2024 IEEE 18th International Conference on Semantic Computing (ICSC) | 979-8-3503-8535-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSC59802.2024.00036

Detection of Latent Gender Biases in Data and Models Using the Approximate Generalized Inverse Method

Takafumi Nakanishi
Department of Data Science
Musashino University
Tokyo, Japan
takafumi.nakanishi@ds.musashino-u.ac.jp

***Abstract*—Gender bias creates inequalities in roles, expectations, and opportunities between males and females. When such biases are incorporated into artificial intelligence models, the corresponding technological solutions and products can further entrench the social biases. Herein, a new method for investigating the extent to which latent biases in text-based training data affect a language model is presented. Potential gender bias is identified by deriving values assigned to male/female words via inverse operations from embedded expressions to the original words using the approximate inverse model explanation (AIME). In particular, AIME constructs approximate generalized inverse operators for black-box models. A biased embedded representation used in machine learning models as an internal representation of word/sentence vectors likely introduces bias into the overall prediction results of such models. The OpenAI text-embedding-ada-002 large language model, which provides embedded expressions, is employed to determine the gender bias included in the proposed method. Experimental results show that the OpenAI text-embedding-ada-002 model is partially gender-biased owing to the training text data. These results are expected to (i) contribute to the development of effective measures preventing gender bias during the design and training of language models, (ii) promote the identification and mitigation of gender bias in future language models, and (iii) provide insights into the effect of language models and their limitations from technical, social, and cultural perspectives.**

Keywords: gender bias, approximate inverse model explanation, language model, bias detection, embedding representation

I. Introduction

Machine learning models, especially high-performance and versatile large language models (LLMs), are extensively employed in various commercial and academic applications, including information retrieval, sentence generation, and question-answering. Therefore, understanding the effects of the potential biases and prejudices in the training data of these models on their output is essential.

Gender bias creates inequalities in roles, expectations, and opportunities between males and females, and the technological solutions and products obtained using artificial intelligence (AI) models that incorporate such biases can further entrench social norms. For example, a gender-biased job-matching AI yields discriminatory solutions, such as less technical job recommendations for female candidates. Therefore, identifying AI models with hidden biases and developing methods to correct such biases is crucial for improving the performance of such models and for social and ethical transformations. In particular, explainable AI (XAI) technology is the key to addressing this issue. Thus far, we have proposed one XAI technique, i.e., the approximate inverse model explanation (AIME) [1], which identifies biases by constructing approximate inverse operators of a black-box model and is thus expected to contribute to building a society that respects diversity.

This study is focused on the word-embedding representation mechanism in machine learning. Embedded representations are used in machine learning models as internal representations of vectors of words or sentences. Bias in this embedded representation likely introduces bias into the overall prediction results of these models. Therefore, a new method is proposed to investigate the extent to which latent biases in text-based training data affect a language model. The proposed method identifies potential gender biases based on values assigned to male/female words derived via inverse operations on expressions embedded in the original words using AIME. In particular, AIME constructs approximate generalized inverse operators for black-box models. In addition, the OpenAI text-embedding-ada-002 model, widely recognized as a representative of the advances in natural language processing (NLP) technology, is employed. The text-embedding-ada-002 model is not categorized as an LLM; however, its fundamental technology, training data size, and OpenAI origin indicate its strong correlations to LLMs. Therefore, approaching bias detection in LLMs appears feasible. A system implementing the proposed method was developed for the OpenAI text-embedding-ada-002 model, which is widely used in NLP, and gender bias was analyzed.

