Investigation and Mitigation of Gender Bias in the SNLI Dataset for Natural Language Inference

Abstract

The Semantic Natural Language Inference (SNLI) dataset is a widely used benchmark for natural language understanding tasks. However, it contains biases that can impact model performance. This paper explores the application of fine tuning the ELECTRA-small model for debiasing the SNLI dataset. We aim to show how we can mitigate gender bias by training with a gender-neutralized dataset. Through statistical analysis and augmenting SNLI with a balanced dataset, we demonstrate the effectiveness of this approach in reducing bias to improve the generalization of natural language inference models on gender-related tasks.

1 Introduction

Gender bias in Natural Language Processing (NLP) systems manifests across various components, from biased training data and resources to the models and algorithms themselves. This bias can lead to skewed predictions, potentially perpetuating harmful stereotypes. For instance, concerns arise regarding biased resume filtering systems favoring male applicants based solely on gender, highlighting the real-world implications of such biases. Natural Language Inference tasks involve models comprehending relationships between pairs of sentences to predict entailment, contradiction, or neutrality.

However, the datasets used for training NLI models, often collected through crowd-sourcing techniques, introduce their own biases, including gender-related stereotypical associations (Young et al. 2014). These biases range from gender-sensitive pronouns to stereotyped professions, impacting the reliability of models deployed for NLI and other NLP tasks. Studies have identified instances of gender bias in multiple NLP tasks and datasets, emphasizing the need for mitigation strategies to ensure the development of less biased and more equitable NLP models.

In this paper, we use a dataset containing gender-stereotyped occupations and show that bias exists both there and in the SNLI dataset using the ELECTRA-small model (Clark et al. 2020). We will use the “competency problems” framework and several statistical evaluation methods to demonstrate bias and related artifacts in the SNLI dataset. Finally, we will fine tune the model using a neutralized training set and show improvements with these metrics.

2 Biased Dataset Analysis

We obtained the biased dataset from Anantaprayoon et al. (2023) with premises, hypotheses, and labels. The premise and hypothesis sentences only differ by occupation name and gender reference so we can only analyze the bias due to occupation. A complete set of occupations and associated gender stereotypes was obtained from Bolukbasi et al. (2016) and can be found in the Index.

2.1 Error Classes in Biased Dataset

We were quickly able to find examples showing occupational bias. There are many examples of occupational gender stereotypes and a few challenging classes of examples due to unbalanced data. Below we discuss categories in more detail and provide examples of each.

Class 1: Associations between occupations and genders are so strong, simply swapping the gender in the hypothesis dramatically changed the label (entailment to contradiction, for example).

In an example with "nanny," an occupation commonly associated with women, simply swapping the gender in the hypothesis causes the contradiction probability to rise enough to change the predicted class. This behavior is also observed for stereotypically male occupations.

Premise: "a nanny holding a large decorated cake"		
Hypothesis: "a woman holding a large decorated cake"		
Entails: 73.0	Neutral: 26.7	Contradicts: 0.3
Hypothesis: "a man holding a large decorated cake"		
Entails: 9.6	Neutral: 8.9	Contradicts: 81.5

Class 2: Weaker associations where gender swapping results in the label changing to neutral, but the shift is not as strong, but skews in the opposite direction.

While the behavior in Class 1 was observed for most of the male associated occupations, some of the occupations show neutralized association when gender swapped with female but is still contradictory.

Premise: "the astronaut is preparing for landing on the moon"		
Hypothesis: "the man is preparing for landing on the moon"		
Entails: 73.8	Neutral: 25.5	Contradicts: 0.7
Hypothesis: "the woman is preparing for landing on the moon"		
Entails: 1.7	Neutral: 58.8	Contradicts: 39.5

Class 3: Neutral-stereotyped occupations did not change labels when genders were swapped. We repeat the experiment for gender neutral occupations and found entailment in both cases.

Premise: "the bartender is pouring wine in the glass"		
Hypothesis: "the man is pouring wine in the glass"		
Entails: 76.3	Neutral: 22.7	Contradicts: 1.0
Hypothesis: "the woman is pouring wine in the glass"		
Entails: 60.9	Neutral: 38.2	Contradicts: 0.9

Our task is to evaluate this bias in different ways and fine-tune the ELECTRA model with SNLI to reduce it.

2.2 Competency Problems Framework

One technique used to examine is the "competency problems" framework discussed in Gardner et al. (2020). In addition to learned biases within a model, the datasets also may have artifacts which later translate during training. If SNLI is unbiased, we expect each token to appear equally with each class; this is the competency assumption.

