Identifying and Measuring Token-Level Sentiment Bias in Pre-trained Language Models with Prompts

Anonymous ACL submission

Abstract

Due to the superior performance, large-scale pre-trained language models (PLMs) have been widely adopted in many aspects of human society. However, we still lack effective tools to understand the potential bias embedded in the black-box models. Recent advances in prompt tuning show the possibility to explore the internal mechanism of the PLMs. In this work, we propose two token-level sentiment tests: Sentiment Association Test (SAT) and Sentiment Shift Test (SST) which utilize the prompt as a probe to detect the latent bias in the PLMs. Our experiments on the collection of sentiment datasets show that both SAT and SST can identify sentiment bias in PLMs and SST is able to quantify the bias. The results also prove that fine-tuning can augment existing bias in PLMs.

1 Introduction

Large-scale pre-trained language models (PLMs), such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT (Radford et al., 2018, 2019; Brown et al., 2020) and T5 (Raffel et al., 2020), have shown competitive performance in many downstream applications in natural language processing. The key to the success of PLMs lies in the unsupervised pre-training on massive unlabeled corpus as well as a large number of parameters in the neural models. While these PLMs have been deployed to a wide variety of products and services such as search engines and chatbots, investigating the fairness of these PLMs has become a growing urgent research agenda.

Recent studies have shown that there are various stereotypical biases related to social factors such as gender (Bhardwaj et al., 2021), race (Iandola et al., 2020), religion (Nadeem et al., 2021), age (Nangia et al., 2020), ethnicity (Groenwold et al., 2020), political identity (McGuffie and Newhouse, 2020), disability (Hutchinson et al., 2020), name (Shwartz et al., 2020) and many more, that are inherited by these PLMs. However, sentiment bias, which characterizes the bias of words towards a particular sentiment polarity, such as *positive*, *negative*, and *neutral*, has not been well studied. Huang et al. (2020) investigated the sentiment bias in texts generated by language models like GPT while overlooking the fact that each individual word may also have sentiment bias in the PLMs.

In this work, we focus on identifying and measuring the sentiment bias of individual words in pre-trained language models. Instead of investigating all the words in the vocabulary, we only select a list of words with confident sentiment polarities from available sentiment lexicons constructed by humans, and design two novel approaches to identify their sentiment bias based on language model prompting: (1) Sentiment Association Test (SAT), where the bias of each word is identified by detecting its association with various positive or negative reviews; (2) Sentiment Shift Test (SST), where the bias of each word is identified by predicting the sentiment polarity shift after appending it multiple times to various sentiment-oriented reviews. Based on these two approaches, we observe that 39.25% out of 400 words considered neutral in the lexicon show a sentiment bias in commonly used PLMs. In addition, by extending the Sentiment Shift Test, we further design a new metric to measure the strength of the sentiment bias for each word. Our contributions are summarized as follows:

- We design two novel sentiment test approaches, SAT and SST, to investigate the token-level sentiment bias from the PLMs, and demonstrate that 39.25% out of 400 neutral words show a sentiment bias in various PLMs.

- We also design a new metric to quantify the sentiment bias of each word by extending our Sentiment Shift Test.

2 Related Work

Stereotypical Bias in Natural Language Processing Nadeem et al. (2021) define bias based on gender, profession, race, and religion, and design formulae to quantify the stereotypical bias along with model meaningfulness of PLMs for sentence-level and discourse-level reasoning. Nangia et al. (2020) define a metric on how likely is the stereotype/anti-stereotype to generate the rest of the sentence. Hutchinson et al. (2020) use the Google Cloud Sentiment model to demonstrate more negative bias in top-k words predicted by BERT when prompted with disability tokens.

Bias in natural language embeddings Many studies explore bias within word embeddings. Bolukbasi et al. (2016) utilize analogy tests and demonstrate that word2vec embeddings reflect gender bias by showing that female names are associated with familial words rather than occupations. Islam et al. (2016) proposed WEAT (Word embedding association test) to show how names can be associated with entities. Zhao et al. (2019) prove that contextualized word embeddings have bias and show how bias propagates to downstream tasks.

