Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230250

The presence of bias based on pre-trained language models

Rui Liu[1,†], Yiqi Wu[2,†], Yufei Xing[1,3,†]

[1]School of Computing and Cyber Security (Oxford Brookes College), Chengdu University of Technology, Chengdu, Sichuan, 610059, China
[2]School of Advanced Technology, Xi'an Jiaotong-liverpool University, Suzhou, Jiangsu, 215123, China

[3]xyf010818@gmail.com
[†]These authors contributed equally.

Abstract. One of the most active fields of machine learning research is natural language processing (NLP). Although existing linguistic machine learning models excel numerically on a variety of linguistic comprehension tasks, they frequently lack implicit bias reduction optimization. To ensure that deep learning models can avoid the traps of implicit bias and that robots can make fair judgments, bias in NLP must be addressed adequately. The ramifications of permitting biased models to reach the actual world are serious. Thus must address this issue as quickly as feasible. This paper conducted data validation-bias experiments on several real datasets to verify the presence of gender bias in the pre-trained models, then proposed a word vector balancing algorithm to modify the actual vector representation of words by biasing the models and verifying the effectiveness of the debiasing method through debiasing experiments to mitigate the gender bias of the models while ensuring data accuracy, thereby improving the fairness of the classification results. Furthermore, this paper provides more accurate information for the future development and use of deep learning.

Keywords: natural language processing, transformer, bias, BERT, word embeddings.

1. Introduction

In recent years, deep learning has been widely used in information retrieval, data mining, and many other fields, as it allows effective decision-making in complex tasks. However, training data contains biased information about disadvantaged groups and ethnic minorities. Information about biases in the dataset is learned and reinforced by the deep learning model during the training process, which can directly affect the use of artificial intelligence systems and present new challenges to developing research and applications of machine learning techniques [1].

Therefore, the study of de-biasing algorithms in deep learning can improve the fairness of classification results, reduce the negative impact of evaluation results, and provide information to inform more accurate assessments, which is expected to play a significant role in the future development and use of deep learning [2]. Bias in NLP has been reported for various tasks [3]. For example, "She is a nurse, he is a doctor" is the outcome of translating "He is a nurse, she is a doctor" into Hungarian and back to English [4]. Occupations like doctor or computer programmer are considered male, while nurse or housewife are considered female [4].

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/8/20230250

The relationships held by word vectors are far removed from the original meaning of any word. The model learned from it reflects that bias leads to injustice. Various undesirable biases may be captured and reproduced in the data in languages produced by various popular pre-trained text encoders such as BERT [5]. These language models train word embeddings on a human-generated corpus [6]. When incorporating biases into word embeddings [6], downstream tasks that rely on these biased word vectors will also produce biased results. Whereas deep learning models learn and amplify the biased information present in the dataset during training, resulting in biased predictions.

Therefore, it is necessary to take measures to reduce these biases, such as gender bias and racial bias. To solve this problem, there has been much research in recent years on reducing bias. However, many debias papers rarely consider the impact of debias on the ability to solve NLP tasks, as models that fail to solve NLP tasks may emerge. In contrast, debias are being mitigated [7]. Most pre-trained algorithms today employ gender-biased word embeddings [6] from human-generated corpora. The more commonly used post-hoc debiasing algorithms today are not suitable for mitigating or further removing gender bias from word embeddings [6,8].

Early hard-debiasing methods, which reduced gender bias by post-processing the gender-related parts of word embeddings, were not considered to be particularly effective, with gender bias still present in the geometry of the embeddings after debiasing, while in recent years it has been found that earlier debiasing algorithms ignored the effect of word frequency statistics on the spatial structure of a word embeddings, and the resulting Double-Hard Debias approach effectively mitigated the negative effect of word frequency features on the gender debiasing of word embeddings [6,9, 10-12].

However, despite the apparent reduction of gender bias in these debias algorithms, it is still present in the current pre-trained model. A new approach to the detection of gender bias has been proposed, which uses BERT as a case study and finds that the use of the attention matrix, w_q and w_k, in the model introduces a more severe problem of gender bias compared to other modules in the model, and that the introduction of various components of the attention mechanism such as residual connectivity enhances gender bias [5, 13].