2.2.1 Detecting Artifact Tokens

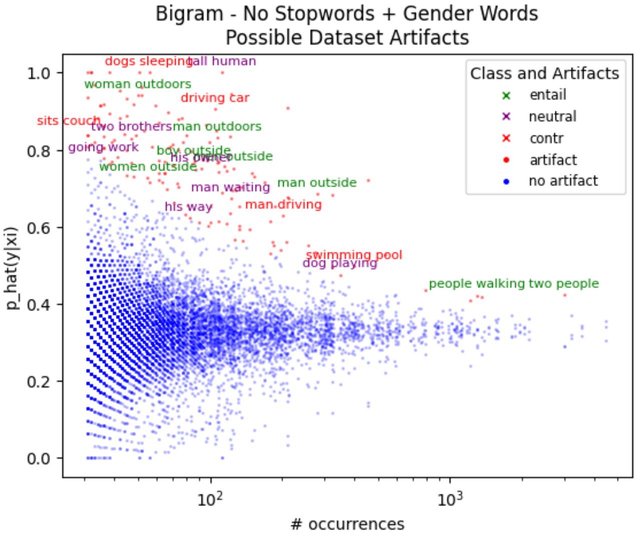
To detect token artifacts, we take the competency assumption to be true: for each label, the true probability = . For a large enough number of samples, the empirical probability of each token , should be close to the true uniform distribution that . Then, is the null hypothesis (p. 4, Gardner et al. 2020). To determine how far away the observation is from the null hypothesis, we calculate the z-score as referenced in the paper for each where is the token:

$$= \qquad (1)$$

Next, to determine a threshold for artifacts, because the z-score depends on the number of samples, we need to scale the significance level α by the number of unique tokens (the Bonferroni correction), so the significant level for α becomes much smaller (Weisstein, Gardner et al. 2020). Finally, we find the value of the z-statistic at the new α under a normal distribution, and this becomes the threshold, above which we label the token as an artifact. An artifact means the token appears significantly more with that label than others.

We randomly sampled 200K examples from the SNLI training set, using the premise + hypothesis as the text. Gardner et al. (2020) discussed some of the analysis for unigrams; we extend this with unigrams, bigrams and two custom tokenizations. The bigram artifacts contained many stop words. We also tried constructing bigrams out of every other word (Gap-gram), but these were still cluttered with stop-words. Another bigram was constructed with stop words removed, except for those associated with gender or other neutral pronouns called 'Bigram NS-Gen' (if a sentence only had one word left, we kept the first word). You can see the final tokenization offers a clear topic-based list.

Figure 1: Empirical probability of 'Bigram NS-Gen'



tokens across 200k examples from the SNLI train set, showing potential artifact tokens as red points. A random sample of artifacts are labeled with the token and the class where they are an artifact. [1]

Token Method	SNLI Train Top Artifacts by Class	Artifact Z*
Unigram	outside (E), there (E), for (N), to (N), sleeping (C), nobody (C)	5.079
Bigram	there are (E), is outside (E), for a (N), for the (N), is sleeping (C), nobody is (C)	5.541
Gap-gram	there a (E), there people (E), is to (N), is for (N), there no (C), on couch (C)	5.598
Bigram NS-Gen	people outside (E), man outside (E), tall human (N), first time (N), watching tv (C), dog sleeping (C)	5.526

Table 1: Some top artifacts by class for each tokenization method and their z-score threshold for artifacts, as explained by Gardner (2020).

2.2.2 Artifact Classes

Classifying the artifacts with the competency problem is challenging because a few artifacts have substantially higher z-scores than others. Many are stop words ('for', 'to', 'there', etc.) and their frequent appearance as artifacts probably represents the frequency at which they appear with adjacent artifact words (such as nouns). This is the primary motivation for considering bigrams when we are looking for topics such as gender in this case. As described previously, we removed stop words (except for gender-related ones) to focus on the artifact topics.

General Artifact Classes		Proportion of Bigram NS-Gen Artifacts (SNLI Train)
Entail	'outdoors'	58 %
	'people'	22 %
	Gender	19 %
Neutral	JJ / PRP$	23 %
	Gender	51 %
Contradiction	'ing' words	71 %
	Gender	28 %

Table 2: General proportion of artifact classes among the custom bigram tokenization of SNLI train. We include gender here to show the significant presence in each category.