Identification and Measurement of Sentiment Bias Huang et al. (2020) propose detection of sentiment bias by varying some sensitive attributes and measuring the sentiment polarity of the generated text using GPT-2. Groenwold et al. (2020) determine sentiment bias for ethnicity by comparing the sentiment of text generated by GPT-2 and find more negative sentiment generated for African American Vernacular English text as compared to Standard American English text. Compared to the above studies, our work investigates and measures the sentiment bias at token level from PLMs.

3 Approach

3.1 Dataset Construction

Our goal is to investigate and measure the sentiment bias of each word in PLMs. Considering that many words may indicate distinct sentiment polarities in different context, we first build a highly confident sentiment lexicon where each word is annotated as *positive*, *negative* or *neutral*. Specifically, we draw strongly positive and negative tokens from the VADER lexicon (Hutto and Gilbert, 2014) where all the words are annotated with sentiment scores from -4 to +4 (-4 being strongly negative and +4 being strongly positive). We draw the neutral words from MPQA opinion corpus (Deng and Wiebe, 2015).We investigate the sentiment bias only on the neutral words, and use the positive and negative words to verify our approaches.

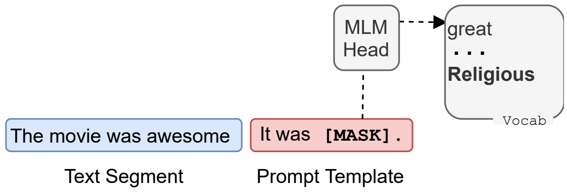

Figure 1: Overview of the prompting approach for SAT.

The sentiment lexicon contains a golden sentiment label of each word. Thus, to detect the sentiment bias, we need to compare the sentiment polarity of each word in PLMs with its golden sentiment label. To predict the sentiment polarity of each word in PLMs, we will leverage a set of sentiment-oriented reviews collected from IMDB (Maas et al., 2011), Amazon Reviews (He and McAuley, 2016), YELP (Asghar, 2016), and SST-2 (Socher et al., 2013). Each review is annotated as positive or negative. We collect 2000 positive reviews and 2000 negative ones. As the reviews span over diverse domains, including movies, food, and products, they can well represent each sentiment polarity.

3.2 Sentiment Bias Identification

Sentiment Association Test Inspired by the Word Embedding Association Test (Islam et al., 2016), we first design a new Sentiment Association Test approach to predict the sentiment polarity of each word in PLMs based on their associations. Our approach is based on the assumption that if a word consistently shows a stronger association to the diverse set of positive (or negative) reviews, it should have a positive (or negative) sentiment polarity. Based on this assumption, we design a language model prompting approach to estimate the association of each word with a review. As Figure 1 shows, given each positive review p_i , we concatenate it with a template-based prompt, "`It was [MASK]`", and feed the whole sequence to a language model encoder. Based on the contextual representation of "`[MASK]`", we predict a probability for each word in the sentiment lexicon as s_{ij}^p where j is the index of the word in the lexicon. Similarly, for each negative review n_k, we apply the same prompt and use the same approach to predict a probability for each word in the lexicon as s_{kj}^n. For each word indexed with j, we determine its sentiment polarity by comparing $mean_i(s_{ij}^p)$ with

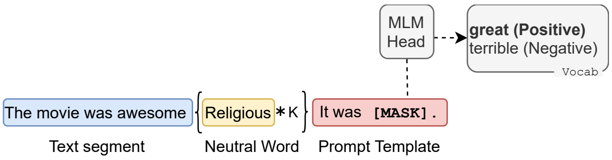

Figure 2: Overview of the prompting approach for SST.

$mean_k(s_{kj}^n) + m * std_k((s_{kj}^n))$ and $mean_k(s_{kj}^n)$ with $mean_i(s_{ij}^p) + m * std_i((s_{ij}^p))$, which denotes the association to positive and negative sentiment polarity, respectively. std denotes the standard deviation, which in this case serves as a dynamic unit to measure the distance of the means between the positive and negative probability mass functions (PMFs) and m shows the strength of the sentiment polarity. With a fixed unit, or no dynamic unit, we can only identify the strongly biased words.

