A Simulation Study Exploring the Role of Pronouns in Gender Stereotypes

Rayyan Merchant and Shengyu Liao

INTRODUCTION

Currently, American society is becoming more and more conscious of its usage of gendered pronouns, coinciding with growing visibility and acceptance of transgender and genderqueer individuals. As such, many have begun shifting their speech to more commonly include gender-neutral 3rd person pronouns, which typically takes the form of the more familiar 'they'. Neopronouns have also become more visible on the Internet, while still being relatively unused by the vast majority of the LGBTQ+ community.

Word embeddings are a powerful framework to represent each word as a vector, such that the geometric relationship between these vectors captures semantic relations between the corresponding words. Recent studies demonstrate that word embeddings capture gender stereotypes in the large corpora of training texts. For example, the vector for the occupation 'architect' would be close to the vector for 'man', whereas the vector for 'nurse' would be closer to 'woman'. The embedding algorithm automatically learns the stereotypes, and they can be problematic if the embedding is then used for sensitive applications such as search ranking, sentiment analysis, or question retrieval. Though lots of research has been devoted to developing algorithms to debias the word embeddings, none of them investigated the role of pronouns in reducing gender stereotypes.

With this in mind, our project aims to simulate 'what if Anglophone society refrained from using gendered third-person pronouns entirely, and instead switched to gender-neutral alternatives. We chose to simulate two options: the (commonly used) 'they', and a neopronoun 'ze'. By checking the associations between gendered words and occupations in text, we can view how gender stereotypes may be impacted by society changing its usage of pronouns.

DATA

For our project, we utilized the Europarl parallel corpus, a corpus extracted from the proceedings of the European Parliament. Since we did not require text in a language other than English, we only used the English-language corpus, with 2,218,201 sentences and 53,974,751 words contained within.

This corpus was then extracted from its XML file format, and tagged using TreeTagger, with the British National Corpus tagset. From the tagged corpus, we then changed all gendered pronouns to either (singular) 'they' or the neopronoun 'ze'. Using TreeTagger, we were able to determine what type of pronoun each one was, and therefore accurately convert them to their equivalent form, despite some forms being the same.

After data preprocessing, three models were trained using Word2Vec: baseline model, they model, and ze model.

METHODS

Given two vectors, their similarity can be measured by their cosine similarity, as in Eq. **1**.

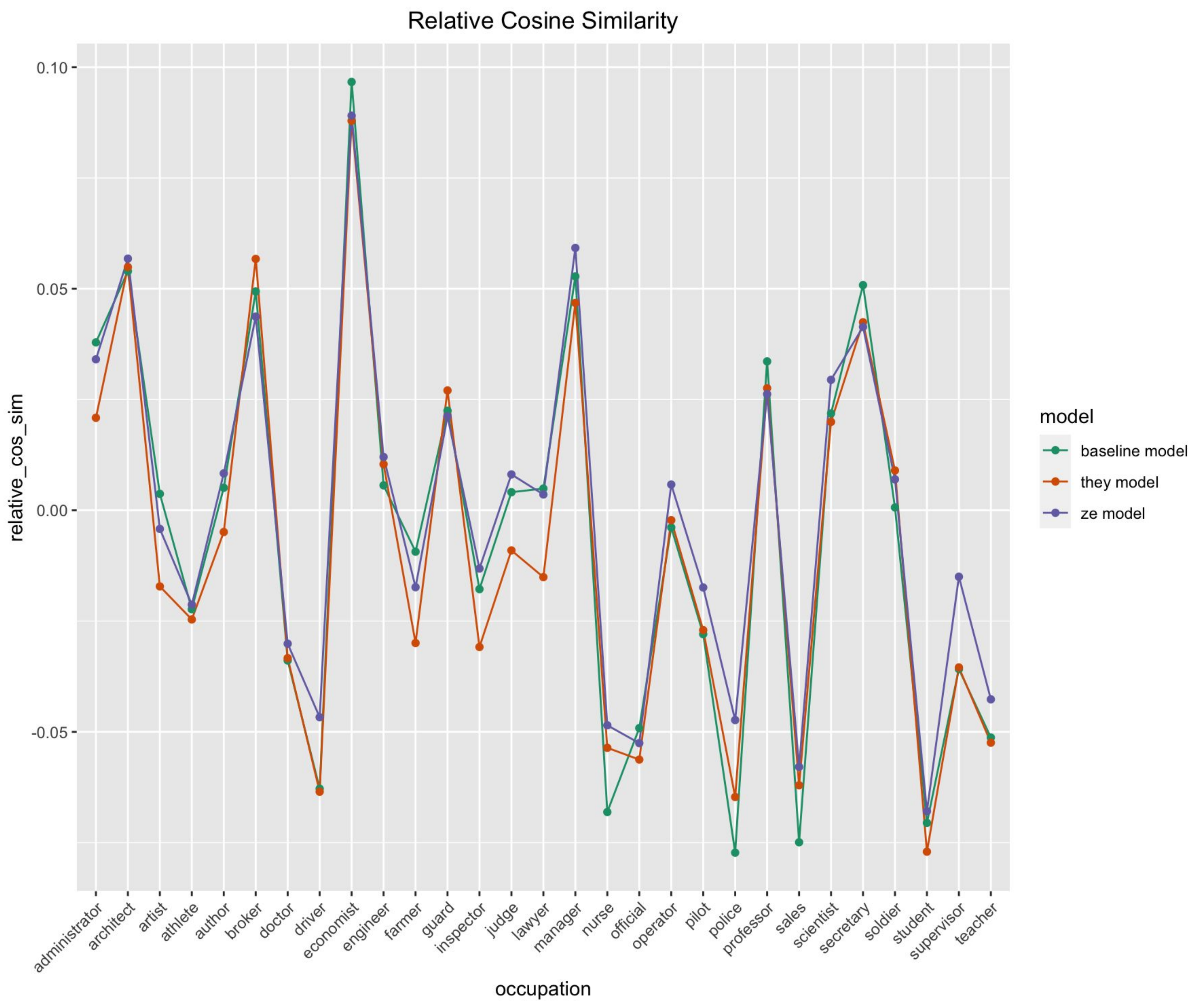
$$cosine(a, b) = \frac{a \cdot b}{|a||b|} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}} \qquad \textbf{[1]}$$

