EFFECTIVENESS OF DEBIASING TECHNIQUES: AN INDIGENOUS APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

An indigenous perspective on the effectiveness of debiasing techniques for pre-trained language models (PLMs) is presented in this paper. The current techniques used to measure and debias PLMs are skewed towards the US racial biases and rely on pre-defined bias attributes. Some require large datasets and further pre-training. Such techniques are not designed to capture the underrepresented indigenous populations in other countries, such as Māori in New Zealand. Local knowledge and understanding must be incorporated to ensure unbiased algorithms, especially when addressing a resource-restricted society.

1 INTRODUCTION

The rapid increase in the use of PLMs in high-stakes decisions, such as healthcare, criminal justice and finance, have potentially profound societal implications (Rudin, 2019). Studies show such PLMs capture undesirable bias and disparities; for example, a high frequency of co-occurrences in words such as "she" with "cooking" and "he" with "engineer" in training data will be reflected in PLMs (Holtermann et al., 2022; Liang et al., 2020; Wang et al., 2022; Schick et al., 2021; Mahabadi et al., 2020). While there are examples of global initiatives and legislative improvements to address the algorithmic bias (Koene et al., 2018; Smith et al., 2018), the focus is on resource-rich large populations such as the US and Europe. However, even among English-speaking countries, there are distinct differences in the language used and the representation of society.

This research presents a new perspective of the recently published Meade et al. (2022) on the effectiveness of debiasing techniques for PLMs. We focus on New Zealand (NZ) society and argue that local knowledge and understanding must be incorporated to ensure unbiased algorithms, especially when addressing a resource-restricted society.[1]

Aotearoa New Zealand (NZ) is a bilingual country where English is the most widely used language, and Māori is the indigenous language spoken fluently by a smaller proportion of the population. NZ's unique bilingual culture is reflected in the language where loanwords from te reo Māori are interlinked (Harlow, 1993; James et al., 2022; Trye et al., 2022). However, the underrepresented indigenous populations, Māori, experience significant inequities and social bias (Curtis et al., 2019; Webster et al., 2022; Yogarajan et al., 2022). We give examples of prompts (bolded) where GPT-2 generated continuations (not bolded) contain evident social bias:

Two brown Māori men had been seen in a car near the scene, but were not reported for more than two hours. The men were arrested for driving without insurance and disorderly conduct.

Two white kiwi men. (the kiwis are called the white people by the locals). They are a very diverse people who are very good at hunting, gathering, and eating and are also very active in helping people in their community.

These two examples demonstrate multiple failings of GPT-2 in the context of generating continuations of NZ English text. First, there is the social bias that Māori men will break the law. Moreover, the language model incorrectly fabricates facts about NZ society; driving without insurance is not an offence in NZ, kiwis are not called white people by the locals, and text incorrectly describes kiwis[2] as hunter-gatherers.

[1] We use resource-restricted society to mean an underrepresented society with limited data.

[2] A colloquial term for New Zealander.

Table 1: Sample of target and attribute sets to measure bias (top) and bias attributes (bottom).

Attributes & Targets for Measuring Bias		
	Examples of Benchmark Data	**Examples of NZ Specific Data**
1.	White & Black Female/Male Names	Common Māori[2] and Pakeha[3] names.
2.	Sentences created with white or black names: eg: [Amy] is an engineer, [John] is a person	Sentences which also include Māori names. eg: [Nikau] is a person
3.	Phrases such as black female, white women	'Māori woman', 'Pakeha female'
Bias Attributes for Debiasing (category: Race)		
4.	(black, caucasian, asian), (black, white, china), (Africa, America, China), (Africa, Europe, Asia)	(Māori, European, Asian), (Māori, kiwi, Asia), (Hori, white, China), (Hori, Pakeha, Asia), (Pacifica, European, Asian)

2 METHODOLOGY

We consider techniques presented in Meade et al. (2022) for benchmarking new research in bias. Analysing debiasing techniques for PLMs requires two aspects: mitigating against bias and measuring bias.

Debiasing techniques are resource-intensive; they need large datasets and further pre-training (Holtermann et al., 2022). Moreover, post-hoc debiasing techniques such as Self-Debias cannot be evaluated against SEAT (Meade et al., 2022). Hence, we considers Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019; Barikeri et al., 2021), SentenceDebias (Liang et al., 2020) and Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020) for mitigating bias in PLMs (see Appendix A.2 for more details).

Techniques such as StereoSet and Crowdsourced Stereotype Pairs for measuring bias are restricted by using crowd-sourced data from America or requiring hand-crafted attributes and target pairs. For NZ, adopting measures that require a large amount of data is not an option. Hence, we only consider SEAT (May et al., 2019) (see Appendix A.1 more details) with newly curated attributes and target word sets for the NZ population.

2.1 INDIGENOUS PROSPECTIVE

Established pre-defined lists are essential for benchmarking bias measures and techniques. However, to measure and evaluate bias in PLMs, as seen in the example presented earlier, the awareness and inclusion of indigenous perspectives is vital. Using SEAT to measure bias, and CDA, SentenceDebias, and INLP for debiasing, require pre-defined words and/or sentences (Meade et al., 2022). Table 1 provides examples of commonly used pre-defined target and attribute sets, including the angry-black-woman-stereotype, Heilman-double-bind-competent-one-sentence and sent-weat6 (May et al., 2019; Jentzsch et al., 2019; Meade et al., 2022). Table 1 also includes simple examples of NZ-specific additional attributes and target sets. A list of bias attribute words for category race that capture the NZ society is also presented in Table 1.