II. Related Works

Blodgett et al. [2] surveyed 146 reports on "bias" analysis in NLP systems and noted that the definition and motivation of bias are ambiguous in many reports. Their study provided directions for research on bias in NLP and encouraged conceptualization of bias to clarify and focus on the actual experiences of communities affected by the NLP system. Bolukbasi et al. [3] surveyed 304 papers on gender bias in NLP and identified the limitations of this method and directions for future research. Zhao et al. [4] proposed a new benchmark for assessing gender bias, WinoBias, in occupation data, which contained two challenging statements requiring linking gendered pronouns to stereotypically male or female occupations collected from the US Department of Labor. Squazzoni et al. [5] investigated gender bias in peer

979-8-3503-8535-9/24/$31.00 ©2024 IEEE
DOI 10.1109/ICSC59802.2024.00036

review using data from 145 diverse research journals and found that manuscripts solely authored/co-authored by women were treated more favorably by reviewers and editors. Rudinger et al. [6] investigated gender bias in a co-reference resolution task, determining whether two or more words or phrases mentioned in a sentence refer to the same entity; they studied whether the English co-reference resolution system showed any bias in resolving references to words or phrases associated with a particular gender (e.g., professional names). Sharma et al. [7] proposed an evaluation method to measure these biases by combining gender-neutral assumptions with gender-specific hypotheses using occupational names for BERT, RoBERTa, and BART to confirm the presence of gender stereotypes. Sun et al. [8] critically discussed the possible sources of gender bias incorporation in NLP systems, including training data, resources, pretrained models, and algorithms, and the limitations of the current de-biasing methods. These previous studies highlight several methods developed for reducing gender bias in NLP.

Several studies on gender bias in embedded language expressions have also been reported. Basta et al. [9] evaluated gender bias in contextualized word embedding and showed that word embedding can retain and even amplify the gender bias present in the current data source. Furthermore, contextualized word embeddings have lesser gender bias than non-contextualized standard word embeddings. Caliskan [10] showed that semantics contain a human-like bias (generated from GloVe word embedding) automatically derived from text corpora and demonstrated how texts capture semantics, cultural stereotypes, and empirical associations. Garg et al. [11] developed an index to identify the evolution of historical trends and social changes in gender stereotypes and attitudes toward minorities in the US in the 20th and 21st centuries, beginning in 1910; the changes in the embedding resembled the population and occupational changes over time. Kurita et al. [12] used BERT to predict masked tokens and measured the bias encoded in the representation using the log-probability bias score. May et al. [13] proposed the Sentence Encoder Association Test, an extension of the Word Embedding Association Test (WEAT), as a new test method for measuring social bias in sentence encoders, such as ELMo and BERT. Swinger et al. [14] developed an algorithm for enumerating biases in word embeddings associated with sensitive characteristics such as race and gender, including publicly available word embeddings, especially the "de-biased" ones. Zhao et al. [15] quantitatively analyzed gender bias in the contextualized word vectors of ELMo and explored methods of reducing it, showing that male entities are much more common than female entities in ELMo training data. Zhao et al. [16] addressed gender bias in visual cognitive tasks involving language, such as captioning, visual question answering, and visual semantic role labeling, and found that activity related to "cooking" in the training set was 33% more likely to involve women than men, indicating a new research direction for this study. To address this issue, a Lagrangian relaxation-based algorithm for collective inference was designed to include corpus-level constraints to adjust existing structured prediction models. The proposed method reduced the magnitude of bias amplification by 47.5% (for multi-label classification) and 40.5% (for visual semantic role labeling), with a negligible loss of recognition performance. Zhao et al. [17] extended WEAT to detect and remove multiclass bias in word embeddings, introduced a new test for detecting multiclass bias, namely, Multiclass WEAT, and proposed a new algorithm called HardWEAT to remove bias. Kotek et al. [18] investigated the behavior of LLMs overwhelmed with gender stereotypes using the WinoBias [4] gender bias dataset. Finally, Nemani et al. [19] provided a comprehensive survey of gender bias in the transformer model, noting that despite the recognition of gender bias, methods of assessing it remain unexplored.

Here, a method for detecting potential bias using AIME that can be uniformly applied to most embedding methods is proposed. The method constructs an approximate inverse operator of the original embedding method and calculates its effects on words associated with male/female via inverse operations. In contrast to previous approaches [14–24], which focused on specific datasets or tasks, the proposed method is more versatile and applicable to various datasets and tasks. In the proposed AIME-based method, potential gender bias is recognized by deriving the value assigned to male/female words.