In this paper, to show that a pre-trained language model contains gender bias, the experiment will take the following procedure: First, texts containing biased content are collected from existing training datasets, converted into word vector representations using existing language training models, and then calculate the cosine similarity between each word vector separately. Based on the experimental output data, analysed the presence of gender bias in the chosen model provided suggestions on ways to mitigate the gender bias in the model. Then, present an equalization algorithm for gender word vector that modifies the actual vector representation of words by equalizing gender-specific word groups, and neutralizing the model distance, thus mitigating bias against non-gender words.

The rest of the paper is structured as follows. A summary of the suggested approaches is provided in Section 2, together with a description of the general study framework and the methodologies employed to carry out the procedure. The experimental design, presentation, and explanation of the results are all shown in Section 3. Section 4 presents the attempted debiasing algorithm. Section 5 talks about the limits of work, and the final section summarize the key findings, the importance of the study, and the next research directions.

2. Methods

2.1. Dataset

OntoNotes dataset aims to annotate a variety of large text corpora (FastText, GloVe, and Bert) in three languages (English, Chinese, and Arabic) [5, 14, 15]. Experience shows that when individual levels like syntactic parsing structures, annotate propositional structures and semantic role labels with a high degree of consistency (inter-tagger agreement) and use other datasets with machine learning models to predict these structures, but with slightly lower accuracy. To generate a comprehensive, accurate corpus that incorporates annotations of various layers of shallow semantic structures in the text, this paper first

DOI: 10.54254/2755-2721/8/20230250

converts the text in the OntoNotes dataset into an embedding vector, splits the dataset into a training set, a validation set, and a testing set, then extracts the gender text labels from them.

2.2. Model

FastText is a fast text classification algorithm that speeds up training and testing while maintaining high accuracy without the need to pre-train word vectors [14]. FastText's final output uses layered SoftMax, which greatly optimizes the speed of the model and introduces n-gram to classify sentences, which retains only important features in the trained model, e.g., the addition of 2-grams allows the model to distinguish between "I love her" and "she loves me." This paper uses this model to obtain word vectors by setting and adjusting a series of parameters, including word embedding dimension, number of data cycles, learning rate, etc., which facilitates the rapid implementation of the model structure in the experiments of this paper.

GloVe is a counting and global statistics-based word representation tool [15]. The model is an unsupervised learning algorithm whose count-based approach has advantages in terms of efficiency in capturing global statistics and is more able to distinguish between relevant and irrelevant words. The experiments use this model to obtain word vector representations of individual words in a sentence, which counts the occurrences of a word before and after a given word as co-occurrence counts, obtains a word frequency co-occurrence matrix between words, and calculates the probability of occurrence of a word in the vicinity of a given word. At first, construct the co-occurrence matrix, then calculate the approximate relationship between the word vector and the co-linearity matrix, and construct the loss function to implement the GloVe model finally.

BERT is a language representation model that has been pre-trained [5]. To create deep bidirectional language representations during pre-training, it is necessary to use the novel masked language model (MLM) rather than the conventional unidirectional language model or shallow stitching of two unidirectional language models. The MLM is used to pre-train bi-directional Transformers to build deep bi-directional language representations, one of this model's key features. Only add an output layer after pre-training for fine-tuning to attain cutting-edge performance in diverse downstream jobs. No task-specific structural adjustments to the BERT are required during this procedure.

2.3. Experiment setting

As shown in figure 1, by collecting texts containing biased content from the existing training dataset, building a fast classifier using FastText, GloVe and Bert, obtaining each word and any N-gram vector, storing the word vector and N-gram vector separately, and then using cosine similarity to measure the similarity of the vectors, and according to the experimental output data, to analyze the presence of gender bias in the selected model.

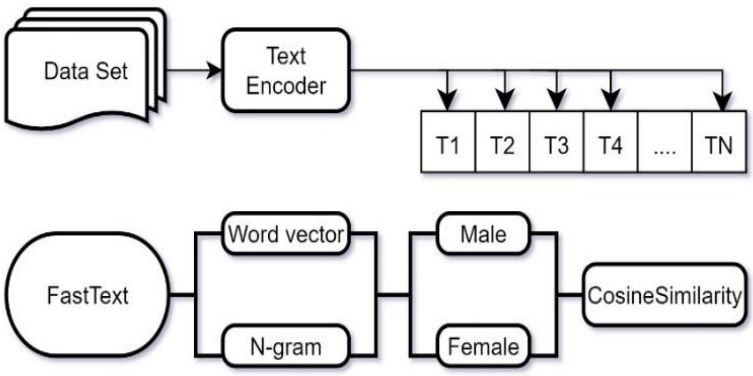


Figure 1. Experimental procedure.

