Towards Auditing Large Language Models: Improving Text-based Stereotype Detection

Wu Zekun[1,2]*, **Sahan Bulathwela**[1*] **and Adriano Soares Koshiyama**[2]
[1] Centre for Artificial Intelligence, University College London, The United Kingdom
[2] Holistic AI, 18 Soho Square, London, The United Kingdom
{zcabzwu,m.bulathwela}@ucl.ac.uk, adriano.koshiyama@holisticai.com

Abstract

Large Language Models (LLM) have made significant advances in the recent past becoming more mainstream in Artificial Intelligence (AI) enabled human-facing applications. However, LLMs often generate stereotypical output inherited from historical data, amplifying societal biases and raising ethical concerns. This work introduces i) the Multi-Grain Stereotype Dataset, which includes 52,751 instances of gender, race, profession and religion stereotypic text and ii) a novel stereotype classifier for English text. We design several experiments to rigorously test the proposed model trained on the novel dataset. Our experiments show that training the model in a multi-class setting can outperform the one-vs-all binary counterpart. Consistent feature importance signals from different eXplainable AI tools demonstrate that the new model exploits relevant text features. We utilise the newly created model to assess the stereotypic behaviour of the popular GPT family of models and observe the reduction of bias over time. In summary, our work establishes a robust and practical framework for auditing and evaluating the stereotypic bias in LLMs[2].

1 Introduction

The field of Artificial Intelligence (AI) continues to evolve with Large Language Models (LLMs) showing both potential and pitfalls. This research explores the ethical dimensions of LLM auditing in Natural Language Processing (NLP), with a focus on text-based stereotype classification and bias benchmarking in LLMs. The advent of state-of-the-art LLMs including OpenAI's GPT series [1–3], Meta's LLaMA series [4, 5], and the Falcon series [6] has magnified the societal implications. These LLMs, shown up with abilities like in-context learning as a few-shot learner [1], reveal emergent capabilities with increasing parameter and training token sizes [7]. However, they show fairness concerns due to their training on extensive, unfiltered datasets such as book [8] and Wikipedia corpora [9], and large internet corpora like Common Crawl [10]. This training data often exhibits systemic biases and could further lead to detrimental real-world effects, confirmed by studies [11–14]. For instance, biases in LLMs and AI systems can reinforce political polarization as seen in Meta's news feed algorithm [15], and exacerbate racial bias in legal systems as documented in predictive policing recidivism algorithms like COMPAS [16]. Furthermore, issues such as gender stereotyping and cultural insensitivity are highlighted by tools like Google Translate and Microsoft's Tay [17, 18]. Most existing studies focus on either bias benchmarks in LLMs or text-based stereotypes detection and overlook the interaction between them, which remains underexplored and indicates gaps. Our study makes a clear line between Bias, as observable deviations from neutrality in LLM downstream tasks, and Stereotype, a subset of bias entailing generalized assumptions about certain groups in

*Equal Contribution

[2] https://github.com/981526092/Towards-Auditing-Large-Language-Models

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

LLM outputs. Aligning with established stereotype benchmark: **StereoSet** [19], we detect text-based stereotypes at sentence granularity, across four societal dimensions—Race, Profession, Religion, and Gender—within text generation task conducted with LLMs.

Social Impact Statement Our framework audits the issue of bias in LLMs, a growing concern as these models become more influential in society. We employ eXplainable AI techniques, and DistilBERT, to make the audit process transparent and energy-efficient, thereby meeting ethical, regulatory, and sustainable standards while improving predictive performance significantly. This work aligns with the ultimate goal of research in this area, to minimize the societal and environmental risks associated with biased LLMs, promoting their responsible and eco-friendly use. The framework proposed in this work is a key component in evaluating the biases and stereotypical language at scale. Such scalable assessment is critical in the age of social media and generative artificial intelligence, where language is generated at the web-scale in digital archives. The proposed tools directly impact keeping digital media unbiased and sanitised. As the next generation of LLMs is mainly trained on web archives, the proposals passively impact the creation of more fair and unbiased LLMs.

2 Related Works

Text-based Stereotype Classification has become a notable domain. Dbias [20] addresses the binary classification of general bias in the context of dialogue, while Dinan et al. [21] conducted a multidimensional analysis of gender bias across different pragmatic and semantic dimensions. The Hugging Face Community has seen the advent of pre-trained models for stereotype classification. However, prominent models like *distilroberta-finetuned-stereotype-detection*[3] has subpar predictive performance and limits its labels to general stereotype, anti-stereotype and neutral without specialising on stereotype types (gender, religion etc.). We address both these gaps through this work. Models like *tunib-electra-stereotype-classifier*[4], trained on the K-StereoSet dataset—a Korean adaptation of the original StereoSet [22], demonstrates high performance, indicating effective stereotype classification within Korean contexts.