III. Detection of Gender Bias Using AIME

For a black-box model that accepts x as the input and $\hat{y}$ as the estimation output, AIME [1] derives explanations by constructing a generalized inverse operator $A^{\dagger}$ that estimates $\hat{\boldsymbol{x}}$ with $\hat{\boldsymbol{y}}$ as the input and then performs inverse operations to elucidate why $\hat{\boldsymbol{y}}$ is derived from input $\boldsymbol{x}$. The concept of this approach stems from backward calculations performed to verify the original calculations. By framing it as an inverse problem, we can achieve a more intuitive and straightforward explanation than the conventional forward problem approach.

Here, $\hat{\boldsymbol{y}}$ is a black-box model that derives an embedded representation from a word x represented by a bag of words, and a generalized inverse operator is constructed to estimate the word $\hat{\boldsymbol{x}}$ from the embedded representation $\hat{\boldsymbol{y}}$. An appropriate number of pairs of $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$ are prepared to construct X, $\hat{Y}$. X is a matrix of vectors for the bag-of-words representations of words. Note that X is a unit matrix if x is always a vector representing only one word (this characteristic is important in the later formulation), whereas $\hat{Y}$ is a matrix that arranges the embedded representations. These relationships are illustrated in Fig. 1.

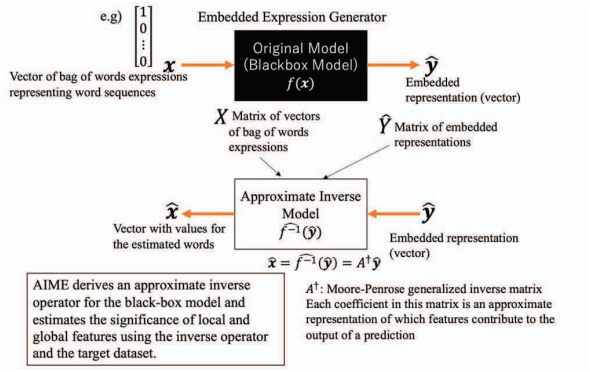

Fig. 1. Conceptual diagram of the proposed method.

$A^{\dagger}$ constitutes the generalized inverse operator and is obtained using the following equation:

$$\begin{aligned} X &= A^{\dagger}\hat{Y}, \\ X\hat{Y}^T &= A^{\dagger}\hat{Y}\hat{Y}^T, \\ X\hat{Y}^T\left(\hat{Y}\hat{Y}^T\right)^{-1} &= A^{\dagger}\left(\hat{Y}\hat{Y}^T\right)\left(\hat{Y}\hat{Y}^T\right)^{-1}, \end{aligned} \tag{1}$$

$$A^{\dagger} = X\hat{Y}^{T}\left(\hat{Y}\hat{Y}^{T}\right)^{-1} = X\hat{Y}^{\dagger}.$$

In other words, if the generalized inverse operator $A^{\dagger}$ is applied to the generated embedded representation $\hat{\boldsymbol{y}}$ ($A^{\dagger}\hat{\boldsymbol{y}}$), then a vector $\hat{\boldsymbol{x}}$ of the bag-of-words representations is obtained, where the original word is represented but only as an approximation. Matrix $A^{\dagger}$ is the Moore–Penrose generalized inverse of matrix A [20, 21]. Matrices $A^{\dagger}$, X, and $\hat{Y}$ do not necessarily need to be squared. The only requirement is that X and $\hat{Y}$ have a common number of data points (or records). Additionally, matrix $\hat{Y}\hat{Y}^{T}$ must be invertible for the computation to be valid. This generalized inverse enables the pseudo-inverse to be computed even when the matrix is not square, thus providing flexibility in handling different matrix dimensions. If the embedded representation $\hat{\boldsymbol{y}}$ contains gender bias, then different values for male and female words exist in $\hat{\boldsymbol{x}}$. Gender bias is detected by obtaining a priori embedded expression for the words and sentences likely to have gender bias and by considering whether these calculations have male/female values or which is greater.

