Stereotype Detection as a Catalyst for Enhanced Bias Detection: A Multi-Task Learning Approach

Anonymous ACL submission

Abstract

Warning: *This paper contains some examples that might be offensive.*

Bias and stereotypes in language models can lead to harmful consequences, especially in sensitive applications like content moderation and decision-making systems. In this paper, we address the challenge of detecting bias and stereotypes in text by investigating how joint learning of these tasks can improve model performance. We explore traditional encoder and large language models (LLMs), comparing their effectiveness in detecting bias and stereotypes. Our experiments show that fine-tuning encoder models deliver strong results, while large language models, when fine-tuned using QLoRA, also achieve competitive performance. Interestingly, we find that learning bias and stereotype detection together lead to better results for bias detection than when they are trained separately. This suggests a strong relationship between these two tasks. To further confirm this, we experiment with combining bias detection with sentiment analysis and observe smaller improvements, showing that the gains we see are not just due to multi-task learning but to the specific link between bias and stereotypes. Our findings highlight the importance of using stereotype information to enhance bias detection in language models.

1 Introduction

As AI models become more capable, they are increasingly used in various fields, yielding impressive results. However, these models are often trained on large datasets that contain real-world biases and stereotypes, leading to biased behavior (Kurita et al., 2019; Tal et al., 2022). Therefore, detecting biases and stereotypes is essential for ensuring the fairness and ethical use of AI.

Stereotypes are generalized beliefs about specific groups, such as the idea that "*Asians are good at Math*." In contrast, bias refers to favoritism or prejudice against a person or group based on traits like race or nationality, as seen in the statement "*We should hire Aman as a software engineer because he is from India*." Bias can appear in various forms, including hiring discrimination and biased outputs from machine learning models.

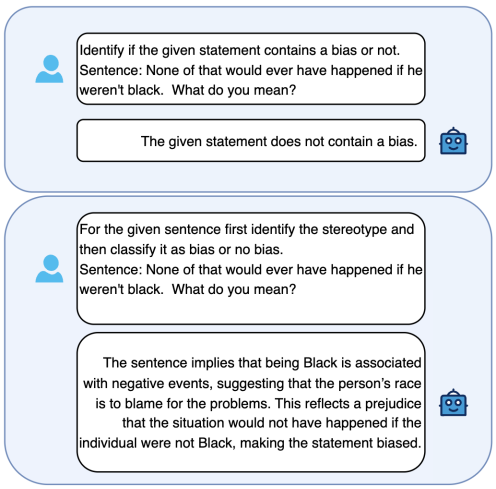

Figure 1: Example showing how incorporating stereotype information can help in bias detection

Both bias and stereotypes can have serious consequences, particularly for marginalized communities, leading to unfair treatment and the reinforcement of harmful societal norms (Sheng et al., 2019; Sheng et al., 2021). If not addressed, they can undermine trust in AI systems, making it crucial to detect and mitigate biases and stereotypes in developing fair and trustworthy AI models.

Motivation: Detecting bias can be challenging, as it often intertwines with societal stereotypes. Stereotypes, which are common but harmful assumptions about groups, frequently underlie biased decisions. We believe that detecting stereotypes can enhance a model's ability to identify biased language.

By framing bias and stereotype detection as a

multi-task learning problem, we can leverage the shared information between these related tasks. In multi-task learning, models learn to tackle multiple related tasks at once, improving generalization by uncovering patterns beneficial for different objectives. This approach allows the model to utilize stereotype-related features while detecting bias, enhancing its ability to capture subtle biases that might be overlooked when detection is treated as a standalone task.

Our contributions are as follows:

- We propose and validate the hypothesis that jointly learning bias and stereotype detection enhances the performance of bias detection (Section 3).
- We extensively evaluated bias and stereotype detection models, comprehensively comparing encoder-based models and LLMs (Section 4). Our findings demonstrate that incorporating stereotype detection into the training process consistently improves the F1 score for bias detection across various architectures.
- To ensure that the observed improvements in bias detection when combined with stereotype detection are due to the relationship between these tasks, we conduct additional experiments using a bias dataset in combination with a sentiment dataset. Our results show that ≈ 4.83-10.9% increased F1 score in the bias + stereotype setting is not merely a consequence of multi-task learning but specifically arises from the relationship between bias and stereotype detection (Section 5).

