Evaluating Gendered and Religious Stereotypes in FastText's Hindi Word Vector Embeddings

Neel Shah
Northwestern University / 643 Library Place, Evanston, IL.
neelshah2023@u.northwestern.edu

Abstract

There has been a surge in research on the social biases inherent in word embedding models, which represent each word by a vector. However, much of this literature has been focused on languages such as English and Spanish. In this paper I quantify the stereotypes captured in Facebook's FastText word embedding model for the widely-spoken Indian language Hindi. I investigate two relationships: between gender and neutral profession terms; and between religions and negative adjectives. Much as with work on other languages, I find that in the Hindi embedding, specific professions are strongly associated with specific genders (such as nursing with women), and that certain negative adjectives are strongly associated with religious groups (such as *terrorist* with Muslims). Given the ever-increasing use of word-embedding models in tasks such as machine translation, the identification (and later, resolution) of these problems becomes imperative.

1 Introduction

Assessing the biases in word embedding models has become a subject of great interest in computational linguistics. Much of this work focuses on high-resource languages such as English, but in recent years the advent of embeddings for lower-resource languages such as Hindi has pushed researchers such as Gupta et al. (2021) in similar directions. This paper, to my knowledge, is the first to turn these methods on to FastText's open-source Hindi embedding model. I conduct two experiments, looking at the link between gender and professions; and between religious groups and negative adjectives.

2 Data

The pre-trained vector embeddings that I analyze were produced by Facebook's FastText using a CBOW model and character n-grams of length 5. These vectors are originally 300-dimensional, and were trained using CommonCrawl and Wikipedia. A detailed discussion of their architecture can be found in Grave et al. (2018).

For simplicity, however, I use a modified version of these embeddings. I make two key changes. First, I only consider the most popular 100,000 tokens, since these encompass most commonly-used words while omitting highly technical or English-loan terms. Secondly, I reduce the dimensionality of the vectors to 50 using Principal Component Analysis (PCA). In particular, I use the sklearn.decomposition.PCA class to accomplish this. This was necessary due to the limited computing power at my disposal, but small-scale replications on the 300-dimensional embeddings produced similar results.

3 Methods

A key benefit of word embeddings is that the geometric relationships between vectors capture semantic relationships between the corresponding words. One such geometric relationship is cosine similarity, which returns $\cos\theta$ where θ is the angle between two vectors. This maps onto semantic similarity, with words like *bat* and *ball* having higher cosine similarities than semantically unrelated ones like *bat* and *arthritis*. This is helpful when assessing biases. The word *doctor* is a seemingly gender-neutral term, but in many embeddings it shows more cosine similarity to *man* than *woman*. The general approach, then, is to compare the cosine similarities of a set of neutral terms such as professions to a set of identity-based terms.

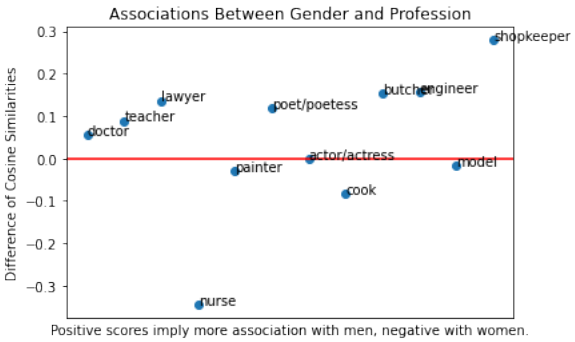
Hindi, however, presents a special case of this problem. Languages such as Hindi exhibit morphological agreement on gender, meaning that the word for a male teacher (*śikasaka*) is different from the one

for a female teacher (*śikasikā*). As a consequence, the word *śikasaka* (male teacher) naturally aligns closer to *ādamī* (man) than *aurata* (woman). This, as Zhou et al. (2019) note, is due to morphological agreement and should not be considered a bias.

In order to rectify this, I ensure my comparisons are aligned where necessary–I compare the word for female teacher (*śikasikā*) to woman (*aurata*), and the word for male teacher (*śikasaka*) to man (*ādamī*).

For the first experiment (gender-profession), I subtract the cosine similarity of the profession to the word *ādamī* (man) from its similarity to the word *aurata* (woman), meaning that a positive score indicates a bias towards men, and vice-versa. The second experiment (religion-adjectives) similarly reports the similarity scores for the words Hindu (*hindū*) and Muslim (*musalamāna*) with respect to the list of negative adjectives.

4 Results



The results of the two experiments can be seen in Table 1 and the figure above.

The first experiment (gender-profession) gives largely predictable results. Words like *nurse* are more closely aligned to women, while words like *doctor* are more closely aligned with men. There are, however, some surprising results–the word *shopkeeper* (*dukānadāra*) exhibits the strongest alignment of all professions to men.

The second experiment also demonstrates the prevalence of stereotypes. As Table 1 shows, negative adjectives common in global stereotypes such as *terrorist* (*ātamkavādī*) show higher similarities to the word Muslim. Common Indian stereotypes–such as the association of Muslims to ghettoes (*basatī*) and dirtiness (*gamdagī*)–also find their way into the embeddings in comparison to the 'control group' of Hindus.

Adjective	Hindu	Muslim
terrorist	0.61	0.67
dirty	0.12	0.19
pakistani	0.51	0.69
traitor	0.56	0.61
patriot	0.57	0.63
angry	0.20	0.23
ghetto	0.47	0.49

Table 1: Religion and Similarity to Chosen Adjectives.

5 Discussion

Thus it is clear that the FastText word embeddings exhibit religious and gendered biases, likely inherited from the training data. FastText's Hindi model is trained using Wikipedia and CommonCrawl, with the latter indiscriminately gathering a corpus using the Hindi-language web. This model is one of the only embeddings for Hindi, meaning that the risk of use in downstream applications such as machine translation is high.

To this end, there has been discussion of methods to debias vector embeddings, such as in Pujari et al. (2019), who use an SVM-based classifier on Hindi, and the more general approaches discussed by Bolukbasi et al. (2016). This is the natural next step that follows a bias-identification project such as mine.

Other work such as Garg et al. (2018) has attempted to take a historical perspective on the biases in word-embeddings, seeing how they evolve over time. The lack of a large, digitized corpus for Hindi makes this difficult, but this is another avenue for further exploration.

References

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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Gauri Gupta, Krithika Ramesh, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation. *arXiv preprint arXiv:2106.08680*.

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.

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*.