DOI: 10.54254/2755-2721/8/20230250

In 2016, Joulin and Grave et al. proposed a word vector and text classification tool, with a performance equivalent to deep learning and considerably faster, which may give a simple and efficient methodology for text classification and representation learning [14]. The degree of similarity between any two-word vectors may then be determined using the cosine similarity algorithm, which is represented as [16]:

$$\text{CosineSimilarity}(\text{u}, \text{v}) = \frac{\text{u.v}}{||u||_2||v||_2} = \cos(\theta) \quad (1)$$

3. Results and analysis

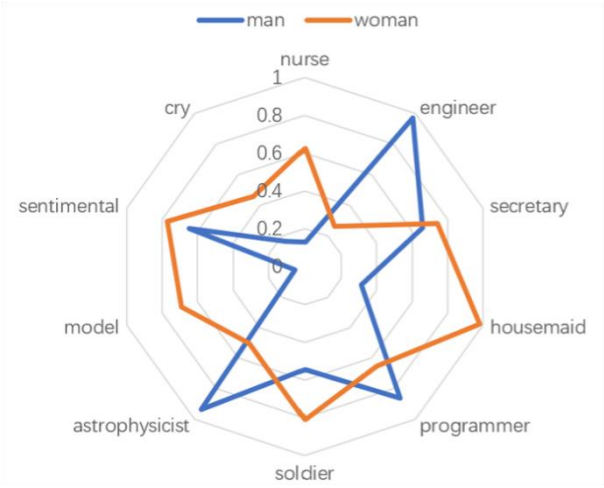


Figure 2. In FastText (The samples of gender bias).

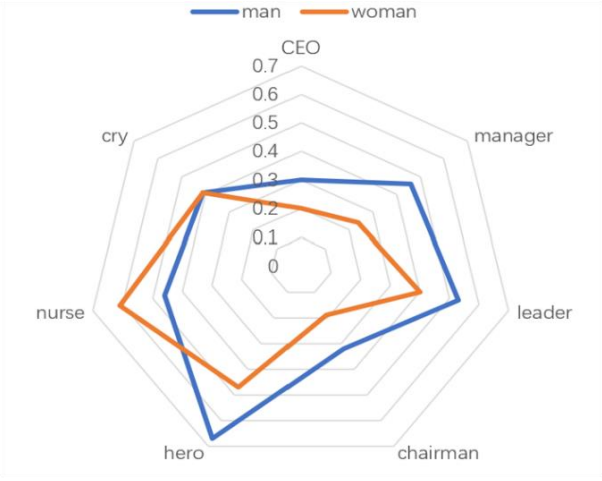


Figure 3. In GloVe (The samples of gender bias).

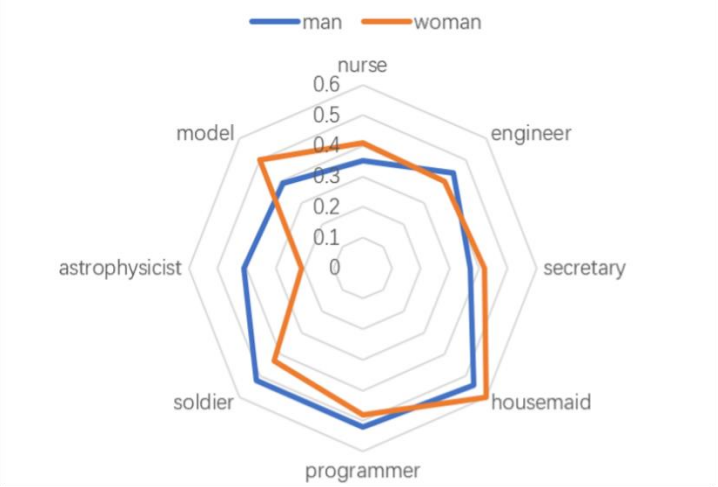


Figure 4. In BERT (The samples of gender bias).