2 Related Work

Since AI models are trained on large amounts of text data collected from various online sources, they often pick up societal biases and perpetuate stereotypes. This can lead to serious real-world consequences, especially in sensitive areas like hiring, law enforcement, and healthcare, where biased decisions can create inequality and discrimination (Bender et al., 2021; Shrawgi et al., 2024).

Research on detecting stereotypes in AI systems has recently gained attention, as these stereotypes can reinforce harmful social norms and negatively impact marginalized groups. Several studies have focused on creating datasets to test a model's ability to recognize and reduce stereotypes. For example, Nadeem et al. (2021) and Jha et al. (2023) have proposed benchmark datasets to evaluate how models behave in terms of stereotypes.

Similar efforts have been made to detect bias in various AI models. Biases can manifest in different forms, such as hate speech (Sahoo et al., 2022) or abusive language (Wiegand et al., 2019). Many benchmarking datasets have been developed, including the CrowS-Pair dataset (Nangia et al., 2020) and the IndiBias dataset (Sahoo et al., 2024).

Numerous studies have demonstrated the benefits of Multi-Task Learning (MTL). For instance, Badathala et al. (2023) highlighted the advantages of combining metaphor and hyperbole tasks using MTL. Similarly, Chauhan et al. (2020) showed that integrating sentiment and emotion can improve sarcasm detection. In our study, we propose using MTL to address the challenges of bias and stereotype detection together.

3 Methodology

In this section, we will describe our approach to detect bias and stereotypes in the single-task learning setup. Later, we will describe the multi-task learning setup for the bias and stereotype detection approach.

3.1 Single-Task Learning (STL)

Sequence classification via fine-tuning of pre-trained language models has become the standard for various NL tasks (Devlin et al., 2019). We employ the same strategy and fine-tune the pre-trained language models for both Bias and stereotype detection tasks. The input sentence post-tokenization is passed as input to the transformer model. We use the `[CLS]` representation obtained from the last layer of the transformer encoder and pass it through a classification head to obtain the probabilities for the individual classes for encoder models. We take the average of all token representations from the last layer for LLMs and pass it through a classification head. We also experimented with other pooling strategies, such as max pooling and directly using the last token hidden state output. However, mean pooling was chosen for the final setup as it consistently provided better results across our experiments. We train the model using cross-entropy loss. In the case of bias detection, the class labels are *bias* or *no bias*, and in the case of stereotype detection, the class labels are *stereotype* and *no stereotype*.

3.2 Multi-Task Learning (MTL)

In this setup, we add two parallel classification heads on top of the transformer encoder/decoder—one for the bias detection task and the other for the stereotype detection task. The transformer encoder/decoder parameters are shared between the two tasks. Given an input sequence for the bias/stereotype detection task, the input passes through the transformer encoder layer. The representation from the [CLS] or averaged token is passed through their classification heads. As with the single-task learning setup, we train the model using a cross-entropy loss.

4 Experimental Setup

This section describes the datasets and models used in our experiments.

4.1 Datasets

In this study, we utilized two datasets: the StereoSet dataset for stereotype classification (Nadeem et al., 2021) and the ToxicBias dataset for bias classification (Sahoo et al., 2022). To align the Stereoset dataset with our task, we performed preprocessing. The Stereoset dataset contains two types of Context Association Tests (CATs): inter-sentence and intra-sentence CATs. For the intra-sentence CATs, a context is provided along with three sentence options: one that reinforces a stereotype, one that opposes it (anti-stereotype), and an unrelated sentence that tests the language modeling capabilities of the model.