Sentiment Shift Test Another intuitive approach to predict the sentiment bias of each word in PLMs is based on the assumption that if a word is negative in PLMs and appended multiple times to a positive review, it's likely that the sentiment of this new sequence might be shifted to neutral or even negative. Based on this assumption, we further design a new Sentiment Shift Test approach to predict the sentiment bias of each word in PLMs. As Figure 2 shows, given a review, we first apply language model prompting to concatenate the review with a prompt "`It was [MASK]`", and predict a sentiment label by comparing the probability of "*great*" and "*terrible*" based on the contextual representation of "`[MASK]`". Then, for each word in the lexicon, we append it K times to the review, and use the same language model prompting approach to predict a sentiment label. We will predict the sentiment bias of each word in PLMs by analyzing the number of sentiment shifts for all positive or negative reviews, i.e., if a word is appended to positive reviews and reduces the accuracy of the model on reviews, this word will have a negative bias.

3.3 Sentiment Bias Quantification

Based on the Sentiment Shift Test, we further design a new metric to quantify the strength of the sentiment bias of each word. Our motivation is that, the less times that a word is appended and the more sentiment labels are shifted after appending it to the reviews, the stronger that the bias will be. Based on this motivation, we design the following metric: $q = \frac{1}{n}\sum_K \frac{(Neg\_Diff - Pos\_Diff)}{K^2}$ where, $Neg\_Diff = A - A'$ is defined as the change of the sentiment classification accuracy on the negative sentiment set after appending a neutral word K times. A and A' denote the accuracy before and after appending the word, respectively. Similarly, Pos_Diff is defined as the change of the accuracy on the positive sentiment set. If a neutral word reduces the negative accuracy, Neg_Diff will be positive and Pos_Diff is more likely to be negative, providing us with a high overall positive value implying a positive sentiment bias.

4 Experimental Results and Discussion

4.1 Experimental Setup

We select 400 words for each of positive, negative, and neutral categories from the sentiment lexicon and perform SAT and SST on the 2,000 positive and 2,000 negative reviews. The experiments are mainly on RoBERTa models as they show a significantly better understanding of sentiment presented in the text than BERT models. We analyze the word-level sentiment bias in both the pre-trained language model and prompt[1]-based fine-tuning model. The prompt-based model follows the training framework from (Gao et al., 2021) which utilizes a set of training instances as demonstrations to help the model make predictions.

4.2 Does the Probability Predicted by the Language Model Indicate the Sentiment Polarity?

We first investigate whether the pre-trained language models are capable of sensing the sentiment in the text by predicting the probabilities on a set of words with strong sentiment polarity. To do so, we use the mean probabilities of positive and negative words on each positive and negative review. Specifically, we first compute the mean probabilities $mean_j(s_{ij}^p)$ of 400 positive words, and $mean_l(s_{il}^p)$ of 400 negative words. Then, we find their differences $mean_j(s_{ij}^p) - mean_l(s_{il}^p)$ on each the positive review s_i^p and negative review s_{kl}^n. The results on the pre-trained language model are shown in Figure 3. One can observe that most positive reviews (blue line) have positive values, and most negative reviews have negative values. The mean value for positive reviews is 8.2e-3, and the mean value for negative reviews is -1.9e-4. We observe the same trend in the prompt-tuned language model, as shown in Figure 4, except that the fine-tuning improves the performance. The mean

[1]We study the impact of different prompt templates and label words in Appendix A.2

value for a positive review is 1.1e-3 and the mean value for negative reviews is -7.9e-4.

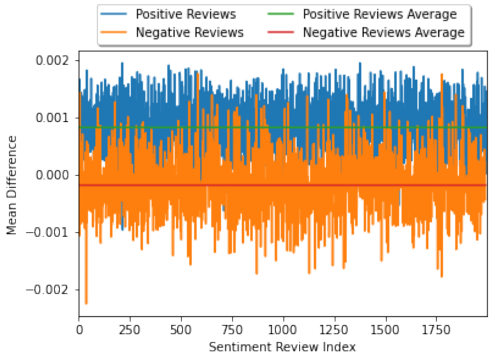

Figure 3: Plot for $mean_j(s^p_{ij}) - mean_l(s^p_{il})$ on positive reviews s^p_i and negative reviews s^n_k for RoBERTa-base.