Figures 2, 3 and 4 illustrate that, in the current pre-trained models, gender inequality has consistently been associated with prejudiced societal norms, including values, beliefs, attitudes, and practices that define social roles and power relations between men and women. Calculate these figures by computing the cosine similarity between the pronouns "He" and "She" with terms identifying different professions. The overall findings show that while males are expected to be breadwinners, women frequently confront significant conventional cultural expectations to be carers and homemakers. This indicates that whereas women are more likely to associate word vectors with indoor or service jobs, males are more likely to associate them with managerial or traditionally high social status employment. Hence, despite equal

DOI: 10.54254/2755-2721/8/20230250

access for men and women to health care and education, many countries still do not allow women and girls to realize their full potential.

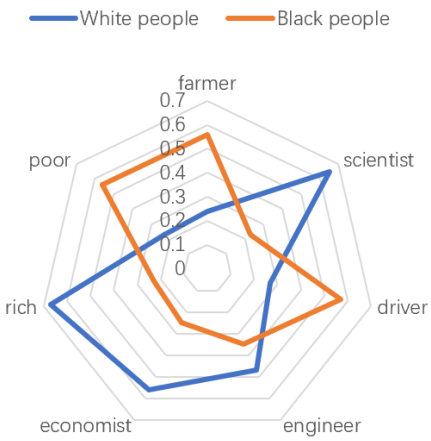


Figure 5. In FastText (The samples of racial bias).

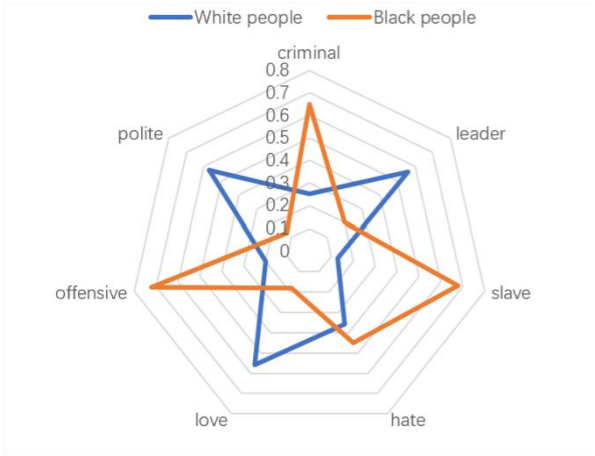


Figure 6. In GloVe (The samples of racial bias).

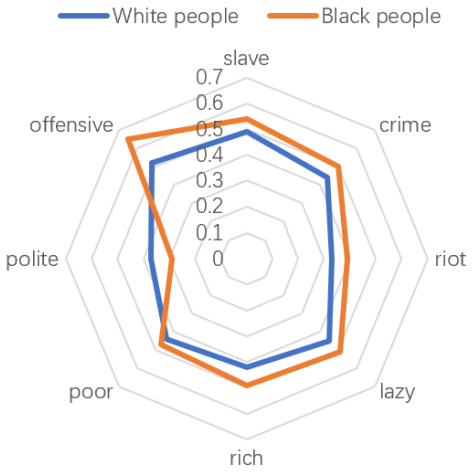


Figure 7. In Bert (The samples of racial bias).

In figures 5, 6 and 7, a clear racial bias among the present-day pre-trained models is evident by counting the cosine similarity between 'black' and 'white' with words denoting different occupations, images and feelings. The results illustrate that Blacks are often perceived as offensive or poor, while Whites are more polite and wealthy. Word vectors reflect racial discrimination that the white race is a noble being, while the black race is a slave symbol of cheap labour. The white race typically has more social power and resources than the black race, and it uses these advantages, frequently in conjunction with coercive methods like discriminatory legislation, low educational attainment, unequal economic opportunities, etc., to maintain the black race's position of inequality.

DOI: 10.54254/2755-2721/8/20230250

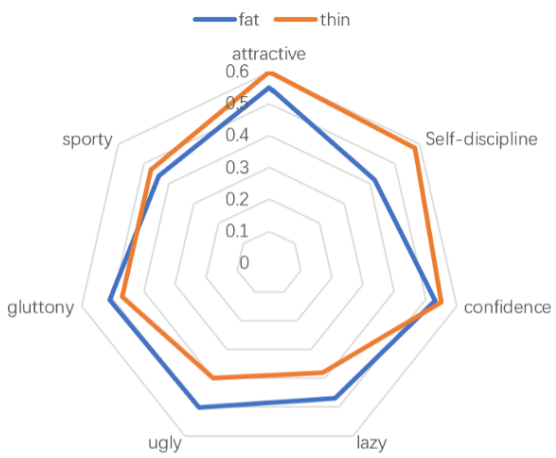


Figure 8. In FastText (The samples of obesity bias).