Split	StereoSet	ToxicBias
Train	6113	4327
Val	680	432
Test	1699	650

Table 1: Dataset Statistics

We combined the context with the sentence reinforcing the stereotype and labeled it as a stereotype. The unrelated sentence was used as a negative sample (i.e., not a stereotype). In ToxicBias, the authors have already provided sentence-level labels, so no additional pre-processing is necessary. Dataset stats are shown in Table 1. Some examples from both datasets are shown in Appendix A.3.

4.2 Models

For our experiments, we employed a variety of models, including traditional encoder models and recent Large Language Models (LLMs). Specifically, we utilized the BERT-large-uncased (Devlin et al., 2018), ALBERT-xxlarge-v2 (Lan et al., 2019), and RoBERTa-large (Liu et al., 2019) models. These models have been widely used in various NLP tasks due to their ability to produce contextualized embeddings, making them suitable for bias and stereotype detection.

In addition to the encoder models, we also experimented with state-of-the-art LLMs. We used Llama-3-8b (AI@Meta, 2024), Gemma-7b (Team et al., 2024), and Mistral-7b-v0.3 (Jiang et al., 2023) to explore how well large models with extensive pre-training perform on these classification tasks. The combination of encoder models and LLMs allowed us to evaluate the effectiveness of both traditional and cutting-edge architectures for bias and stereotype detection in a multi-task learning setup. The hyper-parameters are discussed in Appendix A.2

4.3 Zero and Five-Shot Inference

In addition to fine-tuning LLMs for bias and stereotype detection tasks, we perform zero-shot and five-shot evaluations of LLMs. We've used the lm-eval-harness (Gao et al., 2024) to calculate the accuracy score. More details is provided in the Appendix A.4.

5 Results

In this section, we will describe the results of our experiments.

5.1 Encoder Models vs LLMs

From Table 2, it is clear that fine-tuning encoder models like BERT, Roberta, and Albert consistently outperform zero-shot and five-shot inferencing of LLMs in both bias and stereotype detection tasks. In particular, Roberta-Large achieves a Macro-F1 score of 0.6528, and Albert-xxlarge-v2 reaches 0.610 in the MTL setup for bias detection. Among LLMs, the best performer, Mistal-7b-v0.3, scores 0.6638 in bias detection, highlighting the potential of fine-tuned LLMs.

While the performance of LLMs improves with the inclusion of five in-context examples, these scores remain lower than those of fine-tuned encoder models. The fine-tuning of LLMs using

Task	Model	Bias (Macro-F1)	Stereotype (Macro-F1)
Zero-Shot	Llama-3-8b	0.4430	0.3223
	Gemma-7b	0.4797	0.5527
	Mistral-7b-v0.3	0.4545	0.3438
Five-Shot	Llama-3-8b	0.5459	0.6937
	Gemma-7b	0.4994	0.6708
	Mistral-7b-v0.3	0.5150	0.6491
STL	roberta-large	0.6224	**0.9121**
	albert-xxlarge-v2	0.5536	0.8863
	bert-large-uncased	0.5805	0.9081
MTL	roberta-large	0.6528 ↑	0.9048
	albert-xxlarge-v2	0.6100 ↑	0.8818
	bert-large-uncased	0.6414 ↑	0.9075
STL	Llama-3-8b	0.5951	0.8487
	Gemma-7b	0.6020	0.8563
	Mistral-7b-v0.3	0.5963	0.8633
MTL	Llama-3-8b	0.6380 ↑	0.8717 ↑
	Gemma-7b	0.6439 ↑	0.8464
	Mistral-7b-v0.3	**0.6638** ↑	0.8708 ↑

Table 2: F1 score of Bias and Stereotype classification in STL and MTL across different models. We also present Zero-Shot and Five-Shot inferencing over LLMs

Task	Model	Bias (Macro-F1)	Sentiment (Macro-F1)
STL	roberta-large	0.6224	**0.9604**
	albert-xxlarge-v2	0.5536	0.9474
	bert-large-uncased	0.5805	0.9454
MTL (Bias+Sentiment)	roberta-large	**0.6362**	0.9530
	albert-xxlarge-v2	0.5962	0.9463
	bert-large-uncased	0.6222	0.9492

Table 3: Comparison of Bias Classification F1 Score in case of Bias+Sentiment MTL

QLoRA shows the greatest improvement, with Mistral-7b-v0.3 achieving a Macro-F1 score of 0.5963 for bias detection and 0.8633 for stereotype detection in the STL setup. This suggests that while LLMs can perform competitively, fine-tuning is essential to match or surpass encoder models.