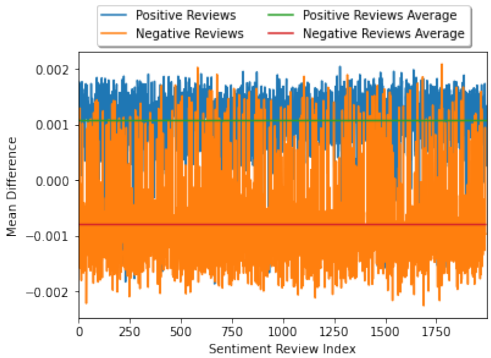

Figure 4: Plot for $mean_j(s^p_{ij}) - mean_l(s^p_{il})$ on positive reviews s^p_i and negative reviews s^n_k for RoBERTa-base-finetuned.

4.3 Sentiment Bias Identification

Sentiment Association Test We perform SAT on the 400 neutral words to identify their potential bias, and show the identified biased words in Table 1 in Appendix A.3. We find that: (1) 41.66% of positive-biased words and 52.7% of negative-biased words are shared by at least two models; (2) The number of negative-biased words is 96.0% higher than the number of positive-biased words; (3) After fine-tuning, the number of biased words drastically increases, 124.6% on RoBERTa-Base and 268% on RoBERTa-Large[2].

Sentiment Shift Test For each of the 400 neutral words from the lexicon, we append it to the reviews for k times where $k \in \{5, 10, 15\}$. Table 2 in Appendix A.3 shows the top-10 most positive and negative-biased words with $k = 5$. The words are ranked by their SST scores. We observe that: (1) The number of identified positive and negative-biased words increase as k increases and the increasing rate decreases as k becomes larger; (2) 70.6% of positive-biased words and 56.8% of negative-biased words are shared by at least two models; (3) For RoBERTa-Large models, 2.25% neutral words simultaneously reduce or increase the accuracy of sentiment classification. We suggest those words are truly neutral. To understand the correlation between SAT and SST, we pick the set of negative and positive-biased words identified by SAT and SST respectively, and find that the two methods share 70% of the negatively biased words and 100% of positively biased words[3].

4.4 Are SST and SAT Effective For Identifying and Measuring Bias?

A large number of overlaps between the identified sentiment-biased words from different models prove there is a shared sentiment trend among them. The large overlaps between SST and SAT show the agreement of the trend identified by two testing methods. Thus, we can claim that the identified trend is a kind of sentiment bias that persists in language models. In addition, we find that the fine-tuning can augment the existing bias in the PLMs as the number of biased words increase in both RoBERTa-Base and RoBERTA-large after prompt-tuning. To understand if the measurement can correctly quantify[4] the sentiment bias, we take the top-50 and bottom-50 words from the negative-biased words from fine-tuned RoBERTa-Base ranked by SST and compare them against all the negative-biased words identified by SAT. We find 76% of the top-50 words agree with the words from SAT and 56% of the bottom-50 words agree with the word from SAT. The much higher agreement rate in the top-50 words ranked by SST proves the effectiveness of the measurement.

5 Conclusion

In this work, we present Sentiment Association Test and Sentiment Shift Test, two prompt-based methods to identify and measure the token level sentiment-bias in PLMs. We perform extensive experiments on collections of positive and negative reviews and prove that there is sentiment bias in PLMs and our proposed tests can identify and quantify the bias.

[2]Averaged on positive and negative biased words.

[3]The number of biased words is the union of the words identified in all models.

[4]The top 10 ranked words with SST scores for RoBERTa-Base and fine-tuned RoBERTa-Base are in Tables 3,4,5,6,7,8.

References

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *CoRR*, abs/1605.05362.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in BERT. *Cogn. Comput.*, 13(4):1008–1018.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1323–1328. The Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating african-american vernacular english in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5877–5883. Association for Computational Linguistics.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517. ACM.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 65–83. Association for Computational Linguistics.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *CoRR*, abs/2005.00813.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.

Forrest N. Iandola, Albert E. Shaw, Ravi Krishna, and Kurt Keutzer. 2020. Squeezebert: What can computer vision teach NLP about efficient neural networks? *CoRR*, abs/2006.11316.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150. The Association for Computer Linguistics.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. *CoRR*, abs/2009.06807.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5356–5371. 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 2020, Online, November 16-20, 2020*, pages 1953–1967. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. *CoRR*, abs/2004.03012.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *CoRR*, abs/1904.03310.