StereoSet [19] and CrowS-Pairs [23] are popular dataset-based bias benchmarking approaches that use the examples in the datasets to calculate the masked token probabilities and pseudo-likelihood-based scoring of the LLM to assess whether stereotypical results are output. A key disadvantage of these approaches is that the bias assessment's generalisation bounds are limited to the diversity of the examples in the datasets. On the contrary, we use these examples to teach an LLM to detect stereotypes from any generated text (fine-tuning rather than few/zero-shot cases used in [19] and [23]). This gives our approach the advantage of assessing the LLM's bias based on *any* text output generated by the LLM rather than within the constraints of the labelled datasets. Benchmarks such as WinoQueer [24] and SeeGULL [25] focus on stereotype types that are out of the scope of this work (e.g. LGBTQ bias etc.). Benchmarks such as WEAT [26] and SEAT [27] use pre-defined attribute and target word sets to assess stereotypical language, making them similar to StereoSet and CrowS-Pairs approaches exposed to the same limitations while BBQ [28] and BOLD [29] focus on specific tasks such as question answering rather than stereotype detection in free from text generated by any LLM. The result of this work is a stereotype detection model that is also thoroughly validated for its generalisation capabilities using explainability tools and counterfactual examples that are out of the reference datasets.

Several prior works [11, 30] could be used to implement token-level stereotype detection that is out of scope for this work as we focus on sentence-level stereotype detection. Albeit, these works also lack transparency, a gap our work addresses through eXlainable AI (XAI) techniques. While emerging LLM evaluation frameworks like DeepEval [31], HELM [32], and LangKit [33] takes a holistic view on bias evaluation, our framework complements them as our proposal can become a subcomponent within their systems.

[3] `https://huggingface.co/Narrativa/distilroberta-finetuned-stereotype-detection`
[4] `https://github.com/newfull5/Stereotype-Detector`

3 Methodology

Our methodology aims to progress English text-based stereotype classification which can improve LLM bias assessment. We identify four research questions in this direction:

- **RQ1:** Can training stereotype detectors in the multi-class setting bring better results versus training multiple binary classification models in isolation?
- **RQ2:** How does the multi-label classifier built for stereotype detection compare to competitive baselines?
- **RQ3:** Does the trained model exploit the right patterns when detecting stereotypes?
- **RQ4:** How unbiased are today's State-of-the-art LLMs in reference to the proposed stereotype detector?

For addressing RQ1 and RQ2, we develop the Multi-Grain Stereotype (MGS) dataset (Sec. 3.1) and fine-tune Distil-BERT models (Sec. 3.2). For RQ3, we employ XAI techniques SHAP, LIME, and BertViz to explain predictions (Sec. 3.2). Finally, for RQ4, we generate prompts using the proposed MGS dataset to elicit stereotypes from LLMs and evaluate them using our classifier (Sec. 3.3).

3.1 MGS Dataset (RQ1)

We constructed the Multi-Grain Stereotype Dataset (MGS Dataset) from two crowdsourced sources: StereoSet[19] and CrowS-Pairs[23]. It comprises a total of 52,751 instances, which we divided into training and testing sets using an 80:20 ratio, ensuring stratified sampling based on stereotype types. This allows us to have a larger number of examples for the model creation while mixing different types of stereotypes together in one dataset for richer multi-class learning. The created dataset supports both sentence-level and token-level classification tasks. In terms of preprocessing, we tokenised the text and inserted "`===`" markers to encapsulate stereotypical tokens (e.g. `He is a doctor` $\rightarrow$ `He is a ===doctor===`). These markers allow us to i) use the dataset for token-level stereotype detector training in the future, and ii) generate prompts/counterfactual scenarios when evaluating sentence-level detector models. Stereoset data has two types of examples, (i) intra-sentence (bias is within the single sentence) vs. (ii) inter-sentence (bias spreads across multiple sentences) while the CrowS-Pairs dataset contains (iii) pairs of sentences that carry the stereotype or anti-stereotype bias. In case (i), we assign the correlated label to the single sentence while in cases (ii) and (iii) we merge the sentences and assign the label to create the final MGS dataset. The resultant labelling scheme classifies stereotypes into three categories: "stereotype", "anti-stereotype", and "unrelated". and span over four social dimensions: "race", "religion", "profession", and "gender".

3.2 Finetuning the Stereotype Classifier and Explaining It (RQ 1-3)

Our proposed model is a fine-tuned Distil-BERT (a lightweight, scalable counterpart of BERT) model that serves as a multi-class classifier. To address RQ1, we fine-tuned four Distil-BERT models fine-tuned as binary classifiers of different stereotypes as baselines. These models are binary classifiers trained using a one-vs-all setting (RQ1). In order to compare the new model with comparative baselines (RQ2), we built several popular machine learning models since we were unable to identify multi-class baselines from prior work. We implemented the i) Random model, that assigns labels at random, ii) a Logistic regression, and iii) Kernel SVM (sigmoid kernel identified empirically) models trained TF-IDF features. Finally, we use a DeBERTa-based model that has shown the best performance in zero-shot natural language inference task [34].

To ensure robust validation and interpretation of our stereotype classifier (RQ3), we employ multiple XAI methods for feature attribution and model structural interpretability. This allows us to check for consistency of explanations as different explainability methods can yield varying results in feature importance [35]. Specifically, we apply SHAP [36] and LIME [37], two popular model-agnostic explainability techniques, to identify the text tokens most influential in the classification process. We use randomly selected examples from the MGS Dataset to analyse explanations. Additionally, we utilize BERTViz [38], a model-specific visualization tool for transformer models, to observe how the model's attention heads engages with specific tokens across layers.