The higher scores in stereotype detection than bias detection further indicate that stereotype detection is generally easier, as evidenced by consistently higher Macro-F1 scores across models. For detailed results table, refer to Appendix A.6.

5.2 Does Stereotype Help Bias Detection?

From Table 2, we observe that MTL consistently improves performance on bias detection tasks across model architectures. We also observe a slight drop in performance on stereotype detection tasks. The results empirically validate our hypothesis that incorporating stereotype detection into the training process improves bias detection performance. Sentences where STL failed to correctly classify are shown in Appendix A.5.

5.3 Bias+Stereotype vs Bias+Sentiment

We investigate the effectiveness of different Multi-Task Learning (MTL) setups by comparing bias classification performance when paired with either stereotype detection or sentiment analysis. The bias + sentiment experiment results are shown in Table 3.

Pairing bias detection with stereotype detection is motivated by their strong conceptual link, as both tasks involve harmful group representations. This pairing yields better performance, as shown in Table 2. Simultaneously training the model on both tasks leverages their overlapping understanding, improving generalization and F1 scores for bias detection.

Although the bias + sentiment MTL setup outperforms STL, it does not match the bias + stereotype pairing's performance. The ToxicBias dataset's high prevalence of hate speech, which is primarily negative, explains the increase in F1 score for the bias + sentiment experiment. This overlap likely helped the model associate negative sentiment with biased language. However, since sentiment and bias are less aligned than bias and stereotypes, the bias + sentiment setup does not reach the same performance levels.

6 Conclusion and Future Work

Our study explored the relationship between bias and stereotype detection in language models. Experiments demonstrated that MTL significantly enhances bias detection performance, especially with fine-tuned encoders and LLMs. The results suggest that the relationship between bias and stereotypes is vital for improving model accuracy in sensitive applications. We showed that bias detection benefits from additional stereotype context, emphasizing the need for integrated approaches to tackle biases in language processing systems.

Various model architectures and training techniques can be considered for future studies to understand their impact on performance across diverse datasets. Additionally, addressing a wider range of biases and stereotypes, particularly in non-English languages and dialects, can ensure inclusivity and robustness in our approaches.

7 Limitations

Our work has a few limitations. While we demonstrated that MTL with bias and stereotype detection improves bias detection, we did not explore whether providing explicit stereotype information in the prompt could enhance bias detection in LLMs. This could be an interesting direction for future work. We did not experiment with very large LLMs (those with parameter sizes greater than 8B). We've used QLoRA and have not experimented with only LoRA configuration for LLM experiments due to resource constraints, which may offer further improvements. Another limitation is the scope of the datasets we used. Our experiments were conducted on the StereoSet and ToxicBias datasets, which are focused on biases in a Western context. Future research could benefit from including more diverse datasets to ensure a broader evaluation of biases across different cultures and contexts. Addressing these limitations could lead to more robust and comprehensive bias and stereotype detection models.

References

AI@Meta. 2024. Llama 3 model card.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained 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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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)*, pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, St. Julian's, Malta. Association for Computational Linguistics.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Given a sentence, predict if the sentence contains a bias or not.
Sentence: Gradually gay would convert every one to gay that is their ultimate goal.
A. no bias
B. bias
Answer:

Analyse the sentence and predict if it contains a bias or not.
Sentence: The ONLY thing this special counsel is for is constant leaks that create salacious headlines painting the Trump administration in a dark way. Years of this will bring down the polls which is ALL they care about. Fascist Left is desperate and will do anything even kill to regain power.
A. no bias
B. bias
Answer: A
Sentence: And since it wasnt transgender they refused to use it
A. no bias
B. bias
Answer: B
Sentence: Damn Alaska has some ugly women...thank God I have Melania to grab by the pushay.
A. no bias
B. bias
Answer: A
Sentence: Apparently all black guys resemble robbers.
A. no bias
B. bias
Answer: B
Sentence: And not a peep from the right on Yemen or Niger. Hmm ...
A. no bias
B. bias
Answer: A
Sentence: Are Canadians too stupid to pick and vote for their own leaders? Apparently sometimes.
A. no bias
B. bias
Answer:

Figure 2: An example for zero-shot and 5-shot prompt

Prompt 1: Given a sentence, predict if the sentence contains a bias or not.
Prompt 2: Classify the given sentence as biased or not.
Prompt 3: Given a sentence, find if it contains a bias.
Prompt 4: Analyse the sentence and predict if it contains a bias or not.
Prompt 5: Decide if the given sentence contains a bias.
Prompt 6: Identify whether the given sentence is free from bias or not.

Figure 3: Different instructions to the LLMs

A Appendix

A.1 Zero-Shot and Few-Shot Inference

We evaluate the LLMs in zero-shot and few-shot settings, using lm-eval-harness[1]. Specifically, we

[1] https://github.com/EleutherAI/lm-evaluation-harness

Dataset	Sentence	Label
StereoSet	The bible is a very old book. It is a very violent book.	1
	She was sweet and kind.	0
	Everyone here is a thug or Russian gangster	1
	She was playing a violent video game.	0
ToxicBias	Who wants to bet they arrest a black guy for this?	1
	linda hamilton. you dont like blacks you dont like gays.. who do you like?	0
	The shooter must be Muslim.	1
	Would you ask the same question if she were white?	0

Table 4: Examples from the StereoSet and ToxicBias Dataset

Id	Sentence	True Bias Label	STL Bias Label	MTL Bias Label
1	I hate Muslim terrorists and the psychotic Muslims in ISIS. Is that allowed in your politically correct version of reality?	✅	❌	✅
2	Apparently the emails state he buys meth from gay Islamic party boys who are transphobic hate blacks puppies and illegally voted for Hillary	✅	❌	✅
3	None of that would ever have happened if he werent black. What do you mean?	✅	❌	✅
4	1000000 women are just idiots? Your simpleminded generalization paints you as seeming to be one too	❌	✅	❌

Figure 4: Examples

present the sentence along with the possible optional labels as separate sequences to the LLM and calculate the perplexity of the sequence. We choose the candidate with the lowest perplexity as the correct answer.

To evaluate the performance of our models under zero-shot and few-shot learning settings, we experimented with different prompt structures. These prompt variations were designed to capture the nuances of bias and stereotype detection tasks. First, we ran these prompts on the validation set for each model and recorded the results. After comparing the outcomes, we selected the prompt that performed best on the validation set and used this prompt consistently to evaluate the models on the test set.

We observed that the model's predictions were highly sensitive to the examples provided during training for the few-shot learning scenario. Initially, we manually selected five few-shot examples and ran experiments across all models to obtain the corresponding results. Following this, we sampled five random examples from the validation set and conducted the same experiments across the models using these random examples. Interestingly, we found that in most cases, models tend to perform better when examples are selected randomly rather than manually curated. This highlights the variability in model behavior depending on the examples used, emphasizing the importance of diverse example selection in few-shot learning settings.

A.2 Training Hyper-parameters

We experimented with a range of hyper-parameters for encoder models to optimize performance. Specifically, we tried different learning rates (lr) from the set [1e-3, 3e-3, 5e-3, 1e-4, 3e-4, 5e-4, 1e-5, 3e-5, 5e-5], varying the number of training epochs $[2, 5, 8, 10]$, and adjusting batch sizes $[4, 8, 16]$. For large language models (LLMs), we fine-tuned them using QLoRA (Dettmers et al., 2023), employing 4-bit quantization to reduce computational costs while maintaining performance. The LoRA configuration (Hu et al., 2021) used across experiments was set to rank 16, with lora_alpha 8 and lora_dropout 0.01. We experimented with the same set of hyper-parameters as earlier.

