Project Proposal CSCI 544: Attenuating Bias in Pre-trained Language Models using MABEL

Aditi Bodhankar
Aditya Anulekh Mantri
Advait Rane
Nithyashree Manohar
Surya Teja CVN
{bodhanka, adityaan, aprane, nithyash, suryatej}@usc.edu

1 Domain and Goals

Pre-trained deep learning models encode harmful social biases, which can have unintended consequences in downstream tasks. Language models inherit social biases when trained on text corpora containing examples of such biases. Large Language Models have been shown to encode biases along gender and religious lines. This has prompted recent research to explore ways to de-bias models and produce fair model outcomes. For our project, we aim to implement one such strategy, MABEL (He et al., 2022), to reduce biases in pre-trained language models through a pre-training step.

MABEL describes an intermediate pre-training step to reduce gender bias in language representations. They use Natural Language Inference (NLI) as a pre-training task and augment the NLI dataset to include samples of the same sentence but with opposite genders. For example, if the sentence, "Woman putting together a wooden shelf." is present in the dataset, they add a new sentence, "Man putting together a wooden shelf". The paper restricts itself to the binary gender.

The model is trained on the NLI task along with an alignment loss to minimize the distance between original sentences and their augmented counterparts, and a contrastive loss based on entailment of inference pairs. They then evaluate the model using intrinsic and extrinsic fairness metrics.

The goals for this project are-

1. Implementing and evaluating MABEL to reduce gender bias in language models, as described in the original paper.

2. Evaluating the use of MABEL to reduce social biases along other dimensions, e.g., racial or religious

3. Evaluating alternative strategies to debias models using MABEL-like augmented datasets e.g. by pre-training on a different task.

2 Related Work

2.1 Bias in NLP

Bias in Natural Language Processing is a pressing issue. Deep learning models are susceptible to bias because they are trained on large amounts of data that may contain implicit biases, leading the model to behave negatively towards unrepresented groups. Many deep learning models exhibit occupational gender stereotypes (Sun et al., 2019). Models predict "He is a doctor" with a higher likelihood than "She is a doctor." Word embeddings encode relationships such that 'man' is to 'woman' as 'computer programmer' is to 'homemaker'. Translating "He is a nurse. She is a doctor" into Hungarian and back to English results in "She is a nurse. He is a doctor."

There are other forms of bias prominent in deep learning models. Embeddings for 'Black' is to 'criminal' as 'Caucasian' is to police. Similarly, 'lawful' is to 'Christianity' as 'terrorist' is to 'Islamic'. AI models are more likely to flag tweets written by African Americans as offensive (Manzini et al., 2019). Research shows that multiple gender stereotypes occur in GPT-3 generated narratives, and can emerge even when prompts do not contain explicit gender cues or stereotype-related content (Lucy and Bamman, 2021).

2.2 Mitigating Bias in NLP

Mitigating bias in NLP can be divided into two broad categories - task-specific and task-agnostic debiasing (He et al., 2022). In the first category, the model learns to discard the influence of sensitive attributes during downstream tasks. In the second category, the model mitigates bias by leveraging textual information from general corpora. For example, mitigating gender bias in NLP models can

involve computing a gender subspace representation and eliminating it from the encoded representations (Bolukbasi et al., 2016). Other methods involve re-training the encoder with higher dropout or equalizing objectives. In this work, we propose to explore task-agnostic methods to mitigate gender bias (Gira et al., 2022).

2.3 Evaluating Bias in NLP

Evaluating Bias in NLP models is as critical as identifying and mitigating bias. There are two types of bias evaluation - *intrinsic* and *extrinsic* (He et al., 2022). Intrinsic evaluation methods directly probe the language model by either measuring the geometry of the embedding space or through likelihood scoring. Extrinsic methods, on the other hand, evaluate bias in models by using the model's predictions for a large population of downstream tasks.

Although intrinsic evaluation metrics are opaque they are popular in contemporary works due to their ease of computation. Compared to intrinsic methods, extrinsic methods are much more compute-intensive and time-consuming. However, extrinsic methods are interpretable and are known to better identify bias in language models and flag social harm.

3 Datasets

We intend to follow the MABEL paper, where the experiments are performed on two renowned NLI datasets, which are Stanford Natural Language Inference (SNLI) (Samuel R. Bowman and Manning, 2015) and the Multi-Genre Natural Language Inference (MNLI) (Adina Williams and Bowman, 2018). The MNLI dataset [1] is modelled on the SNLI corpus [2], which gives a range of genres of spoken and written text and supports a distinctive cross-genre generalization evaluation.

We extract the sentence pairs with an entailment relationship as a pre-processing step. This includes a hypothesis sentence that can be inferred to be true based on a premise sentence. Since gender attribute is the prime focus, we extract all entailment pairs that contain at least one gendered term in either the premise or the hypothesis from an NLI dataset. For a sensitive attribute term in the sequence, we swap the word along the opposite bias direction, i.e., by changing "girl" to "boy", while the non-attribute words remain the same. This approach is followed for each sentence in every entailment pair. We also intend to perform a few suitable data-cleaning activities for the application.

[1] https://cims.nyu.edu/~sbowman/multinli/

[2] https://nlp.stanford.edu/projects/snli/

4 Technical Challenge

One of our main challenges is encoding social biases that are easily compounded in downstream tasks. Evaluating bias can be a very dynamic area of recent research and will need further improvement before it can be turned into a measurable outcome. Additionally, we perceive that some parts of this project are reductive, for instance, binary genders. We do not intend to enforce such reductive discourses, and maybe try to expand to a comprehensive solution to the best of our efforts. As a part of our project, we do not intend to find a complete solution to fix biases in models but will proceed to study ways to reduce these biased outcomes.

As for computational needs, the paper employs 4x NVIDIA GeForce RTX 3090 GPUs for implementing MABEL. We will be downgrading this, which can increase the training and inference time. The paper also highlights a concern for the learning rate, α, where our goal will be to optimize the learning rate to balance the trade-off between the fairer stereotype score and the language modeling ability.

5 Individual Contributions

Task	Assignee
Report/Proposal	Everyone
Data Preprocessing	Surya Teja & Aditi
Model Coding	Advait and Aditya Anulekh
Evaluation Metrics	Nithyashree
Analysis of Results	Everyone

References

Nikita Nangia Adina Williams and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.*

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is

to computer programmer as woman is to homemaker? debiasing word embeddings.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Potts Samuel R. Bowman, Gabor Angeli and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.