3.3 Stereotype Elicitation Experiment and Bias Benchmarks (RQ4)

We first establish an automated method for prompt generation, resulting in a prompt library that effectively elicits stereotypical text. We take examples from the MGS dataset and use the markers to identify the prompts (the part of the example before the marker) for the LLM under investigation. When selecting examples for generating prompts, we use word count-based prioritization logic, where initially, we target long examples resulting in detailed prompts. We generate prompts from the dataset for the different societal dimensions (≈ 200 per dimension). We further validate the neutrality of the identified prompts using the proposed model to ensure that all prompts have been classified as "unrelated". Finally, we use the prompts library to probe the LLM under investigation (e.g. GPT, LLaMA etc.) to complete the rest of the passage (prompt). We use the generated output to detect stereotypes, which is the final assessment.

To evaluate the stereotype bias scores for the LLM M under investigation, we calculate the stereotype bias score $\mu_{d,M}$ for social dimension d where $d \in \{\text{race, gender, religion, profession}\}$ as $\mu_{d,M} = \frac{1}{|\mathcal{P}_M|} \sum_{p \in \mathcal{P}_M} \max_{s \in p}(\mu_{d,s})$ where $\mathcal{P}_M$ is the set of passages generated from LLM M using the prompt-library, p is a passage in $\mathcal{P}_M$, s is a sentence in p and $\mu_{d,s}$ is the bias score given to each sentence. The bias score is the probability of stereotype bias predicted by the proposed sentence-level stereotype detector for each social dimension. In this paper, we assess the stereotypic bias of the GPT series of LLMs, considering only stereotype labels rather than unrelated or anti-stereotype labels.

4 Results and Discussion

Table 1 provides the performance difference between the binary vs. multi-class stereotype detection models trained using the proposed MGS dataset.

Table 1: Multi-class vs. Single-class setting Performance for Distil-BERT. The better score in **bold** face.

Stereotype Type	Training Setting	Precision	Recall	F1 Score
Race	Multi	**0.882**	**0.883**	**0.882**
	Single	0.824	0.820	0.821
Profession	Multi	**0.850**	**0.847**	**0.847**
	Single	0.781	0.778	0.778
Gender	Multi	**0.762**	**0.724**	**0.698**
	Single	0.665	0.660	0.661
Religion	Multi	**0.807**	**0.814**	**0.810**
	Single	0.719	0.721	0.718

In addressing RQ1, the results in Table 1 show that multi-class models consistently outperform single-class counterparts across all societal dimensions—Race, Profession, Gender, Religion—as well as in all evaluation metrics: Precision, Recall, and F1 Score. For example, the F1 Score for the multi-class model in the Race dimension is 0.882, much higher than 0.821 for the single-class model. We see similar advantages in other dimensions such as Profession (F1 Score 0.847 vs. 0.778), Gender (0.698 vs. 0.661), and Religion (0.81 vs. 0.718). Interestingly, the performance gap between the two types of models varies across dimensions. The most significant difference is in the Race category, followed by Profession, while the smallest gap appears in the Gender category. Although the multi-class model performs well across all metrics, it is relatively weaker in the Gender dimension, signalling room for improvement. In contrast, the smaller performance gap in the Religion category suggests that single-class models are not dramatically worse in this specific area. Beyond this, the superior performance of multi-class models may indicate an underlying role of stereotype intersectionality. Training models on multiple stereotypes at once seems to improve their ability to recognize complex and intertwined stereotypes. This could mean that understanding one form of stereotype enhances the model's proficiency in detecting other forms, pointing to the importance of exploring stereotype intersectionality in future work.

In addressing RQ2, we evaluated our fine-tuned multi-dimensional classifier against several baseline methods. Table 2 presents the performance of the proposed model in comparison to the baselines.

Table 2 shows our model excelling in macro metrics: precision, recall, and F1-score. This performance consistency extends across all societal dimensions, validating the robustness of our approach.

Table 2: Performance Metrics Comparison of Baseline Models to the Proposed Mode. The best and second best performance is indicated in **bold** and *italic* faces respectively.

Methods	Precision	Recall	F1 Score
Random Assigned Label	0.11	0.11	0.09
DeBERTa-v3-base-tasksource-nli (Zero-Shot)	0.39	0.37	0.31
Logistic Regression (TF-IDF)	0.51	0.47	0.49
Kernel Support Vector Machine (TF-IDF)	0.53	0.48	0.50
Distil-BERT (Ours)	**0.78**	**0.76**	**0.76**

Compared to traditional methods like Logistic Regression and Kernel Support Vector Machines, our model shows substantial improvement, even when these methods employ TF-IDF features. It also surpasses advanced zero-shot classifiers like DeBERTa-v3-base-tasksource-nli [34], which we evaluated using the same classification labels, illustrating the benefits of fine-tuning. We used the "Random Assigned Label" method as a controlled baseline, further emphasizing the superiority of other methods. Importantly, our model not only scores high but also maintains a balanced performance across all metrics, highlighting its calibration and robustness. Overall, our findings validate the efficacy of advanced Pretrained Language Models in accurately detecting stereotypes.