The association between the female/male words and occupation words is calculated as follows: Construct a group vector by averaging the vectors for each word in the female/male group; construct a vector for each occupation word by averaging the vectors for each word and its plural form; then calculate the similarity between the group vector and each occupation word vector.

The relative cosine similarity, which captures the relative strength of association of an occupation word with respect to two groups, is as described in Eq. **2**, where v_m is the average vector for each occupation word, v_1 is the average vector for group one (female), and v_2 is the average vector for group two (male). The more positive (negative) that the relative cosine similarity is, the more associated the occupation word is toward group two (one). The closer that the value is to zero, the less biased the occupation word is, thus the less the gender stereotype there exists.

$$relative\ cosine\ similarity = cosine(v_m, v_2) - cosine(v_m, v_1) \qquad \textbf{[2]}$$

RESULTS



RESULTS (cont.)

Because we conducted three paired t-tests, the new Bonferroni-corrected alpha level is 0.05/3 = 0.017: a *p*-value has now to be below 0.017 to be treated as significant.

- they model vs. baseline model: no significant difference, though they model is more negative and farther from zero than baseline model, as tested in two-tailed paired t-test [T(28) = 1.78, *p* = 0.09; they model mean: -0.00882 (SD 0.0428); baseline model mean: -0.00557 (SD 0.0452)].
- ze model vs. baseline model: ze model is less negative and closer to zero than baseline model, though this difference was not statistically significant (near significant), as tested in two-tailed paired t-test [T(28) = -2.40, *p* = 0.02; ze model mean: -0.00125 (SD 0.0394); baseline model mean: -0.00557 (SD 0.0452)].
- ze model vs. they model: ze model is less negative and closer to zero than they model [ze model mean: -0.00125 (SD 0.0394); they model mean: -0.00882 (SD 0.0428)]. This difference was significant as tested in two-tailed paired t-test [T(28) = -5.02, *p* < 0.001]. Cohen's d was 1.32, which indicates a large effect size.

CONCLUSIONS & LIMITATIONS

- The use of the singular 'they' did not affect gender stereotypes in any significant manner.
- Replacing gendered pronouns with the neopronoun 'ze' resulted in a decrease in the gender bias of occupations as compared to the baseline model. This change was near statistically significant.
- The use of the neopronoun 'ze' led to significantly less 'occupational' gender bias compared to the use of the singular 'they', with a large effect size.
- Our models are limited in that they are decontextualized. People can use different sets of pronouns in different situations, but our models failed to capture the nuances in this.
- Some occupation words are polysemous, e.g., judge, pilot, which may affect the results. Thus, future research may need to take PoS information into consideration.

REFERENCES

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of the Tenth Machine Translation Summit*, 79–86.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceeding of the First International Conference on Learning Representations: Workshop Track*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems, Volume 2*, 3111–3119.