IV. Method

A. Overview of the Proposed Method

An overview of the proposed method is presented in Fig. 2. The scheme consists of two major phases: an approximate inverse operator generation phase and a generalized inverse operator derivation phase. The approximate inverse operator generation phase generates an approximate inverse model of the embedded expression generator by preparing a base vocabulary group, embedding the vocabulary group, and finally computing $A^{\dagger}$ from the base vocabulary group as a combined matrix X of the vectors of bag-of-words expressions and a combined matrix $\hat{Y}$ of the embeddings. The generalized inverse operator derivation phase uses the approximate inverse model generated in the former phase (approximate inverse operator generation) to perform the inverse operation and embed words and sentences likely to have gender stereotypes. This method can construct approximate inverse operators using an available set of vocabularies (that can be expressed by the embedded expression generator) and a set of embedded expressions for each vocabulary. The vocabulary set can be a subset of the vocabulary supported by the original embedded expression generator. However, the maximum possible number of vocabulary sets with similar distributions must be created to derive approximate inverse operators with high accuracy.

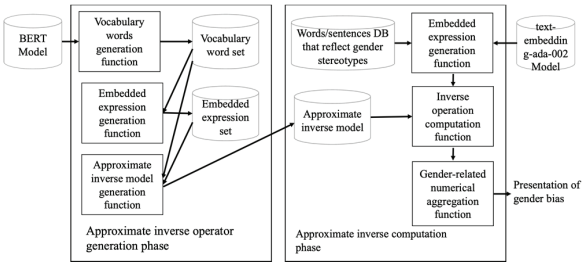

Fig. 2. Overview of the proposed method.

The embedding model text-embedding-ada-002 published by OpenAI converts text into vector embeddings and may share some concepts with other architectures, such as the GPT series. Thus, this model may be employed as a reference for analyzing the bias status of OpenAI LLMs. The embedded expression generation function, which appears in two places in Fig. 2, generates the embedded expressions of words and word sequences using the text-embedding-ada-002 model, enabling the detection of latent gender bias in this model. The bias-detection experiments were performed specifically for Japanese (as explained later).

B. Approximate Inverse Operator Generation Phase

In the approximate inverse operator generation phase, an approximate inverse operator of the target embedded expression generation function is constructed.

1) Vocabulary word generation function: The vocabulary word generation function extracts all the words that can be expressed by the target embedded expression generation function as vocabulary words. Extracting all the words from the OpenAI text-embedding-ada-002 model was impossible. Therefore, the cl-tohoku/bert-base-japanese-whole-word-masking model of BERT was used, described as the "BERT Model" in Fig. 2. This model is typically used as a Japanese model for BERT; moreover, as it is trained using data derived from the Japanese version of Wikipedia, the model is not biased toward any field. The vocabulary words of this model can be treated as those of text-embedding-ada-002, which is an unbiased approximation. Using this function, we obtained 32,000 vocabulary words, which were then represented as a vector of the bag-of-words form and arranged to obtain the identity matrix X.

2) Embedded expression generation function: This function embeds words using the OpenAI text-embedding-ada-002 model. The 32,000 words obtained using the vocabulary word generation function were converted into embedded representations by this function, followed by the generation of matrix $\hat{Y}$.

3) Approximate inverse model generation function: The approximate inverse operator $A^{\dagger}$ of the text-embedding-ada-002 model was derived using matrices X and $\hat{Y}$, which were described in 1) and 2), respectively, and the formulas presented in Section III. As X is an identity matrix, the approximate inverse operator $A^{\dagger}$ can be obtained using the following relationship:

$$A^{\dagger} = \hat{Y}^{T}\left(\hat{Y}\hat{Y}^{T}\right)^{-1} = \hat{Y}^{\dagger}, \tag{2}$$

In other words, the Moore–Penrose [20, 21] generalized inverse of $\hat{Y}$ is the approximate inverse operator $A^{\dagger}$.

C. Approximate Inverse Computation Phase

In the approximate inverse computation phase, the approximate inverse operator constructed in Section IV-B can be used to transform the embedded expressions representing the gender-biased words into bag-of-words expressions, and gender bias can be identified based on their values, which may be related to men/women, or differences in size.