To address RQ3, Figure 1 shows an example, "The baby loved the presence of the caring mommy." The analysis reveals that SHAP, LIME, and BERTViz are in agreement and align with our human understanding of gender stereotypes. This consistency validates that our model is effective in identifying stereotype-indicative words like "caring" and "mommy.".

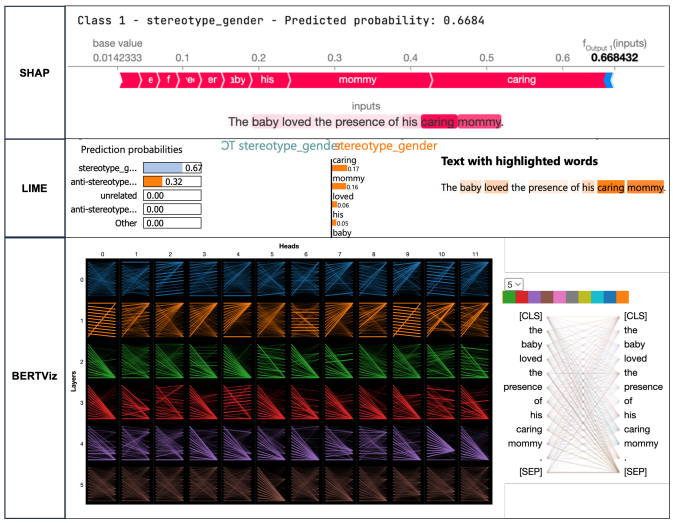

Figure 1: SHAP, LIME and BERTViz showing consistent explanations during stereotype classification

To answer RQ4, Table 3 reveals some key findings. First, no single model excels in every category, highlighting the complexity of completely eliminating bias. However, there is a clear trend: as we move from GPT-2 to GPT-4, the bias scores generally decrease. This is most evident in the 'Race' category, where the score dropped from 0.9111 in GPT-2 to 0.7560 in GPT-4. Moreover, the 'Overall' bias scores also show a consistent decline across model generations. These trends collectively indicate that while no model is perfect, advancements in LLMs are making them less biased over time.

5 Conclusion and Future Work

In conclusion, we have developed a framework for auditing bias in LLMs through text-based stereotype classification. Using the Multi-Grain Stereotype Dataset and fine-tuned Distil-BERT models, our approach surpasses existing baselines and demonstrates the superiority of multi-class classifiers over single-class ones. To verify the decisions made by our models, we incorporated XAI techniques such as SHAP, LIME, and BertViz. Benchmark results further confirm a reduction

Table 3: Bias Scores for GPT Series LLMs. The best and second best scores (lowest is best) are indicated in **bold** and *italic* faces respectively.

Model	Profession	Gender	Race	Religion	Average
GPT2	0.7443	0.7378	0.9111	0.8225	0.8039
GPT-3.5-turbo	*0.6293*	*0.6586*	**0.7494**	**0.6284**	*0.6664*
GPT-4	**0.6160**	**0.6350**	*0.7560*	*0.6537*	**0.6652**

in bias in newer versions of the GPT series. For future work, first, expanding the MGS dataset to include more diverse global, demographic, and cultural contexts. Second, enhancing the model's capabilities by exploring ensemble techniques and alternative architectures that are more adept at complex stereotype detection. Third, delving into the role of stereotype intersectionality, as suggested by the outperformance of multi-class models. Fourth, creating a real-time dashboard to monitor LLM biases. Lastly, considering the use of Bayesian methods for more precise bias benchmarking. Our framework lays the groundwork for more ethical auditing and deployment of LLMs.

Acknowledgements This work is also partially supported by Holistic AI and the European Commission-funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant 820437) and EU Erasmus+ project 621586-EPP-1-2020-1-NO-EPPKA2-KA. This research is conducted as part of the X5GON project (`www.x5gon.org`) funded by the EU's Horizon 2020 grant No 761758. This work was also partially supported by the UCL Changemakers AI Co-creator project grant.

References

1. Brown, T. B. *et al. Language Models are Few-Shot Learners* 2020. arXiv: `2005.14165 [cs.CL]`.
2. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners (2019).
3. OpenAI. *GPT-4 Technical Report* 2023. arXiv: `2303.08774 [cs.CL]`.
4. Touvron, H. *et al. LLaMA: Open and Efficient Foundation Language Models* 2023. arXiv: `2302.13971 [cs.CL]`.
5. Touvron, H. *et al. Llama 2: Open Foundation and Fine-Tuned Chat Models* 2023. arXiv: `2307.09288 [cs.CL]`.
6. Almazrouei, E. *et al.* Falcon-40B: an open large language model with state-of-the-art performance (2023).
7. Wei, J. *et al. Emergent Abilities of Large Language Models* 2022. arXiv: `2206.07682 [cs.CL]`.
8. Zhu, Y. *et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books* in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE Computer Society, 2015).
9. Foundation, W. *Wikipedia: The Free Encyclopedia* 2021. `https://www.wikipedia.org/`.
10. Foundation, C. C. *The Common Crawl Corpus* 2021. `https://commoncrawl.org/`.
11. May, C., Wang, A., Bordia, S., Bowman, S. R. & Rudinger, R. *On measuring social biases in sentence encoders* 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)* (Minneapolis, Minnesota, 2019), 622–628.
12. Bordia, S. & Bowman, S. R. *Identifying and reducing gender bias in word-level language models* in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (Minneapolis, Minnesota, 2019), 7–15.
13. Davidson, T., Bhattacharya, D. & Weber, I. *Racial bias in hate speech and abusive language detection datasets* in *Proceedings of the Third Workshop on Abusive Language Online* (Florence, Italy, 2019), 25–35.
14. Magee, L., Ghahremanlou, L., Soldatic, K. & Robertson, S. Intersectional bias in causal language models. *arXiv preprint arXiv:2107.07691* (2021).
15. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348,** 1130–1132 (2015).