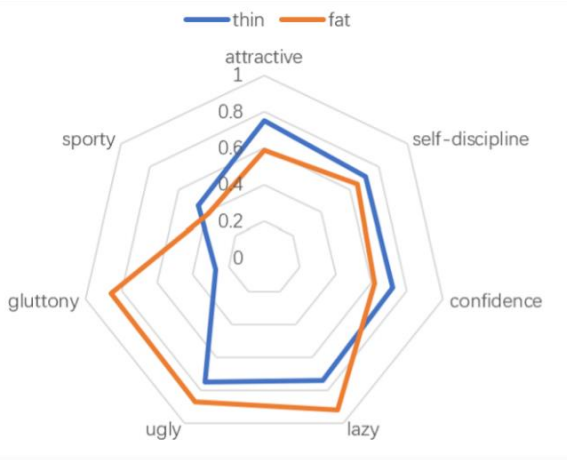


Figure 9. In GloVe (The samples of obesity bias).

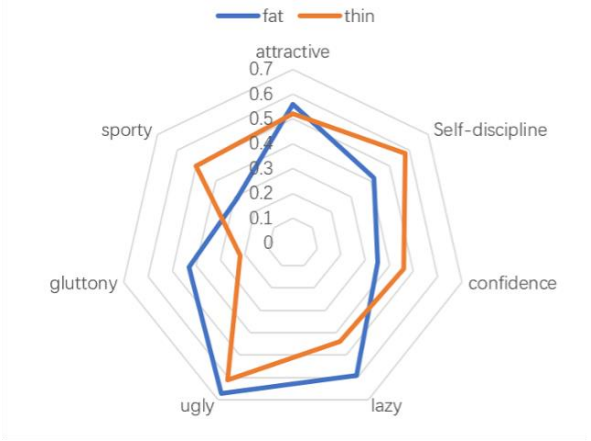


Figure 10. In Bert (The samples of racial bias).

Figures 8, 9 and 10 indicate that in the current pre-training model, obesity is often easily equated with laziness, gluttony, lack of willpower and lack of self-discipline, associating certain arbitrary moral judgements with body size, socially devaluing and stigmatizing obese people, and leading to prejudice, negative stereotyping and discrimination against these people. The discrimination against fatness in human society seems to be a genetically embedded cognitive bias, involving the denial, discrimination and stigmatization of the 'obese', even the 'judgement of character' of a person based on their size, and these negative prejudices against obese people permeate a wide range of contexts, creating inequalities in the workplace, healthcare and educational opportunities.

As can be seen from the results of these experiments, all three models monitor different levels of bias concerning gender, race and body size. Among them, the FastText model shows serious bias in both gender and race, while the bias present in the GloVe model is mainly presented in the dimensions of race and body size; and even in the latest Bert model, stereotypes in gender, race and body size are presented.

4. Debias algorithm

For some specific words, such as 'actor' and 'actress' or 'grandmother' and 'grandfather,' there should be a gender difference. Other terms, such as 'secretary' and 'manager,' should not be gendered and should be treated differently when dealing with them.