1) Embedded expression generation function: This function converts and embeds words and sentences that reflect gender stereotypes and are stored in the word/sentence database (DB) into expressions using the OpenAI text-

embedding-ada-002 model. Accordingly, each word/sentence in DB is represented by a vector $\hat{\boldsymbol{y}}$.

2) Inverse operation computation function: $A^{\dagger}$, derived in Section IV-B, acts on $\hat{\boldsymbol{y}}$, i.e., $(A^{\dagger}\hat{\boldsymbol{y}})$, and the resultant word vector has the bag-of-words form. If the male- and female-related components of this vector have significantly different values, then gender bias is confirmed.

3) Gender-related numerical aggregation function: For the aforementioned vector, the averages of the element values corresponding to predetermined male- and female-related words were determined. We aimed to identify the presence and extent of latent gender bias in the model by observing the magnitude of the difference between these average values for male and female words.

These sequential steps were followed to identify the latent gender bias in the OpenAI text-embedding-ada-002 model, the target model used in this study.

V. Experiments

A. Experimental Environment

This experiment was specifically focused on the Japanese owing to the birth origin and cultural familiarity of the author; however, the method and its results can be generalized for people from diverse races and ethnic backgrounds. Future work may include a comparison of the gender bias in different countries.

This experiment was conducted to detect potential gender bias in the OpenAI embedding model. The model has been trained on diverse datasets and is therefore suitable for detecting biases in texts with diverse contexts and backgrounds owing to its ability to generate one of the highest quality embeddings currently available, generality, and open source/public availability, which enable validation and replication of experiments. However, because extracting all vocabulary from this model is not feasible, the cl-tohoku/bert-base-japanese-whole-word-masking model of BERT, which yielded 32,000 vocabulary words, was adopted. This BERT model, trained on Japanese Wikipedia, is popular for Japanese NLP tasks. As Wikipedia articles are collaboratively edited by a diverse group of users, reaching a consensus from various perspectives is presumed to have minimal bias. This vocabulary size is adequate for obtaining an approximate inverse operator. In addition, compared to LIME [22] and SHAP [23], AIME [1] produces simple and clear male-/female-related word values.

In Experiment 2 (Section V-C), the keywords—generated interactively using the web version of ChatGPT 4.0—utilized to reflect gender stereotypes were stored in a DB called "Words/sentences DB that reflects gender stereotypes." The keywords were grouped into four categories: occupation, family roles, adjectives, and hobbies and activities, with several words selected for each category. The selected keywords are listed in Table I. Among the 32,000 vocabulary words, "man (男)," "male (男性)," "boy (男子)," and "male child (男の子)" were used as words related to men, and "woman (女)," "female (女性)," "girls (女子)," and "young girl (女の子)" were used as words related to women. As described in IV-C-3, to identify gender bias, the mean of the values corresponding to "man (male)," "male (male)," "boy (male)," and "male child (boy)" was calculated as the value for men, while that of the values corresponding to "woman (female)," "female (female)," "girls (girls)," and "young girl (female)" was determined as the value for women. Subsequently, these two mean values were compared to locate any male/female value. The experimental system was implemented in Python 3.10.12—using the packages openai 0.28.0, matplotlib 3.7.1, numpy 1.23.5, pandas 1.5.3, seaborn 0.12.2, and transformers 4.33.2—and run on Google Colab Pro+.

TABLE I. List of Selected Keywords for Detecting Gender Bias

Occupation	**Family roles**	**Adjectives**	**Hobbies and activities**
Nurse	Mother	Strong	Cooking
Engineer	Father	Gentle	Sports
Teacher	Wife	Dependable	Reading
Soldier	Husband	Emotional	Car repair
Doctor	Older sister	Proactive	Traveling
Secretary	Younger brother	Passive	Gardening
Pilot	Younger sister	Independent	Shopping
Cook	Older brother	Submissive	Fishing
Hairdresser	Grandfather	Aggressive	Hiking
Architect	Grandmother	Calm	Watching movies
Writer			
Accountant			
Police officer			
Dancer			
Journalist			
Programmer			
Truck driver			
Pharmacist			
Researcher			
Artist			