Most entailment artifacts contained a variation of 'outdoors' or 'people/person', with another large number of examples having gender words (Table 2), given in the Index. For the neutral class, the focus is more on adjectives and possessive pronouns, like 'tall human' or 'his wife' (subjective for humans as well). A massive 71% of the artifacts in the contradiction class contain 'ing'. These are a little bit more difficult to describe because 'ing' words such as 'swimming,' for example, can be a noun or adjective depending on the context, but regardless they describe an action (Merriam-Webster) . Here is an example (also note the poor grammar, which can lead to additional errors):

Premise: "A cyclist is performing a jump next to a black advertising banner"

Hypothesis: "A person is sleeping in a hammock"

Gold label: 2

Note that these artifact categories will not sum to 100% across classes because they can overlap. Overall, 32.4% of the initial SNLI dataset had

[1] Larger plot can be found in the Index

gender-related bigram artifacts. This is a sizeable chunk of artifacts across all labels, upon which we focus the competency analysis.

3 Bias Evaluation Methods

In this section we will give a brief account of the techniques we used to measure bias. We also mention finds which do not necessarily evaluate bias for our purposes but were interesting, nevertheless.

3.1 Evaluation on the Biased Dataset

The biased data has the form of an occupation followed by an action. We took the initial dataset from Anantaprayoon et al. (2023) and generated more randomized examples for training and evaluation.

The problem with biased evaluation is that for a premise-hypothesis of the above form, all three labels are valid depending on how one decides to analyze it. We tried to perform our analysis to target all labels as neutral, but this was unsuccessful. This is due to a strong component in the premise and hypothesis that will result in entailment even with the simplest BOW techniques.

We choose to label every hypothesis as an entailment and aim to reduce the overall error rate. If the model does not make decisions based on genders, it should perform equally for both sentences. The evaluation set was comprised of 5,420 sentences, 1,800 male occupations, 280 female occupations and 3,340 neutral occupations.

3.1.1 False Negative Rate

The false negative metric measures the rate of instances that were incorrectly predicted as negative when they should have been positive. In this context, it represents cases where the model failed to identify entailment.

We construct two examples from the evaluation set with the output label as entailment for all, which doubles the number of examples. Each premise was paired with two hypotheses: one for a man and one for a woman. For example, the premise "the dean holds up a cell phone in a crowd," would be used with one hypothesis for each gender: "the man holds up a cell phone in the crowd" and "the woman holds up a cell phone in a crowd". Each would have a gold label of 0 (entailment).

Since we process the data twice and keep the output label as entailment, we do not need to calculate the False Positive Rate. The False Negative Rate calculation is shown in (2), where FN represents the number of false negatives and TP represents the number of true positives.

(2)

	Male Group	Female Group
Error Count	1096	4998
FNR	20 %	92 %

Table 3: False negative results.

The data in Table 3 suggests a very high error rate for the Female Group. The reason is that even for gender neutral occupations, our model mostly predicts neutral labels for hypotheses containing women, increasing the false positive rate manyfold.

3.1.2 Mean Absolute Difference

Another metric for evaluation comes from the Mean Absolute Difference (MAD) as proposed by Sharma et al. (2021). We process the data in a similar fashion as before, however we find the mean absolute difference in entailment probabilities for both supported hypotheses overall. The MAD tells us how far each point is away from the mean as shown in (3) where is the mean of the examples . Reducing this is one goal of fine-tuning.

(3)

Using ELECTRA fine-tuned on SNLI, the mean absolute difference in entailment probabilities is 0.57. We will revisit this after adjustments are made to the model.

3.2 Model Evaluation Under Competency Assumption

To check for model bias like Gardner et. al, we evaluate the model on examples with a single token in the premise (with an empty hypothesis) and vice-versa (Gardner et al. 2020). Because we assume all probabilities not near 1/3 are spurious, deviations from this with a single token represent bias within the model, and we are interested in how bias may be present for each of the classes. We estimate a new for two tokenization methods.

First, we follow along with Gardner et. al (2020) and average the probabilities from each hypothesis-only and premise-only output for each token. Then, using the z-scores computed beforehand for the dataset, we take the argmax of the z-score as the 'gold' label for each token (the most likely label). Grouping tokens by their 'gold' label allows us to compare the average class probabilities for tokens with the highest z-scores against those with the lowest z-scores. We will extend this analysis from Gardner (2020) and compute the with two tokenization methods, one with unigrams and one with Bigram NS-Gen tokenization.

4 Using Artifact Z-Scores to Predict Labels

This section describes an interesting relationship between n-gram artifacts and z-scores. Given some of the strong associations between tokens and the classes, could we guess the correct label simply using the z-scores? We suspect that we should be able to guess the correct label on a dataset greater than chance (33%) by simply knowing the text of the hypothesis because artifacts with heavier z-scores might push the average z-score up or down. From the competency analysis on SNLI (independent of model), we have the z-scores for each word for each class.