A Appendix

A.1 Implementation Details

We use the pre-trained RoBERTa-Base and RoBERTa-Large models from Huggingface. We use the Adam optimizer with a learning rate of 1e-5 and batch size 2 to train our models. Each epoch takes about 30 mins and we run the experiment on one Tesla P40.

A.2 Impact of Prompt Templates

As the prompt plays an important role in our work, we experiment with different types of templates, such as "`the review was [MASK]`" and "`I [MASK] it.`" but we find that adding more context affects the models' ability to identify bias. The context dominates the prediction. Thus, we decide to use the simple template "`It was [MASK].`" in all of our experiments.

A.3 Experiment Results

Table 1: This table shows the top-10 most biased words identified by SAT on various PLMs. The % column shows the percentage of identified biased words in all the neutral words (400).

Model	Threshold	Positive Words	%	Negative Words	%
RoBERTa-Base	0.5*Standard Deviation	modular,shone,Tours	0.75	obvious,deadly,vanilla, cheese,speculation,systematic, conjecture,sleepy,economics,vodka	17.0
RoBERTa-Base	1*Standard Deviation	shone	0.25	cheese,speculation,turkey, skeletal,outright,bacterial, carrots,tires,dialog,attitudes	3.75
RoBERTa-Base	1.5*Standard Deviation	-	0.0	cheese,speculation,carrots	0.75
Finetuned RoBERTa-Base	0.5*Standard Deviation	PEOPLE,sovereignty, embodiment,predominant, Indeed,incorporate,touch	1.75	vanilla,skeletal,speculation, implicit,cheese,sleepy,overweight, pitched,judgement,conjecture	36.25
Finetuned RoBERTa-Base	1*Standard Deviation	sovereignty	0.25	vanilla,skeletal,speculation, cheese,sleepy,overweight,conjecture, rural,Possible,turkey	24.0
Finetuned RoBERTa-Base	1.5*Standard Deviation	-	0.0	vanilla,skeletal,speculation, cheese,sleepy,overweight,conjecture, rural,Possible,turkey	16.25
RoBERTa-Large	0.5*Standard Deviation	shone	0.25	deadly,screaming,cheese, speculation,overtime,overweight, implicit,systematic,vodka,conjecture	12.5
RoBERTa-Large	1*Standard Deviation	shone	0.25	cheese,speculation,conjecture, bacterial,tires,signals, ander,Minor	2.0
RoBERTa-Large	1.5*Standard Deviation	-	0.0	bacterial	0.25
Finetuned RoBERTa-Large	0.5*Standard Deviation	precious,shone,servings, embodiment,Indeed	1.25	familiar,implicit,screaming, overweight,cheese,sleepy,speculation, bacterial,turkey,conjecture	29.5
Finetuned RoBERTa-Large	1*Standard Deviation	shone	0.25	implicit,overweight,sleepy,speculation, bacterial,conjecture,patriarchal, Possible,vodka,glare	16.5
Finetuned RoBERTa-Large	1.5*Standard Deviation	-	0.0	implicit,overweight,sleepy, speculation,bacterial,turkey,conjecture, patriarchal,Possible,vodka	9.0

Table 2: This table shows the top-10 most biased words identified by SST on various PLMs, with k=5. The % column shows the percentage of identified biased words in all the neutral words (400).