B. Experimental Results: Verification of Words Likely to Have Gender Stereotypes

In this experiment, we detected gender bias in occupations, family roles, adjectives, hobbies, and activities, with gender stereographs determined by interacting with ChatGPT 4.0 (Table I). The results for occupations are shown in Fig. 4. Overall, gender bias was relatively small, except in some occupations, as the difference between male and female bias was small. For nurses, although the values were negative for both males and females, the absolute values were higher for males than for females, suggesting that the bias against male nurses may be high. However, traditionally, in Japan, nursing is often seen as a female occupation, which may suggest certain stereotypes and expectations. Hairdressing is traditionally viewed as a female profession as well, and accordingly, the female values were greater than the male values, possibly reflecting this stereotype. In addition, truck drivers exhibited large values for males, and this difference was the largest in the list, indicating strong male bias in truck driving. The bias results related to family roles are shown in Fig. 5. Words related to family roles showed a strong bias toward gender corresponding to that role—a trend widely recognized and attributed to several factors, including roles and relationships within a family, social expectations and stereotypes, and cultural backgrounds.

The bias results regarding adjectives are shown in Fig. 6. For "strong," social stereotypes dictate that males are strong; for "gentle," positive and negative biases exist for males and females, respectively. "Dependable" had positive values for both males and females but larger values for males, possibly

due to leadership expectations. “Emotional” had negative values for males and slightly positive values for females. “Independent” had a strong positive bias toward females, backed by the evaluation of female independence. “Submissiveness” had positive and negative values for males and females, respectively, possibly due to expectations of submissiveness and independence. The negative values for “calm” for both genders indicate its gender-independent nature.

The bias results for hobbies and activities are presented in Fig. 7. “Cooking” showed a strong positive bias toward females. “Sports” had a strong positive bias toward males. “Car repair” had positive values for both genders but a strong positive bias for males. “Travel” had negative values for both genders. “Gardening” had negative values for both genders but particularly strong negative values for females. “Shopping” had negative values for males and was mostly neutral for females. “Fishing” and “watching movies” exhibited strong positive biases toward females. The strong positive bias observed toward females in “fishing” was surprising considering the traditional image of fishing in Japan. However, this result may be due to the recent increase in the number of females adopting fishing as a hobby. In addition, the results were similar to the biased stereotypes held by the Japanese.

The OpenAI text-embedding-ada-002 model can generate embedded expressions in sentences, suggesting the feasibility of confirming latent gender bias in a text using the proposed method. The sentences “This strategic thinking is essential to the success of the project,” “I would like to try a new knitting pattern,” and “I attend Pilates classes three times a week” were embedded into the OpenAI text-embedding-ada-002 model, and approximate inverse operators were utilized to derive the male/female values, as shown in Fig. 8. The sentence “This strategic thinking is essential to the success of the project” exhibits male bias. The sentence “I would like to try a new knitting pattern” has positive values for both males and females; however, a strong bias toward females is present. The sentence “I attend Pilates classes three times a week” has female bias. Although this is only one example, detecting gender bias in the sentences is possible.

Overall, these results indicate that although the probability of gender bias in the OpenAI text-embedding-ada-002 model is insignificant, it yields biased results for certain examples and events.

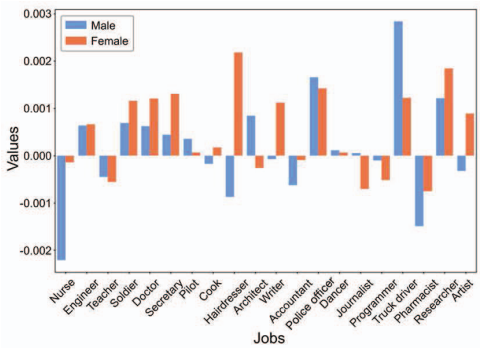

Fig. 3. Gender bias in the text-embedding-ada-002 model for occupations.