DOI: 10.54254/2755-2721/8/20230250

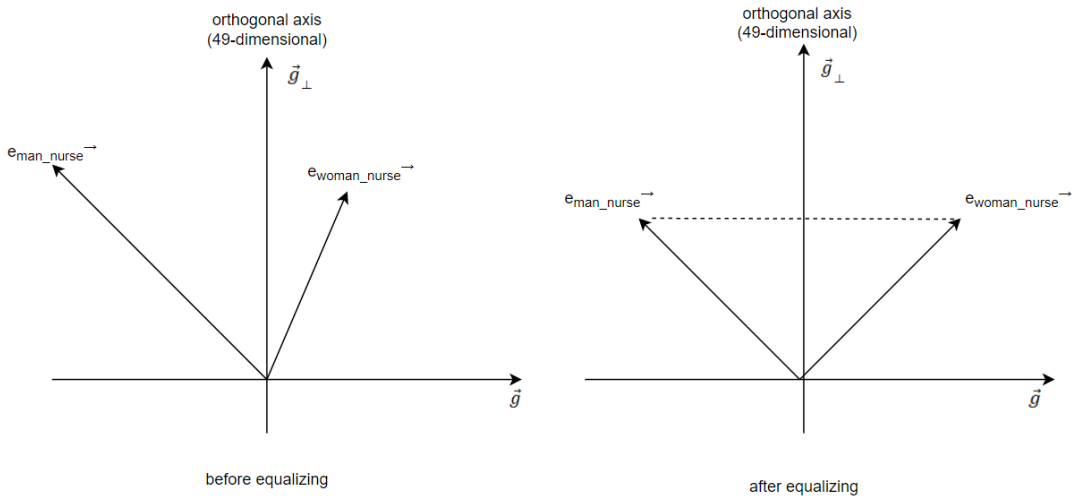


Figure 11. Attempt to debias using a balancing algorithm (using nurse as an example).

Figure 11 illustrates the process of attempting to de-bias gender in this paper. If using a 50-dimensional word embedding, then the 50-dimensional space can be divided into two parts: the bias direction g, which represents gender bias, and g⊥, which is the remaining 49 dimensions. Orthogonalize the 49-dimensional g⊥ to g at an angle of 90 degrees.

In neutralizing the bias, a vector such as enurse, with the bias direction (g) zeroed, is taken to obtain the de-biased enurse. Afterwards, equalize the gender-specific phrases to ensure that a particular pair of words is equally distant from g⊥ in 49 dimensions. For example, if "woman" is significantly closer to nurse in "man" and "woman", then use a balanced approach to make the distance between the word "nurse" and the words "woman" and "man" equal. As shown in Figure 4, define the bias direction of g using more than just male and female, and calculate the average distance between multiple gender vectors, e.g., g1 = ewoman - eman, g2 = emother - efather, g3 = egirl - eboy, after which the average g = avg (g1, g2, g3), which will allow a better estimation of the "gender" dimension of the word embedding space.

Based on previous research, the de-biasing process for this equilibrium algorithm is derived as follows [17]:

$$u = \frac{e_{w1}+e_{w2}}{2} \tag{2}$$

$$u_B = \frac{u.bias_axis}{||bias_axis||_2^2} * bias_axis \tag{3}$$

$$u_\perp = u - u_B \tag{4}$$

$$e_{w1B} = \frac{e_{w1}.bias_axis}{||bias_axis||_2^2} * bias_axis \tag{5}$$

$$e_{w2B} = \frac{e_{w2}.bias_{axis}}{||bias_{axis}||_2^2} * bias_axis \tag{6}$$

$$e_{w1B}^{corrected} = \sqrt{|1 - ||u_\perp||_2^2|} * \frac{e_{w1B}-u_B}{||(e_{w1}-u_\perp)-u_B||_2^2} \tag{7}$$

$$e_{w2B}^{corrected} = \sqrt{|1 - ||u_\perp||_2^2|} * \frac{e_{w2B}-u_B}{||(e_{w2}-u_\perp)-u_B||_2^2} \tag{8}$$

$$e_1 = e_{w1B}^{corrected} + u_\perp \tag{9}$$

DOI: 10.54254/2755-2721/8/20230250

$$e_2 = e_{w2B}^{corrected} + u_{\perp} \tag{10}$$

5. Discussion

This paper only verifies the presence of gender bias in current pre-trained language models (e.g., FastText, GloVe and Bert) by calculating the cosine similarity from word vectors and does not detect various biases in the models from different perspectives using multiple methods in terms of multidimensionality. This fairness is fragile and short-lived and may disappear once the model is affected by the latest data. Also, it did not generate data after using the balanced vector debiasing algorithm to demonstrate the general effectiveness of the algorithm in removing gender bias from existing pre-trained language models.

6. Conclusion

By converting text containing biased content from existing training data into word vectors and calculating the remaining cosine similarity, this study demonstrates that there is a clear gender bias in the FastText, GloVe and Bert models, which define occupations such as models and secretaries as female, and correspondingly define engineers and astrophysicists as male, with the stereotype of men in more leadership occupations being particularly evident. Meanwhile, this paper uses a method of calculating word vector pre-similarity to compensate for the fact that earlier Debias methods ignored the effect of word frequency statistics on the spatial structure of word embeddings, and then, by using Python to automatically train and test the language models and using a balanced vector approach to tune the new language models.