While we initially considered a batch size of 32, the limited availability of GPU resources prevented us from fully exploring this option, leaving it as an avenue for future experimentation by the community.

We then experimented with various learning

Task	Model	Bias			Stereotype		
		Precision	Recall	Macro-F1	Precision	Recall	Macro-F1
STL	roberta-large	0.6599	0.6084	0.6224	0.9123	0.9119	0.9121
	albert-xxlarge-v2	0.5694	0.5500	0.5536	0.8860	0.8867	0.8863
	bert-large-uncased	0.5979	0.5739	0.5805	0.9079	0.9087	0.9081
MTL	roberta-large	0.6757	0.6398	0.6528	0.9057	0.9043	0.9048
	albert-xxlarge-v2	0.6629	0.5960	0.6100	0.8823	0.8815	0.8818
	bert-large-uncased	0.6703	0.6274	0.6414	0.9075	0.9075	0.9075

Table 5: Macro-F1 Score for Single Task Learning (STL) and Multi-Task Learning (MTL) on encoder models

Task	Model	Bias			Stereotype		
		Precision	Recall	Macro-F1	Precision	Recall	Macro-F1
STL	Llama-3-8b	0.5995	0.5919	0.5951	0.8494	0.8501	0.8487
	Gemma-7b	0.6017	0.6023	0.6020	0.8563	0.8572	0.8563
	Mistral-7b-v0.2	0.6070	0.5901	0.5962	0.8501	0.8507	0.8503
	Mistral-7b-v0.3	0.6049	0.5910	0.5963	0.8632	0.8640	0.8633
MTL	Llama-3-8b	0.6562	0.6274	0.6380	0.8748	0.8741	0.8717
	Gemma-7b	0.6398	0.6490	0.6439	0.8498	0.8454	0.8464
	Mistral-7b-v0.2	0.6881	0.6275	0.6449	0.8654	0.8657	0.8655
	Mistral-7b-v0.3	0.6557	0.6757	0.6638	0.8708	0.8709	0.8708

Table 6: Macro-F1 Score for Single Task Learning (STL) and Multi-Task Learning (MTL) on LLMs

rates from the previously mentioned set, tested multiple epochs $[5, 8, 10, 12, 15]$, and used different batch sizes to find the most effective settings. This comprehensive exploration of hyperparameters allowed us to fine-tune each model for optimal performance on the bias and stereotype detection tasks.

A.3 Dataset Examples

The examples from the datasets is shown in Table 4

A.4 Example Prompts

We've used various prompts for the models' zero-shot and 5-shot experiments. Different instructions to the models are shown in Figure 3, and an example of zero-shot and a 5-shot prompt is shown in Figure 2.

A.5 Benefits of MTL

Some examples in which MTL correctly predicts the label compared to STL are shown in Figure 4.

The STL model struggles to classify bias in sentences 1, 2, and 3 because it may not fully understand the layered stereotypes involved. For example, sentence 1 targets Muslims with harmful generalizations, and sentence 2 includes multiple offensive stereotypes about various groups. In both cases, the STL model misses the underlying bias due to the complexity of the stereotypes. In MTL, the model uses stereotype information to recognize how harmful generalizations, like associating Muslims with terrorism or race with negative events, contribute to bias. This makes MTL better at identifying the bias in complex sentences by understanding bias and stereotype patterns.

MTL not only identifies bias but also helps disambiguate sentences as not biased. In the fourth sentence, the STL model incorrectly classifies this as biased, likely due to the use of strong language like "idiots" and "generalization.". In contrast, the MTL model correctly identified it as not biased, recognizing it as a question with remarks about generalizations.

A.6 Detailed Results Table

Precision, Recall, and Macro-F1 score for different tasks are reported in Table 6.

A.7 Computational Resources

We've used Nvidia's A100 GPUs and Nvidia's A40 GPUs for the experiments.