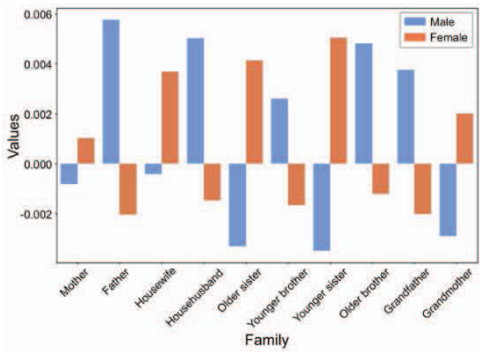

Fig. 4. Gender bias in the text-embedding-ada-002 model for family roles.

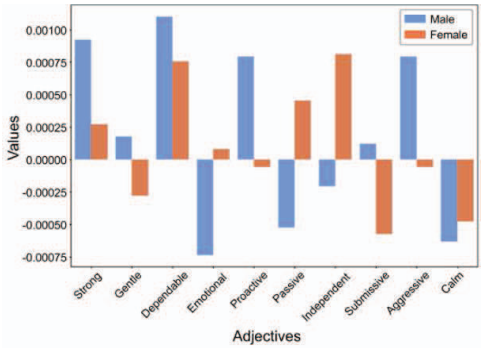

Fig. 5. Gender bias in the text-embedding-ada-002 model for adjectives.

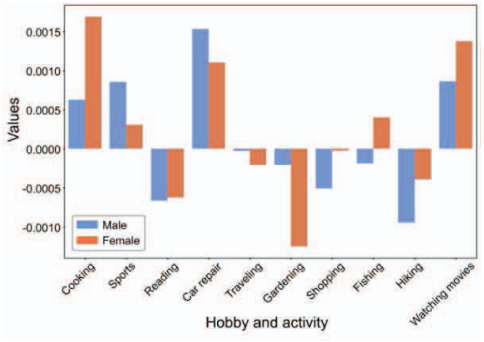

Fig. 6. Gender bias in the text-embedding-ada-002 model for hobbies and activities.

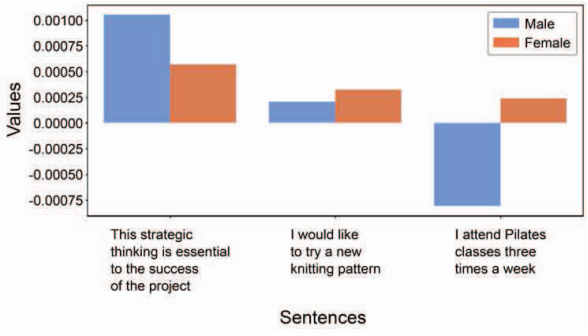

Fig. 7. Gender bias in the text-embedding-ada-002 model for sentences that may have gender stereotypes.

VI. Conclusions

In this study, AIME was applied to develop a new method for detecting potential biases in language models and data. The text-embedding-ada-002 multilingual model of OpenAI was employed to conduct experiments specifically on Japanese subjects. The results obtained in these experiments reveal that the OpenAI text-embedding-ada-002 model has a low overall gender bias. However, some words and phrases are more biased toward either gender. In particular, the results of Experiment 2 revealed that bias is present in words and sentences related to gender stereotypes; however, the values were small, indicating the challenges in determining the presence of large biases. Notably, the AIME method cannot capture the entire OpenAI text-embedding-ada-002 dataset. Thus, discrepancies in the findings may emerge owing to the approximation of the inverse operator. Nevertheless, the small differences observed between males and females suggest that some degree of gender bias exists in the model, highlighting a feasible method of considering bias while designing and training language models. Although this study provides valuable insights into gender biases in language models, it has certain limitations. For instance, our method primarily focuses on gender bias, without considering other potential biases. Additionally, using a specific model and dataset may not capture the full spectrum of biases present in diverse datasets. This method also requires a set of words representing the overall distribution, even if it is hypothetical, thus highlighting a major limitation of the proposed method.

Given these limitations, future work should focus on detecting other biases, such as racial and ethnic, age, religious, and geographic and nationality biases. Furthermore, developing remedial measures to address such biases is crucial. In addition, research on detecting differences in bias across countries and applying this method to other language models is necessary. The development of technology to eliminate bias without the loss of accuracy is also a major issue. Employing these efforts would result in the development of fair and transparent language models and bias-aware machine learning algorithms.