This paper predicts that the theory will still apply to other models such as ELETRA and XLNet but has not tested the method in situations such as racial bias and religious bias, once apply the method to these scenarios above, the calibration templates need to be regenerated. However, there are some shortcomings in the investigation, like many de-emphasis' papers, which do not consider the impact of de-emphasis on the ability to solve NLP tasks. Today, language models will continue to play a significant role in society, for reducing prejudice, discrimination and other problems, it is necessary to continue exploring the technical aspects. The experimental results suggest that gender equity is possible if a gender-neutral corpus is used to refine the language model. Therefore, future debias algorithms based on pre-trained models should minimize the reliance of language models on a corpus of human-generated text, increase the computational speed of model training, improve the requirements for memory training, adapt training methods appropriately, and strive to eliminate bias while also need to be aware of any significant degradation in model performance and address the root causes of bias.

References

[1] Cheng L, Ge S, Liu H. "Toward Understanding Bias Correlations for Mitigation in NLP" [J]. arXiv preprint arXiv:2205.12391, 2022.

[2] Mandis I S. "Reducing Racial and Gender Bias in Machine Learning and Natural Language Processing Tasks Using a GAN Approach" [J]. International Journal of High School Research, 2021, 3(6): 17-24.

[3] Schick T, Udupa S, Schütze H. "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp" [J]. Transactions of the Association for Computational Linguistics, 2021, 9: 1408-1424.

[4] Sun T, Gaut A, Tang S, et al. "Mitigating gender bias in natural language processing: Literature review" [J]. arXiv preprint arXiv:1906.08976, 2019.

[5] Devlin J, Chang M W, Lee K, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding" [J]. arXiv preprint arXiv:1810.04805, 2018.

[6] Wang S, Zhou W, Jiang C. "A survey of word embeddings based on deep learning" [J]. Computing, 2020, 102: 717-740.

[7] Birknerová Z, Benková E, Lajčin D, et al. "IDENTIFICATION OF THE EFFECTS OF NEURO-LINGUISTIC PROGRAMMING TECHNIQUES IN MANAGERIAL ENVIRONMENT IN

DOI: 10.54254/2755-2721/8/20230250

THE CONTEXT OF GENDER DIFFERENCES" [J]. Polish Journal of Management Studies, 2022, 26(2): 60-74.

[8] Alabdulmohsin I M, Lucic M. "A near-optimal algorithm for debiasing trained machine learning models" [J]. Advances in Neural Information Processing Systems, 2021, 34: 8072-8084.

[9] Gonen H, Goldberg Y. "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them" [J]. arXiv preprint arXiv:1903.03862, 2019.

[10] Mu J, Bhat S, Viswanath P. "All-but-the-top: Simple and effective postprocessing for word representations" [J]. arXiv preprint arXiv:1702.01417, 2017.

[11] Gong C, He D, Tan X, et al. "Frage: Frequency-agnostic word representation" [J]. Advances in neural information processing systems, 2018, 31.

[12] Wang T, Lin X V, Rajani N F, et al. "Double-hard debias: tailoring word embeddings for gender bias mitigation" [J]. arXiv preprint arXiv:2005.00965, 2020.

[13] Li B, Peng H, Sainju R, et al. "Detecting gender bias in transformer-based models: a case study on Bert" [J]. arXiv preprint arXiv:2110.15733, 2021.

[14] Joulin A, Grave E, Bojanowski P, et al. "Fasttext. zip: Compressing text classification models" [J]. arXiv preprint arXiv:1612.03651, 2016.

[15] Pennington J, Socher R, Manning C D. "Glove: Global vectors for word representation" [C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[16] Zhestiankin B, Ponomareva M. "Zhestyatsky at SemEval-2021 Task 2: ReLU over Cosine Similarity for BERT Fine-tuning" [J]. arXiv preprint arXiv:2104.06439, 2021.

[17] Bolukbasi T, Chang K W, Zou J Y, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings" [J]. Advances in neural information processing systems, 2016, 29.