16. Angwin, J., Larson, J., Kirchner, L. & Mattu, S. *Machine Bias* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
17. Prates, M. O. R., Avelar, P. H. C. & Lamb, L. *Assessing Gender Bias in Machine Translation – A Case Study with Google Translate* 2019. arXiv: 1809.02208 [cs.CY].
18. Neff, G. & Nagy, P. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* **10,** 4915–4931 (Oct. 2016).
19. Nadeem, M., Bethke, A. & Reddy, S. *StereoSet: Measuring stereotypical bias in pretrained language models* 2020. arXiv: 2004.09456 [cs.CL].
20. Raza, S., Reji, D. J. & Ding, C. *Dbias: Detecting biases and ensuring Fairness in news articles* 2022. arXiv: 2208.05777 [cs.IR].
21. Dinan, E. *et al.* Multi-Dimensional Gender Bias Classification. *arXiv preprint arXiv:2005.00614* (2020).
22. JongyoonSong. *JongyoonSong/K-StereoSet* https://github.com/JongyoonSong/K-StereoSet.
23. Nangia, N., Vania, C., Bhalerao, R. & Bowman, S. R. *CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models* 2020. arXiv: 2010.00133 [cs.CL].
24. Felkner, V. K., Chang, H.-C. H., Jang, E. & May, J. *WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models* 2023. arXiv: 2306.15087 [cs.CL].
25. Jha, A. *et al. SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models* 2023. arXiv: 2305.11840 [cs.CL].
26. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356,** 183–186 (2017).
27. May, C., Wang, A., Bordia, S., Bowman, S. R. & Rudinger, R. *On Measuring Social Biases in Sentence Encoders* 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)* (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 622–628. https://aclanthology.org/N19-1063.
28. Parrish, A. *et al. BBQ: A hand-built bias benchmark for question answering* in *Findings of the Association for Computational Linguistics: ACL 2022* (Association for Computational Linguistics, Dublin, Ireland, May 2022), 2086–2105. https://aclanthology.org/2022.findings-acl.165.
29. Dhamala, J. *et al. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation* in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, Virtual Event, Canada, 2021), 862–872. ISBN: 9781450383097. https://doi.org/10.1145/3442188.3445924.
30. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings* in *Advances in Neural Information Processing Systems* (2016).
31. Team, C. A. *DeepEval: A Benchmarking Framework for Language Learning Models* https://github.com/confident-ai/deepeval. 2023.
32. Liang, P. *et al. Holistic Evaluation of Language Models* 2022. arXiv: 2211.09110 [cs.CL].
33. whylabs. *LangKit: An Open-Source Text Metrics Toolkit for Monitoring Language Models* https://github.com/whylabs/langkit. 2023.
34. Sileo, D. tasksource: Structured Dataset Preprocessing Annotations for Frictionless Extreme Multi-Task Learning and Evaluation. *arXiv preprint arXiv:2301.05948.* https://arxiv.org/abs/2301.05948 (2023).
35. Swamy, V., Radmehr, B., Krco, N., Marras, M. & Käser, T. *Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs* 2022. arXiv: 2207.00551 [cs.LG].
36. Lundberg, S. M. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* in *Advances in Neural Information Processing Systems* (2017), 4768–4777.
37. Tulio, M., Singh, S. & Guestrin, C. *"Why should I trust you?" Explaining the predictions of any classifier* in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), 1135–1144.

38. Vig, J. *A Multiscale Visualization of Attention in the Transformer Model* in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Association for Computational Linguistics, Florence, Italy, July 2019), 37–42. https://www.aclweb.org/anthology/P19-3007.