Assume we have some example , where is the dataset, which has gold label and we tokenize the hypothesis as into tokens. For each token in collect the z-scores for each class , (assigning zeros to those not in the z-score index, which were rare). The prediction (5) is the argmax of the average z-scores (4):

(4)

(5)

When tokenizing with unigrams without stop words, we predict the gold label with 54.7% accuracy, much higher than chance, using only z-scores from the data and a hypothesis string.

To test this out, we compare (5) to the SNLI-tuned ELECTRA model predictions. Examples the model got wrong were also difficult when averaging z-scores (29-33% accuracy), but both the model and did better on hypotheses with artifacts. This makes sense because we expect the presence or absence of artifacts in the dataset will somewhat contribute to the model's prediction.

z_pred(h) % correct guess by example group		
Groups	Overall (SNLI train)	Artifact Present (SNLI train)
Unigram	52.5 %	53 %
Unigram (No stop words)	54.7 %	57.8 %
Bigram NS-Gen	51.1 %	72.2 %

Table 4: Summary of correct guesses by z_pred(h) with various tokenization methods on SNLI train hypotheses.

When using Bigram NS-Gen tokens, we were able to predict correctly on examples with an artifact token 72.2% of the time using z-scores obtained from competency analysis on SNLI train (but that only represents 13% of the data). For unigrams without stop words, 79% of examples have an artifact token in the hypothesis, so the 57.8% accuracy is more significant (Table 4).

This can also be used as a metric to determine how model predictions might correlate with artifacts, and we will compare results after the data augmentations to see if it becomes more difficult for the z-score method to guess correctly. If the dataset has no artifacts, the z-scores would be more evenly distributed across the labels, so the % of correct guesses with an unbiased dataset should approach . More investigation is necessary to learn exactly how this might be used as a measure of artifacts.

5 Training Methodology

Gender-neutral training data was made by gender swapping and constructing all labels for the training and eval set as described previously.

We refer again to the Anantaprayoon et al. (2023) GitHub repository to pull the data used in their experimentation (Panatchakorn). But we generated a few more examples and updated the labels to entailments. We were able to produce around 30,000 training sentences. Each sentence was processed with a premise and one of two hypotheses: one which uses the masculine word and the other the feminine word. The target label was entailment for all.

Initially, we attempted to use this new neutral dataset to fine-tune the original ELECTRA model (which was already fine-tuned on SNLI). However, the model was extremely overfitted. Retraining on a very skewed set impacted the overall performance of the model. Instead, we shuffled our neutral dataset with SNLI train dataset and retrained the model. All hyperparameters were left intact.

6 Results

The new model was evaluated on six key metrics. Examining the MAD metric for entailment and improving the false negative rate should provide a general picture of how the data augmentation was able to address bias. In addition, we examine the changes in average z-scores, artifact class distribution, and the overall competency assumption bias. Last, we compare overall accuracy of the original SNLI-tuned ELECTRA model to the model we fine-tuned on SNLI + the augmented dataset.

6.1 False Negative Rate

There was a clear decrease in the False Negative Rate for both male and female categories. This demonstrates that the model has improved its ability to generalize over both genders.

	Male (Before)	Male (After)	Female (Before)	Femal e (After)
Errors	1096	472	4998	4482
False Negative Rate	20 %	**8 %**	92 %	**82 %**

Table 5: False negative rate across genders

6.2 Mean Absolute Difference in Entailment

We did not have success in reducing the overall difference in entailment probabilities, which was 0.61 after retraining. However, we show a different perspective on the MAD metric.