Model	K	Positive Words	%	Negative Words	%
RoBERTa-base	5	shone , dominant , clout , intrigue , globalization , uncover , Saint , Circle , Beans , exercised	57	deadly , judgement , screaming , plight , appeal , bacterial , glare , obligations , obligation , overweight	31
RoBERTa-base	10	anyways , utilizes , shone , dominant , attaches , intrigue , clout , uncover , reflecting , globalization	58	deadly , judgement , screaming , undergoing , glare , appeal , opinions , speculation , plight , referee	34
RoBERTa-base	15	anyways , clout , dominant , intrigue , utilizes , attaches , shone , uncover , incorporate , reflecting	59.2	deadly , judgement , screaming , speculation , undergoing , glare , counselling , referee , subsequently , Possible	34.5
Finetuned RoBERTa-base	5	shone , extensive , Saint , Indeed , systematic , Awareness , concerted , insights , dominant , renewable	44.5	deadly , judgement , overweight , speculation , glare , appeal , patriarchal , tobacco , adversity , notion	17.6
Finetuned RoBERTa-base	10	shone , extensive , Saint , quite , dominant , concerted , renewable , Awareness , insights , Indeed	44	deadly , judgement , overweight , speculation , appeal , glare , tobacco , screaming , speculate , patriarchal	27.5
Finetuned RoBERTa-base	15	extensive , shone , Saint , quite , dominant , insights , incorporate , participants , concerted , entirely	45	judgement , deadly , speculation , overweight , appeal , notified , speculate , counselling , glare , screaming	30
RoBERta-large	5	renewable , exercised , sovereignty , precious , Jordanian , intrigue , modular , comedy , Destiny , Episcopal	63.2	deadly , judgement , appeal , disposition , skeletal , patriarchal , corrective , December , systematic , convict	23.17
RoBERta-large	10	Destiny , precious , renewable , Rapid , sovereignty , Jordanian , Awareness , intrigue , incorporate , Episcopal	60.2	deadly , judgement , disposition , appeal , patriarchal , screaming , Pricing , skeletal , plight , glare	23
RoBERta-large	15	Awareness , Destiny , incorporate , Episcopal , precious , renewable , intrigue , Jordanian , sovereignty , olive	59.5	deadly , judgement , disposition , appeal , screaming , patriarchal , glare , counselling , skeletal , Confederate	22
Finetuned RoBERTa-large	5	familiar , shone , comedy , vanilla , renewable , exercised , consistency , insights , chocolate , convertible	44.5	judgement, patriarchal, assumption, Minor, overweight, conjecture, distance, bacterial, appeal, stall	44.8
Finetuned RoBERTa-large	10	familiar , insights , vanilla , consistency , shone , olive , chocolate , exercised , correctness , silver	48.1	judgement , deadly , disposition , patriarchal , assumption , distance , appeal , conjecture , overweight , impacts	41.8
Finetuned RoBERTa-large	15	familiar , insights , vanilla , consistency , shone , olive , correctness , silver , incorporate , extensive	49.8	judgement , deadly , disposition , assumption , conjecture , distance , appeal , patriarchal , glare , overweight	39

Table 3: Neutral words obtained using SST on the finetuned Roberta-Base

Word	Score
Whites	-0.0
quarter	-0.01
System	-0.04
Posts	-0.05
bucks	-0.05
Religious	0.0
downright	0.01
outcome	0.01
Count	0.02
events	0.02

Table 4: Top 10 most positive-biased words using SST on finetuned Roberta-Base

Word	Score
shone	20.62
extensive	15.17
Saint	13.24
Indeed	9.67
Awareness	9.39
systematic	8.92
concerted	8.81
dominant	8.63
insights	8.62
renewable	8.18

Table 5: Top 10 most negative-biased words using SST on finetuned Roberta-Base

Word	Score
deadly	-24.48
judgement	-19.13
overweight	-14.57
speculation	-13.97
glare	-11.59
appeal	-9.94
patriarchal	-8.4
tobacco	-7.92
screaming	-6.66
replacing	-6.31

Table 6: Top 10 most positive-biased words using SST on finetuned Roberta-Base

Word	Score
shone	15.52
dominant	12.67
clout	12.49
intrigue	11.37
globalization	10.59
uncover	10.49
Circle	9.71
Saint	9.26
Beans	9.03
reflecting	8.73

Table 7: Top 10 most negative-biased words using SST on finetuned Roberta-Base

Word	Score
deadly	-36.74
judgement	-30.35
screaming	-18.26
plight	-16.29
glare	-14.28
bacterial	-14.24
appeal	-10.89
obligations	-10.87
referee	-10.46
obligation	-10.14

Table 8: Neutral words obtained using SST on the pretrained Roberta-Base

Word	Score
puppy	-0.04
silver	-0.07
expectation	-0.09
Productions	-0.1
bucket	-0.12
Hindu	0.02
Fiscal	0.03
stances	0.05
notion	0.06
supplies	0.1