Occupational Gender stereotypes in Indic Languages

Neeraja Kirtane
Manipal Institute of Technology
MAHE, Manipal, India
kirtane.neeraja@gmail.com

Tanvi Anand
Manipal Institute of Technology
MAHE, Manipal, India
tanviaanand@gmail.com

Abstract

As the use of Natural Language Processing systems increases, there is a need to address the gender bias in these systems. While research is being done in the English language to quantify and eliminate bias, it is not the same for Indic Languages. Most Indic languages are gendered, i.e., each noun is assigned a gender. Therefore, evaluation differs from what is done in English. In this paper, we evaluate and find occupational bias in Indic languages. We prepare a dataset of gender neutral occupations and gendered occupation pairs, in Hindi and Marathi.

1 Introduction

Extensive research is being done in the English language to address the bias inherently present in NLP systems. Bolukbasi et al. (2016) paper was one of the first papers to address gender bias in the English language and attempted to debias it. However, this work is relatively nascent in Hindi, Marathi, and other Indic languages. Pujari et al. (2019) attempted to debias word vectors in the Hindi Language using the debiasing techniques used in Bolukbasi et al. (2016). Hindi is a language that is spoken by approximately 420 million people around the world. Around 120 million people speak Marathi. Unlike English, Hindi and Marathi are gendered languages, i.e., every noun is assigned a gender. In this paper, we find occupational gender stereotypes in Hindi and Marathi embeddings. Embeddings are commonly used in machine learning applications. Therefore, there is a need to identify the bias present and possibly try to mitigate it. We prepared the first-of-its-kind dataset for both gendered and gender-neutral occupations in Hindi and Marathi to enable further study in this area [1]. We propose new metrics to measure bias for both gendered and gender-neutral occupations.

2 Dataset

We asked two native speakers of Hindi and Marathi to prepare a list of 167 unique occupations in Hindi and 90 in Marathi. This dataset D consists of two parts D_{gen} and D_{neu}, for Hindi and Marathi each. Hindi and Marathi being gendered languages, some occupations have different names for their male and female counterparts. D_{gen} consists of multiple d_i, each being a word pair of a particular occupation. For example, in Hindi, लेखक "male writer" and लेखिका "female writer". D_{neu} consists of those occupations that have the same name irrespective of gender. For example, in Hindi वकील "lawyer". We further use S_{gen}, a collection of male and female seed words in Hindi and Marathi. For example, in Marathi तो "he" and ती "she".

3 Methodology

Since D_{gen} and D_{neu} are fundamentally different datasets, we use two different ways for evaluating bias. We use pre-trained ULMFiT language model (Howard and Ruder, 2018) for our embeddings. The perplexity of these embeddings is better than others for Indic Languages (Arora, 2020).

[1] https://github.com/neeraja1504/occupation_corpus

3.1 Method for gendered occupation nouns

We hypothesize that for one occupation pair, there should be a similar relation between the gendered words and the respective occupation word. If there is a gap, we know that bias exists. Building on the works of Zhao et al. (2020), our proposed formula is:

$$iBias = \frac{1}{|M|} \cdot \sum_{\vec{m} \in M} cos(\vec{O}_m, \vec{m}) - \frac{1}{|F|} \cdot \sum_{\vec{f} \in F} cos(\vec{O}_f, \vec{f}) \tag{1}$$

The score obtained should be ideally close to zero. O_m and O_f are male and female occupation word vectors respectively obtained D_{gen}, out of one occupation pair. M and F are are the male and female counterparts of S_{gen}

3.2 Metrics for non-gendered occupation nouns

Since we did not find a significant correlation in our data, Bolukbasi et al. (2016)'s Principal Component Analysis (PCA) method did not show notable results. PCA is often used as a dimensionality-reduction technique (Wold et al., 1987). So, it should be used mainly for variables which are strongly correlated. If the relationship is weak between variables, PCA will not work well to reduce the data. We found a weak relationship in our case. Hence, we suggest a modified evaluation metrics to measure bias in neutral occupation words.

3.2.1 Indic Direct Bias

A particular occupation ideally should not deviate towards either of the genders. Therefore, the difference between the similarities of one occupation with each of the genders should be negligible. We propose the formula:

$$iDB = \frac{1}{|M|} \cdot \sum_{\vec{m} \in M} cos(\vec{O}, \vec{m}) - \frac{1}{|F|} \cdot \sum_{\vec{f} \in F} cos(\vec{O}, \vec{f}) \tag{2}$$

The score obtained should be ideally close to zero. Here $\vec{O}$ is the occupation word vector obtained from D_{neu}. M and F are are the male and female counterparts of S_{gen}.

4 Results

Table 1 shows the results for gendered occupations, and Table 2 shows the results for gender-neutral occupations. After calculating the respective metrics, the sign of the metric shows the type of bias present. If the sign is negative, there is a female bias. If the sign is positive, there is a male bias.

Language	Occupation Pair	Translation	iBias Score	Male or female Bias
Hindi	नर्तक, नर्तकी	dancer	0.0501	Male Bias
Hindi	अभिनेता, अभिनेत्री	actor	-0.0155	Female Bias
Marathi	लेखक, लेखिका	writer	-0.0128	Female Bias
Marathi	गायक, गायिका	singer	0.0088	Male Bias

Table 1: Results for gendered occupations

5 Conclusion and Future Work

In this work, we propose two metrics to quantify inherent gender bias present in Hindi and Marathi word embeddings. We intend to work on debiasing the existing bias. Bias variation based on embeddings used is also another scope of the study. There are other biases based on religion, race, caste, etc., which we also want to look into further.

Language	Occupation	Translation	iDB	Male or female bias
Hindi	किसान	farmer	0.0012	Male Bias
Hindi	मज़दूर	wage worker	-0.0035	Female Bias
Marathi	आचारी	cook	-0.0006	Female Bias
Marathi	तंत्रज्ञ	technician	0.0139	Male Bias

Table 2: Results for gender neutral occupations

References

Gaurav Arora. 2020. inltk: Natural language toolkit for indic languages. *arXiv preprint arXiv:2009.12534*.

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*, 29:4349–4357.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender bias in natural language processing across human languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54.

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.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Haiyang Zhang, Alison Sneyd, and Mark Stevenson. 2020. Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. *arXiv preprint arXiv:2010.02847*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*.

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. *arXiv preprint arXiv:1909.02224*.