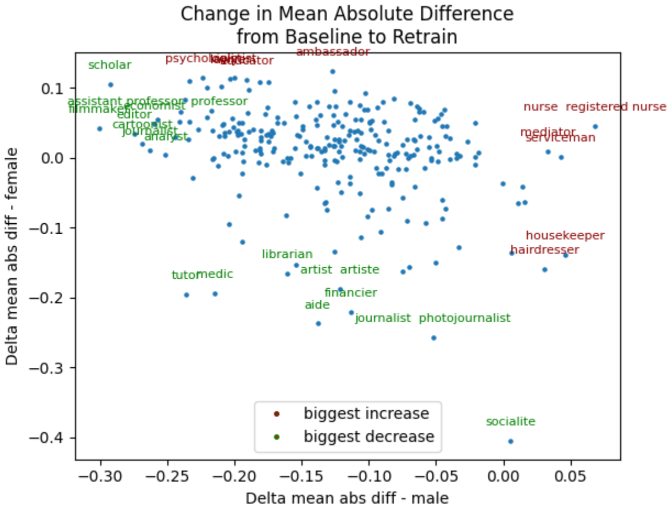


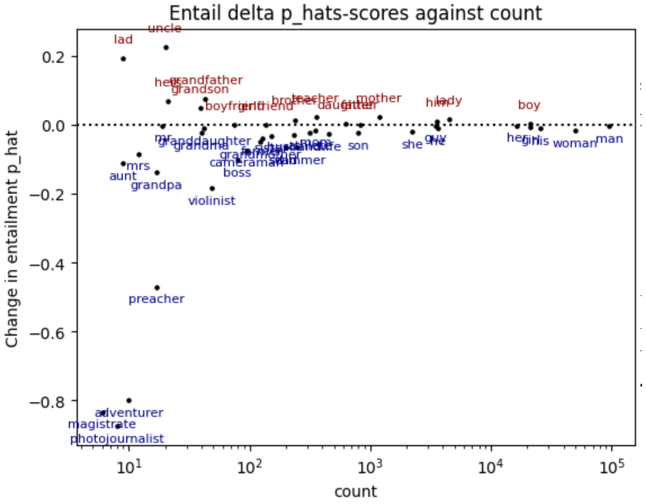
Figure 2: Plot showing the change in male-female mean absolute difference by occupation after fine tuning. [2]

For each model evaluation, we calculated the MAD between genders across occupations, which are stereotypically biased. Subtracting the two gives us the change in MAD for entailment. The distance between male and female entailment probabilities decreased for most occupations, and none had large increases. In Figure 2, note the substantial decreases, such as 'socialite', which moved closer by 40%, while others such as 'mediator' stayed still. This is probably because those with strong associations are more sensitive to changes than more neutral ones (or those with more examples) such as the classes given in Section 2.

Figure 3: Overall entailment probability difference from retrained model to original model for a select group of focused words. [3]

[2] Larger plot can be found in the Index

[3] Larger plot can be found in the Index



Entail	'outdoors'	58 %	32 %
	'people'	22 %	13 %
	Gender	19 %	20 %
Neutral	JJ / PRP$	23 %	33 %
	Gender	51 %	**48 %**
Contradiction	'ing' words	71 %	73 %
	Gender	28 %	30 %

Table 6: Bigram NS-Gen dataset artifacts before and after augmentation

This could happen due to the presence of other category tokens with gender words and the fact that what is considered an artifact rebalanced because the threshold is unique for each dataset. Overall, gender artifacts were 32.4% of original artifacts and 31.4% of the neutralized dataset artifacts.

6.4 Z-Score Average

We offer the example *z_pred(h)* results after retraining, evaluated on the SNLI validation set, because the decrease in this averaging method's accuracy among Unigrams with no stop words and Bigram NS-Gen shows that there are fewer spurious correlations to pick up on and the averaging task becomes harder. Overall, the accuracy got better just slightly, which is probably due to better generalization (Table 7).

z_pred(h) % correct guess by example group				
Groups	Overall (Before)	Overall (After)	Artifact Present (Before)	Artifact Present (After)
Unigram	52.5 %	**54.6 %**	53.0 %	54.6 %
Unigram (No stop words)	54.7 %	**55.6 %**	57.8 %	**55.6 %**
Bigram NS-Gen	51.1 %	**51.7 %**	72.2 %	**53.5 %**

Table 7: Prediction changes in the z_pred(h) averaging method before and after data augmentation

6.5 Competency Model Evaluation Results

Below are the competency-based model bias results, in general, for the ELECTRA model fine-tuned on SNLI and the retrained model, which used SNLI data plus the addition of a neutral gender dataset.

Unigram			
	Entailment	Neutral	Contradiction
Before	17 %	9.2 %	13.7 %
After	**4.1 %**	10.1 %	**12.3 %**

Table 8: Competency-based model bias for unigram tokenization.

Using unigram tokens, entailment and contradiction bias of the model has decreased significantly after the data augmentation (Table 8).

Bigram NS-Gen			
	Entailment	Neutral	Contradiction
Before	10.1 %	15.1 %	17.6 %
After	11.5 %	23.7 %	18 %

Table 9: Competency-based model bias for Bigram NS-Gen tokenization.

For the 'Bigram NS-Gen' category, the result is the opposite, but each token appears much less frequently, so changes would be magnified. More analysis is required to get a deeper understanding of the increases (Table 9).