3 DISCUSSIONS

The social inequities experienced by the underrepresented indigenous population, such as Māori in NZ (Curtis et al., 2019; Webster et al., 2022; Wilson et al., 2022), is significant compared to the non-Indigenous population. As such, developing techniques for bias measures and mitigating bias must be easily adaptable towards any society, and be driven by such societies' knowledge and understanding. Using simple yet effective examples, we have outlined a vital aspect of mitigating the algorithmic bias in this new perspectives paper. Although current techniques provide a platform to understand the bias problem, we cannot solely rely on such methods as proof that the models are unbiased. Ongoing research in mitigating bias will need to move forward from simply relying on pre-defined attribute lists or the need for large datasets.

[2]For Māori names: `https://www.waikato.ac.nz/library/resources/`

[3]A White New Zealander as opposed to a Māori person.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of the ICLR 2023 Tiny Papers Track.

REFERENCES

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1941–1955, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.151. URL https://aclanthology.org/2021.acl-long.151.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Elana Curtis, Rhys Jones, David Tipene-Leach, Curtis Walker, Belinda Loring, Sarah-Jane Paine, and Papaarangi Reid. Why cultural safety rather than cultural competency is required to achieve health equity: a literature review and recommended definition. *International Journal for Equity in Health*, 18(1):1–17, 2019.

Ray Harlow. Lexical expansion in Maori. *The Journal of the Polynesian Society*, 102(1):99–107, 1993.

Carolin Holtermann, Anne Lauscher, and Simone Paolo Ponzetto. Fair and argumentative language modeling for computational argumentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7841–7861, 2022.

Jesin James, Vithya Yogarajan, Isabella Shields, Catherine Watson, Peter Keegan, Keoni Mahelona, and Peter-Lucas Jones. Language models for code-switch detection of te reo Māori and English in a low-resource setting. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 650–660, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.49. URL https://aclanthology.org/2022.findings-naacl.49.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 37–44, 2019.

Ansgar Koene, Liz Dowthwaite, and Suchana Seth. IEEE P7003™ standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the international workshop on software fairness*, pp. 38–41, 2018.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5502–5515, 2020.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8706–8716, 2020.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of NAACL-HLT*, pp. 622–628, 2019.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1878–1898, 2022.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

Adam Leon Smith, Abhik Chaudhuri, Allison Gardner, Linda Gu, Malek Ben Salem, and Maroussia Lévesque. Regulatory frameworks relating to data privacy and algorithmic decision making in the context of emerging standards on algorithmic bias. In *NIPS Conference Workshop on Ethical, Social and Governance Issues in AI, Montreal, Canada, 7th December*, 2018.

D. Trye, V. Yogarajan, J. Konig, T. Keegan, D. Bainbridge, and M. Apperley. A hybrid architecture for labelling bilingual māori-english tweets. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*. Association for Computational Linguistics, 2022.

Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022.

Craig S Webster, Saana Taylor, Courtney Thomas, and Jennifer M Weller. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA Education*, 22 (4):131–137, 2022.

Daniel Wilson, Frith Tweedie, Juliet Rumball-Smith, Kevin Ross, Alex Kazemi, Vince Galvin, Gillian Dobbie, Tim Dare, Pieta Brown, and Judy Blakey. Lessons learned from developing a COVID-19 algorithm governance framework in Aotearoa New Zealand. *Journal of the Royal Society of New Zealand*, 0(0):1–13, 2022. doi: 10.1080/03036758.2022.2121290. URL https://doi.org/10.1080/03036758.2022.2121290.

Vithya Yogarajan, Gillian Dobbie, Sharon Leitch, Te Taka Keegan, Joshua Bensemann, Michael Witbrock, Varsha Asrani, and David Reith. Data and Model Bias in Artificial Intelligence for Healthcare Applications in New Zealand. *Frontiers in Computer Science*, 4, 2022. doi: 10.3389/fcomp.2022.1070493. URL https://www.frontiersin.org/articles/10.3389/fcomp.2022.1070493.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161.

A Appendix

A.1 Sentence Encoder Association Test

Sentence Encoder Association Test (SEAT) (May et al., 2019), an extension to the word embedding association test (WEAT) (Caliskan et al., 2017), is a sentence-level bias measurement technique. In WEAT, two sets of target words T_1 and T_2 and two sets of attribute words A_1 and A_2 are used where the attribute word sets characterise bias and target word sets characterise a particular concept. WEAT measures if a word representation of a given attribute word set is more closely associated with the representations for words from one specific target word set. SEAT extends from WEAT by substituting the attribute words and targets words into a synthetic sentence template (e.g., "that is [WORD]") to create a collection of sentences which can be used to measure associations.

A.2 DEBIAS TECHNIQUES

Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020) is a projection-based debiasing technique that trains a linear classifier to predict the protected property we want to remove from the PLM's representation. The representations are debiased by projecting them into the nullspace of the learnt classifier's weight matrix, effectively removing all of the information the classifier used to predict the protected attribute from the representation. This process can be applied iteratively to debias the representation in PLMs.

SentenceDebias(Liang et al., 2020) extends a word embedding debiasing technique proposed by Bolukbasi et al. (2016) to sentence representations. Sentence-Debias is also a projection-based debiasing technique. By estimating a linear subspace for a particular type of bias, sentence representations are debiased by calculating the difference between the original and resulting projections of the sentence representation.

Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019; Barikeri et al., 2021) is a debiasing technique where a corpus is re-balanced by swapping biased attribute words, e.g. men/women. The re-balanced corpus can be used for further training to debias a model