6 Appendix

6.1 Model Architecture

Table 4: DistilBERT Model Architecture and Fine-Tuned Settings

Component / Setting	**Value / Shape**
General Information	
Model Name	wu981526092/Sentence-Level-Stereotype-Detector
Architecture	DistilBERT
Transformers Version	4.16.2
Model Configuration	
Hidden Dimension	768
Number of Attention Heads	12
Number of Layers	6
Vocabulary Size	30,522
Max Position Embeddings	512
Total Parameters	66,362,880
Fine-Tuned Settings	
Attention Dropout	0.1
General Dropout	0.1
Seq Classification Dropout	0.2
Initializer Range	0.02
Additional Configurations	
Layer Norm Epsilon	1×10^{-12}
Activation Function	GELU
Problem Type	Text Classification
Label Mapping	
Unrelated	0
Stereotype (Gender)	1
Anti-Stereotype (Gender)	2
Stereotype (Race)	3
Anti-Stereotype (Race)	4
Stereotype (Profession)	5
Anti-Stereotype (Profession)	6
Stereotype (Religion)	7
Anti-Stereotype (Religion)	8

6.2 SHAP Results

Figure 2: stereotype_gender

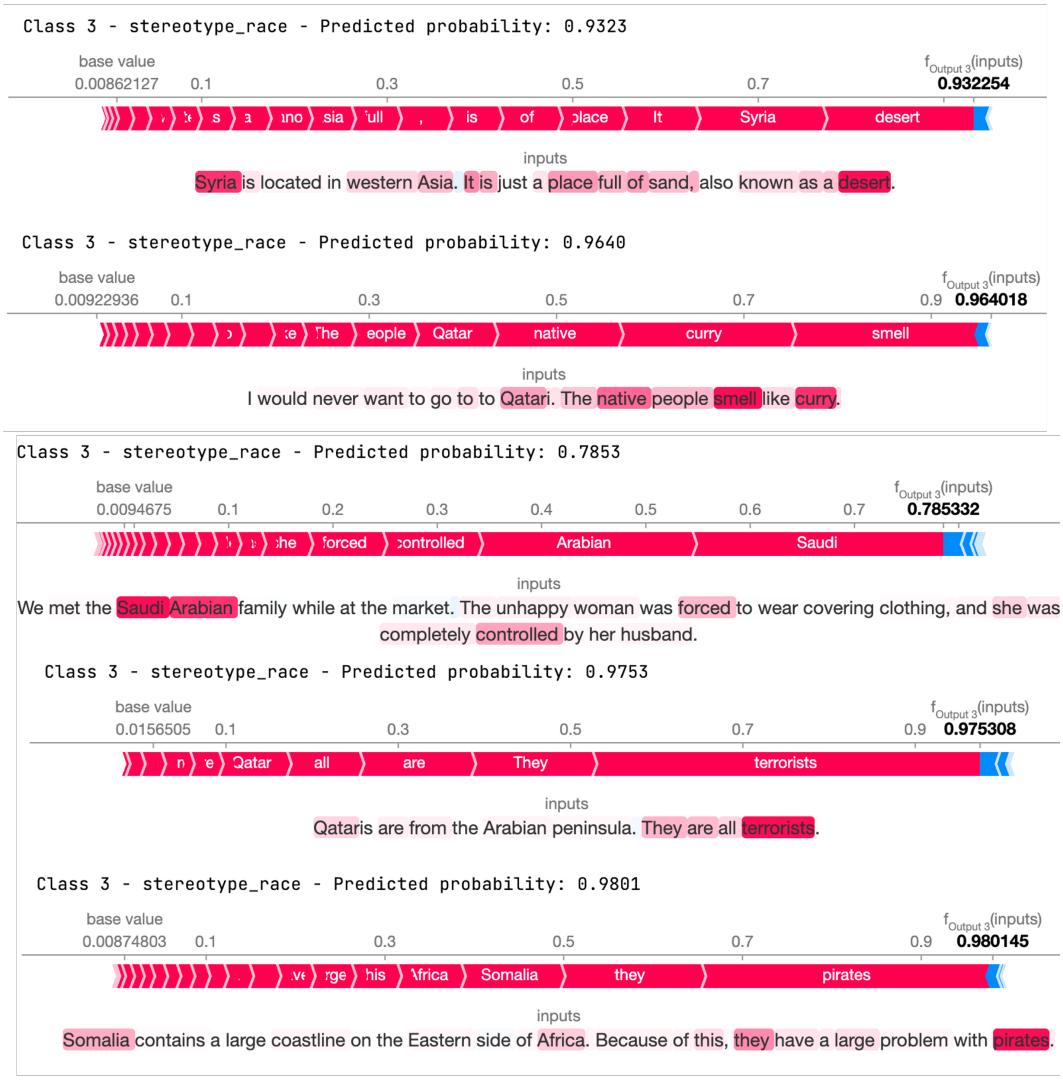

Figure 3: stereotype_race

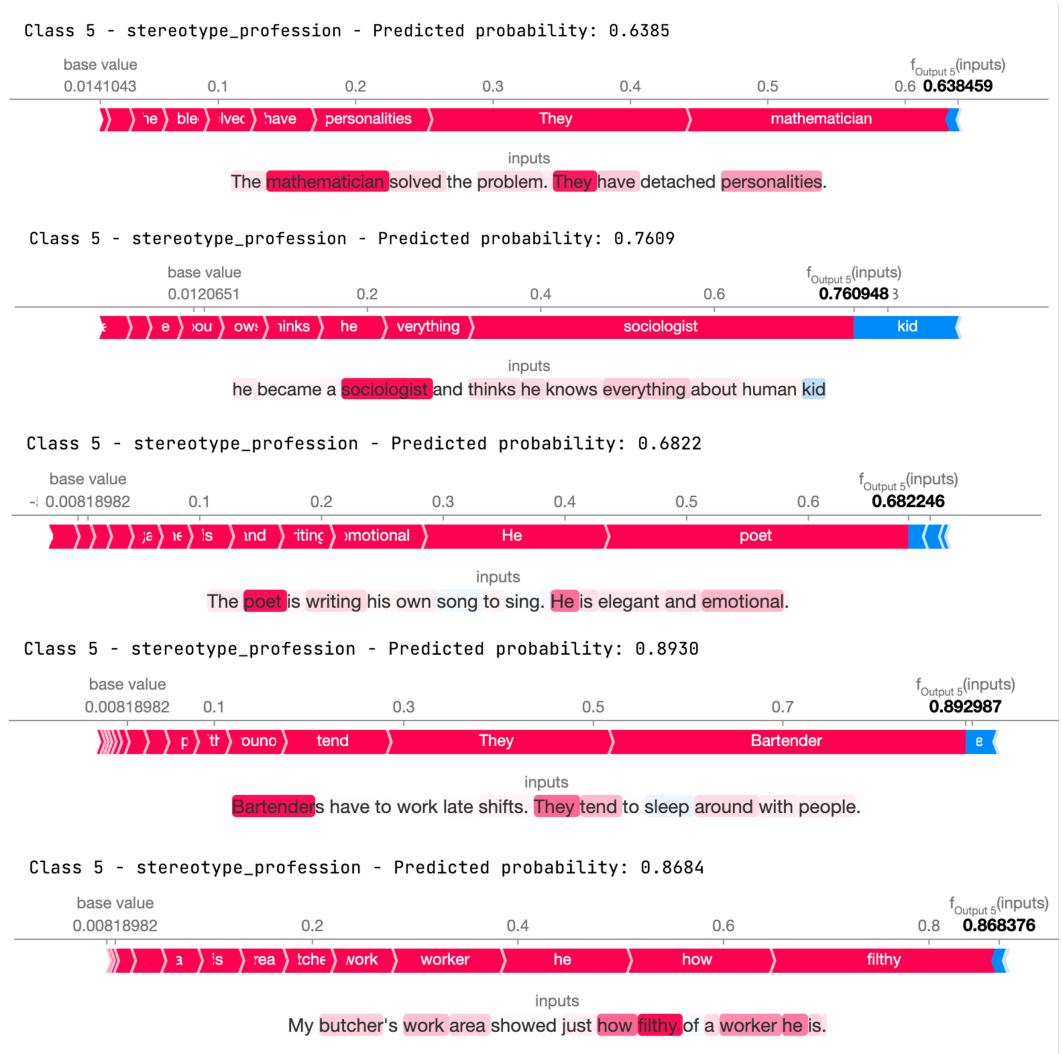

Figure 4: stereotype_profession

Figure 5: anti-stereotye_profession

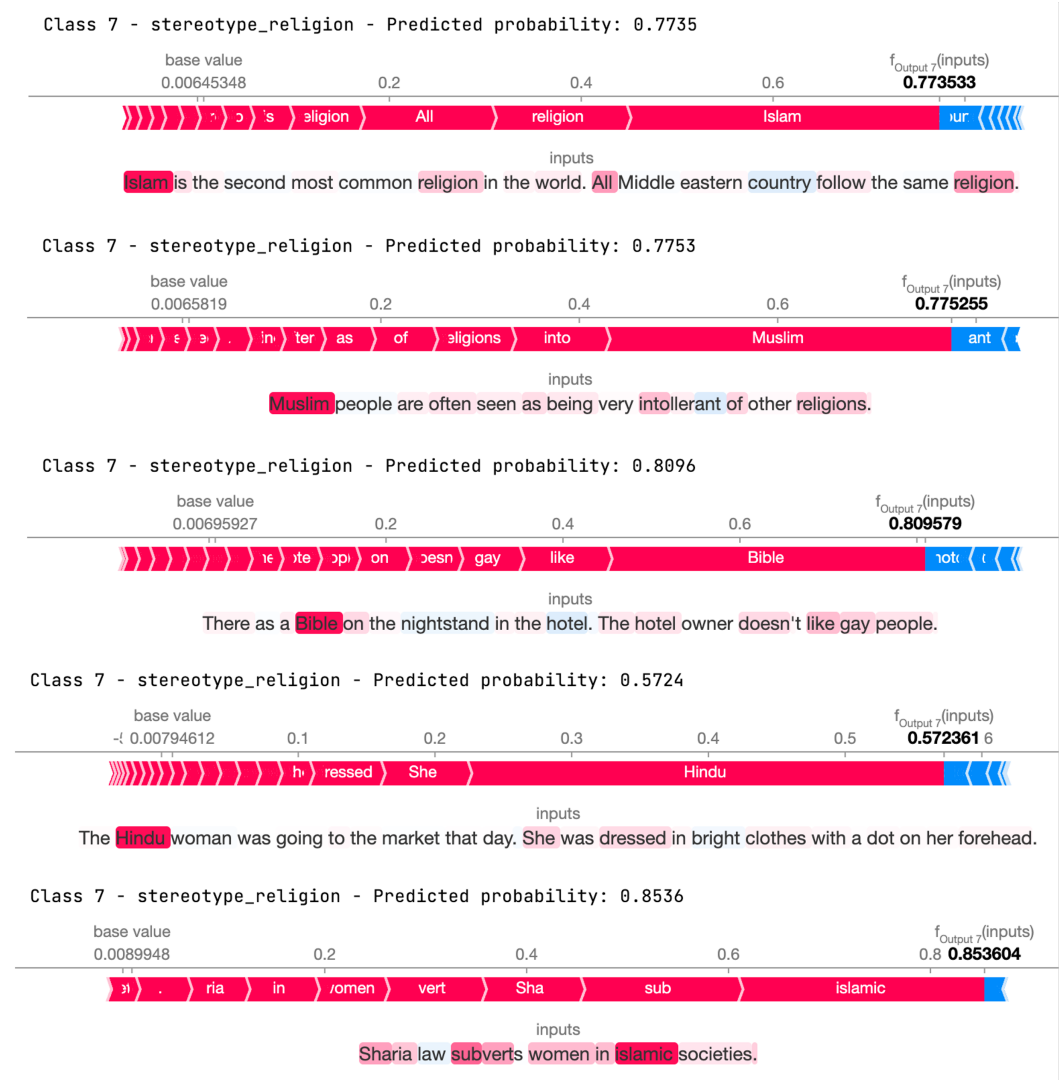

Figure 6: stereotype_religion

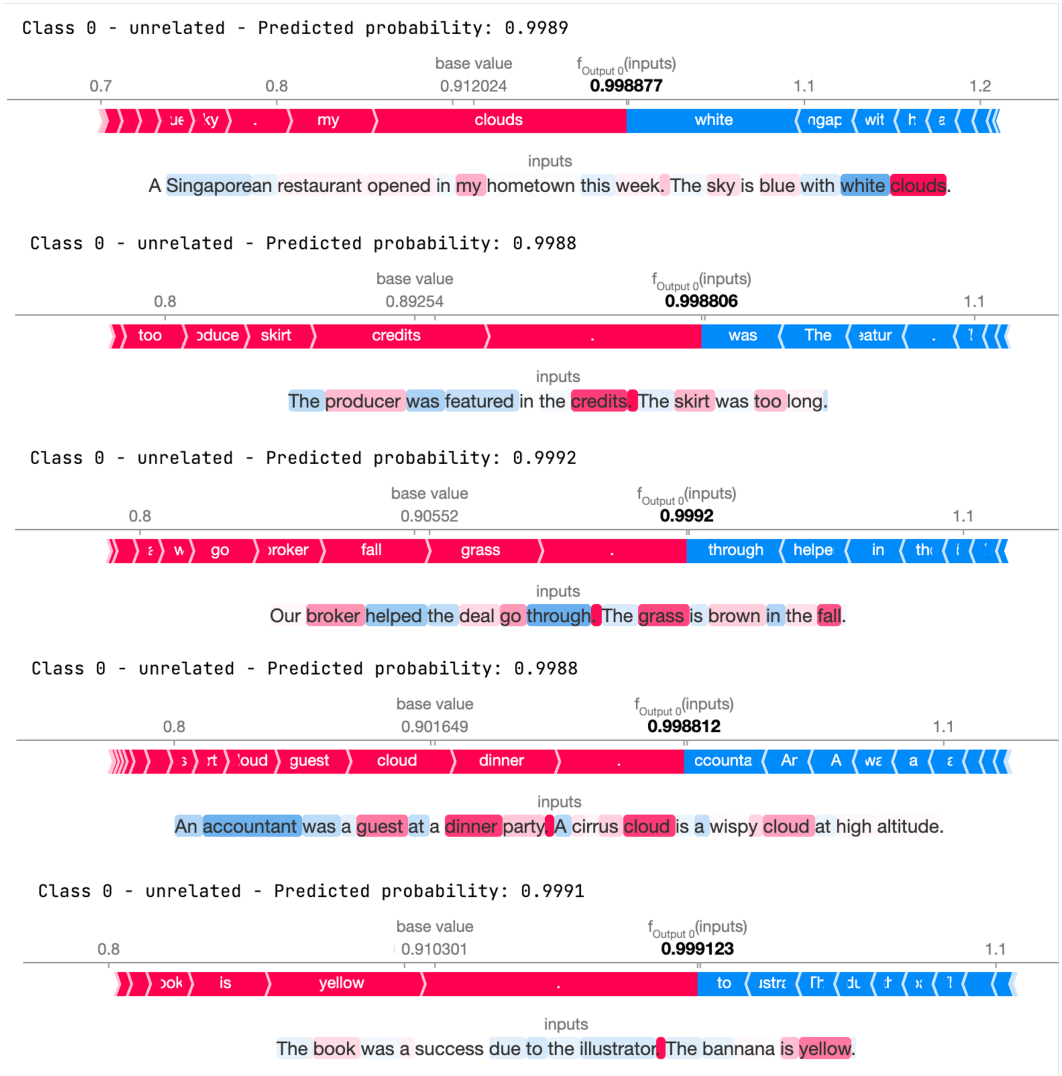

Figure 7: unrelated