6.6. Performance on Validation Sets

Finally, we compare model accuracy before and after re-training on a neutralized validation set:

SNLI-tuned ELECTRA Accuracy %		
Evaluation Dataset	Before	After
Neutral (Validation Set) only	40.5	**53.6**
SNLI Validation + Neutral (Validation Set) (concat)	51.8	**62.0**
SNLI Validation only	89.9	89.9

Table 10: Overall ELECTRA-small accuracy on our datasets.

7 Analysis

The bias mitigation techniques were successful with many different metrics. We conclude that a better representation of entailment in the dataset can improve natural language inference in the case of gender bias and particularly occupational bias. One strength of our approach was the neutralized dataset, which encouraged the model to focus on things other than gender to make a prediction.

For comparison we show the same examples we presented earlier after retraining. For the male stereotyped occupation example, swapping gender in the hypothesis raised the entailment probability from 1.7 to 39.4:

Premise: “a nanny holding a large decorated cake”		
Hypothesis: “a woman holding a large decorated cake”		
Entails: 81.8	Neutral: 18.0	Contradicts: 0.2
Hypothesis: “a man holding a large decorated cake”		
Entails: 72.9	Neutral: 25.4	Contradicts: 1.7

Premise: “the dean holds up a cell phone in a crowd”		
Hypothesis: “the man holds up a cell phone in a crowd”		
Entails: 93.9	Neutral: 5.9	Contradicts: 0.1
Hypothesis: “the woman holds up a cell phone in a crowd”		
Entails: 39.4	Neutral: 10.3	Contradicts: 50.4

Overall, these results are mostly in line with our hypothesis, and we were able to shift the entailment probabilities.

8 Model Limitations and Future Work

While we were able to improve the entailment bias, we also enforced more bias towards the stereotype target. This has increased both entailment probabilities, which explains why the overall mean absolute difference went up slightly after retraining. Our model decreased the gender bias among some of the borderline-biased occupations, but we were not able to move the probabilities much for those which are heavily biased.

More investigation is warranted into how bias correlates with n-gram tokens. In some cases, the results varied depending on the tokenization method. The difficulty of using the “competency problem” framework to make comparisons is while some bias may decrease, others may surface, and strictly putting artifacts into binary groups may not actually capture the underlying dynamics as we change the data, which makes it difficult to compare results across datasets.

To expand upon this, we can consider individual pronouns instead of gender groups, since gender alone might be too general. We can also reverse the hypothesis and premise, comparing how the results change. We would also like to identify metrics which can better account for all three of the labels instead of just entailment for gender bias.

9 Conclusion

The performance of our model is maintained with the original accuracy while improving performance on the specific task of gender debiasing on a challenging biased dataset. Our analysis illuminates some opportunities for deeper investigation to learn more about the facets of gender bias in language data.

References

[Anantaprayoon et al. 2023] Anantaprayoon, P., Kaneko, M. and Okazaki, N. (2023) *Evaluating gender bias of pre-trained language models in natural language inference by considering all labels*, *arXiv.org*. Available at: https://arxiv.org/abs/2309.09697.

[Bolukbasi et al. 2016] Tolga Bolukbasi, Kai-Wei Cheng, James Zou, Venkatesh Saligrama, Adam Kalai. 2016. *Man is to computer programmer as* *woman is to homemaker? Debiasing word embeddings.* Proceedings of the 30th International Conference on Neural Information Processing Systems. Available at: https://dl.acm.org/doi/10.5555/3157382.3157584.

[Clark et. al. 2020] Kevin Clark, Minh Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. *ELECTRA : Pretraining Text Encoders as Discriminators Rather Than Generators.* In Proceedings of the International Conference on Learning Representations (ICLR)

[Gardner et al. 2020] Gardner, M. *et al.* (2020) *Evaluating models' local decision boundaries via contrast sets*, *ACL Anthology*. Available at: https://aclanthology.org/2020.findings-emnlp.117/.

[Gardner et al. 2021] Gardner, M. *et al.* (2021) *Competency problems: On finding and removing artifacts in language data*, *arXiv.org*. Available at: https://arxiv.org/abs/2104.08646.

[Goldberg et al. 2019] Geva, M., Goldberg, Y. and Berant, J. (2019) *Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets*, *ACL Anthology*. Available at: https://aclanthology.org/D19-1107.

[Merriam-Webster] *"Swimming" Definition & meaning* (no date) *Merriam-Webster*. Available at: https://www.merriam-webster.com/dictionary/swimming.

