$$P_{person} = P(Person = [MASK]|PM) \quad (1)$$

$$P_{prior} = P(Person = [MASK]|PPM) \quad (2)$$

Gender	Mean	Normalised Mean
Feminine	-4.173	-1.235
Neutral	-2.575	0.0
Masculine	-1.382	0.922

Table 1: Bias in Hindi language model

The Occupation Gender Bias (OGB), the bias evaluation metric, is then calculated as follows,

$$OGB = log(\frac{P_{person}}{P_{prior}}) \quad (3)$$

The metric measures the bias associated with professions and genders, of the pre-trained language models. The mean of this metric is calculated across the feminine and masculine nouns of the obtained log score for all professions. We also calculate the same for the gender-neutral nouns for comparison.

4 Results and Conclusion

Table 1 shows the mean of the OGB score over masculine, feminine and neutral nouns. We normalize the scores to compare the values. We observe that feminine nouns have a negative association bias while the masculine nouns have a positive association bias compared to the neutral nouns after normalising as shown in Table 1. The opposite signs indicate a presence of bias in these models.

We plan to reduce the above calculated bias by various mitigation strategies. We also intend to work on various other Indic Languages. Many of these languages are gendered, and a novel evaluation metric for such languages is needed.

References

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020a. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020b. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Neeraja Kirtane and Tanvi Anand. 2022. Mitigating gender stereotypes in Hindi and Marathi. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 145–150, Seattle, Washington. Association for Computational Linguistics.

Arun K Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 450–456.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online. Association for Computational Linguistics.

Clarice Wang, Kathryn Wang, Andrew Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2021. User acceptance of gender stereotypes in automated career recommendations. *arXiv preprint arXiv:2106.07112*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.