Acknowledgment

The author would like to thank Editage [www.editage.com] for English language editing. The author acknowledges the role of OpenAI's ChatGPT AI system in facilitating discussions and inspiring innovative ideas during the writing process. During the preparation of this manuscript, DeepL and ChatGPT were used to improve readability and language. After using these tools, the author reviewed and edited the content as required. The author takes full responsibility for the published content.

References

[1] T. Nakanishi, "Approximate Inverse Model Explanations (AIME): Unveiling local and global insights in machine learning models," IEEE Access, vol. 11, pp. 101020–101044, 2023. DOI: 10.1109/ACCESS.2023.3314336.

[2] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A critical survey of 'bias' in NLP," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 5454–5476, 2020. Association for Computational Linguistics.

[3] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," Adv. Neural Inf. Proc. Syst., vol. 29, 2016.

[4] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in arXiv preprint, arXiv:1804.06876, 2018.

[5] F. Squazzoni, G. Bravo, M. Farjam, A. Marusic, B. Mehmani, M. Willis, A. Birukou, P. Dondio, and F. Grimaldo, "Peer review and gender bias: A study on 145 scholarly journals," Sci. Adv., vol. 7, no. 2, eabd0299, 2021. DOI:10.1126/sciadv.abd0299.

[6] R. Rudinger, J. Naradowsky, B. Leonard, and B. V. Durme, "Gender bias in coreference resolution," in arXiv preprint, arXiv:1804.09301, 2018.

[7] S. Sharma, M. Dey, and K. Sinha, "Evaluating gender bias in natural language inference," in arXiv preprint, arXiv:2105.05541, 2021.

[8] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K. W. Chang, and W. Y. Wang, "Mitigating gender bias in natural language processing: Literature review," in arXiv preprint, arXiv:1906.08976, 2019.

[9] Basta, M. R. Costa-Jussà, and N. Casas, "Evaluating the underlying gender bias in contextualized word embeddings," arXiv preprint, arXiv:1904.08783, 2019.

[10] A. Caliskan, "Semantics derived automatically from language corpora contain human-like biases," Science, vol. 356, pp. 183–186, 2017. DOI: 10.1126/science.aal4230.

[11] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," Proc. Natl. Acad. Sci., vol. 115, no. 16, pp. E3635–E3644, 2018.

[12] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," arXiv preprint, arXiv:1906.07337, 2019.

[13] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," arXiv Preprint, arXiv:1903.10561, 2019.

[14] N. Swinger, M. De-Arteaga, N. T. Heffernan IV, M. D. Leiserson, and A. T. Kalai, "What are the biases in my word embedding?" in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), New York, NY, USA, 2019, pp. 305–311. DOI: 10.1145/3306618.3314270.

[15] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K. W. Chang, "Gender bias in contextualized word embeddings," arXiv preprint, arXiv:1707.09457, 2019.

[16] J. Zhao, T. Wang, and M. Yatskar, V. Ordonez, and K. W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," arXiv preprint, 2017.

[17] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K. W. Chang, "Learning gender-neutral word embeddings," arXiv preprint, arXiv:1809.01496, 2018.

[18] H. Kotek, R. Dockum, and D. Q. Sun, "Gender bias and stereotypes in large language models," arXiv preprint, arXiv:2308.14921, 2023.

[19] P. Nemani, Y. D. Joel, P. Vijay, and L. F. F. Liza, "Gender bias in transformer models: A comprehensive survey," arXiv preprint arXiv:2306.10530, 2023.

[20] E. H. Moore, "On the reciprocal of the general algebraic matrix," Bull. Am. Math. Soc., vol. 26, pp. 294–300, 1920.

[21] R. Penrose, "A generalized inverse for matrices" in Mathematical Proceedings of the Cambridge Philosophical Society. Cambridge: Cambridge University Press, vol. 51, no. 3, 1955, pp. 406-413. DOI: 10.1017/S0305004100030401.

[22] M. Ribeiro, S. Singh and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier", Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 97-101, 2016.

[23] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions", Proc. Neural Inf. Process. Syst. (NIPS), vol. 30, pp. 10, 2017, [online] Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.