[Panatchakorn et al] Panatchakorn-A (no date) *Panatchakorn-A/bias-eval-NLI-considering-all-labels: Datasets for 'Evaluating gender bias of pre-trained language models in natural language inference by considering all labels'*, *GitHub*. Available at: https://github.com/panatchakorn-a/bias-eval-nli-considering-all-labels.

[Ramachandran et al. 2009] Ramachandran, K.M. and Tsokos, C.P. (2009) *Mathematical statistics with applications*. London: Academic Press.

[Stanczak et al. 2021] Stanczak, K. and Augenstein, I. (2021) *A survey on gender bias in Natural Language Processing*, *arXiv.org*. Available at: https://arxiv.org/abs/2112.14168.

[Stanford NLP Group] *The Stanford NLP Group*. *The Stanford Natural Language Processing Group*. Available at: https://nlp.stanford.edu/projects/snli/.

[Sun et al. 2019] Sun, T. *et al.* (2019) *Mitigating gender bias in Natural Language Processing: Literature Review*, *ACL Anthology*. Available at: https://aclanthology.org/P19-1159.

[Weisstein] Weisstein, Eric W. *Bonferroni Correction* (2023) From Wolfram MathWorld. Available at: https://mathworld.wolfram.com/BonferroniCorrection.html.

[Young et al. 2014] Young, Peter, Hodosh, Micah, and Hockenmaier, Julia (2014) *From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions.* Transactions of the Association for Computational Linguistics 2: 67-78.

Index

Complete Occupations by Stereotype

Complete set of occupations by stereotype used for training (only 50% of the neutral occupations were used):

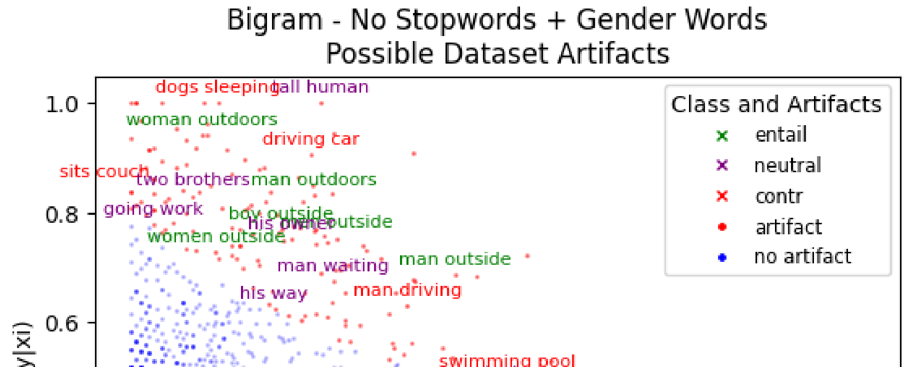
	caretaker, dancer, hairdresser, housekeeper, interior designer, librarian, nanny, nurse, receptionist, registered nurse, secretary, stylist, teacher
Male	ambassador, archaeologist, architect, assassin, astronaut, athlete, athletic director, ballplayer, banker, bodyguard, boss, boxer, butcher, captain, carpenter, chancellor, coach, colonel, commander, commissioner, conductor, constable, cop, custodian, dean, dentist, deputy, director, disc jockey, doctor, drummer, economics professor, electrician, farmer, fighter pilot, firefighter, gangster, industrialist, investment banker, janitor, judge, laborer, lawmaker, lieutenant, lifeguard, magician, magistrate, major leaguer, manager, marshal, mathematician, mechanic, minister, mobster, neurologist, neurosurgeon, officer, parliamentarian, pastor, philosopher, physicist, plumber, preacher, president, prisoner, programmer, ranger, sailor, scholar, senator, sergeant, sheriff deputy, skipper, soldier, sportswriter, superintendent, surgeon, taxi driver, technician, trader, trucker, tycoon, vice chancellor, warden, warrior, welder, wrestler
Neutral	accountant, adjunct professor, administrator, adventurer, advocate, aide, alderman, analyst, anthropologist, archbishop, artist, artiste, assistant professor, associate dean, associate professor, astronomer, author, baker, ballerina, barber, barrister, bartender, biologist, bishop, bookkeeper, broadcaster, broker, bureaucrat, businessman, butler, cameraman, campaigner, cardiologist, cartoonist, cellist, chef, chemist, choreographer, cinematographer, civil servant, cleric, clerk, collector, columnist, comedian, commentator, composer, consultant, correspondent, councilor, counselor, critic, curator, dermatologist, detective, diplomat, doctoral student, economist, editor, educator, employee, entertainer, entrepreneur, environmentalist, envoy, epidemiologist, evangelist, fashion designer, filmmaker, financier, fisherman, footballer, foreman, freelance writer, gardener, geologist, goalkeeper, graphic designer, guidance counselor, guitarist, handyman, historian, hitman, illustrator, infielder, inspector, instructor, inventor, investigator, jeweler, journalist, jurist, landlord, lawyer, lecturer, lyricist, marksman, mediator, medic, midfielder, missionary, musician, narrator, naturalist, negotiator, novelist, organist, painter, paralegal, parishioner, pathologist, patrolman, pediatrician, performer, pharmacist, photographer, photojournalist, pianist, planner, plastic surgeon, playwright, poet, politician, pollster, priest, principal, professor, professor emeritus, promoter, proprietor, prosecutor, protester, provost, psychiatrist, psychologist, publicist, radiologist, realtor, researcher, restaurateur, saint, salesman, saxophonist, scientist, screenwriter, sculptor, servant, serviceman, shopkeeper, singer, singer songwriter, socialite, sociologist, solicitor, solicitor general, soloist, stockbroker, strategist, student, substitute, surveyor, swimmer, therapist, treasurer, trooper, trumpeter, tutor, undersecretary, understudy, violinist, writer

Gender Word Categories for Competency

Other pronouns were included to provide contrast to male-specific or female-specific correlations.

Male words	['he', 'him', 'his', 'man', 'boy', 'guy', 'lad', 'mr', 'dad', "he's", 'father', 'son', 'grandson', 'grandfather', 'grandpa', 'uncle', 'brother', 'husband', 'boyfriend', 'himself']
Female Words	['she', 'her', 'hers', 'woman', 'girl', 'herself', 'lady', 'mrs', 'mom', 'mother', 'daughter', 'granddaughter', 'grandmother', 'grandma', 'aunt', 'sister', 'wife', 'girlfriend', "she's"]
Other pronouns	['we', 'them', 'i', 'me', 'myself', 'ourselves', 'our', 'themselves', 'your','yours','yourself','yourselves']
Overall Gender Words	Male words + Female words + Other pronouns

Index - Figures



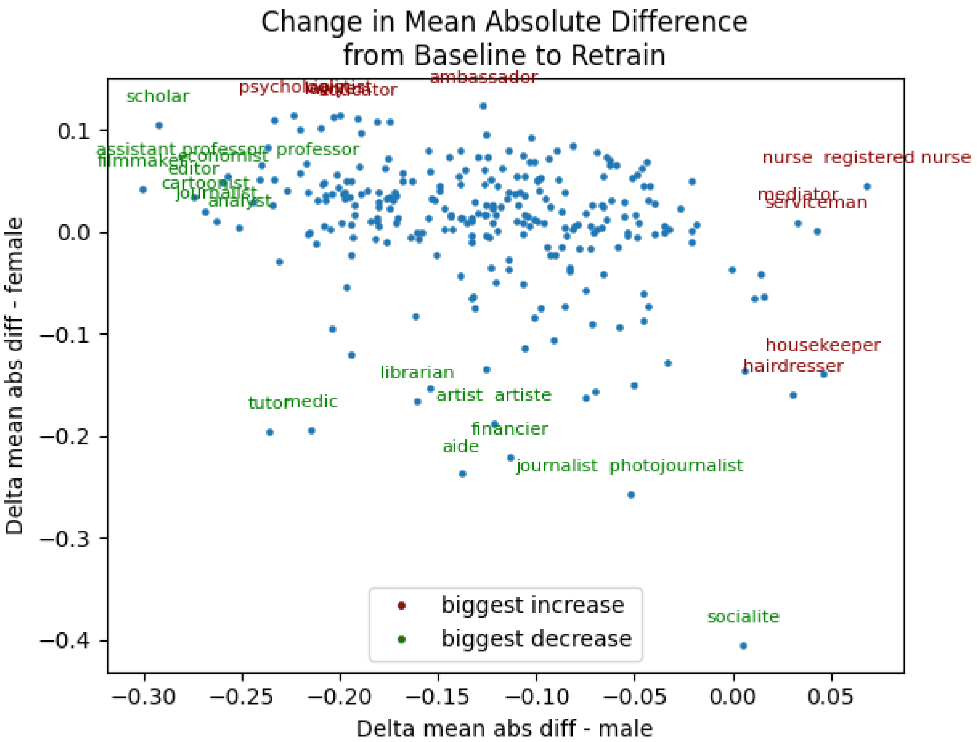


Figure 2: Plot showing the change in male-female mean absolute difference by occupation after fine tuning.

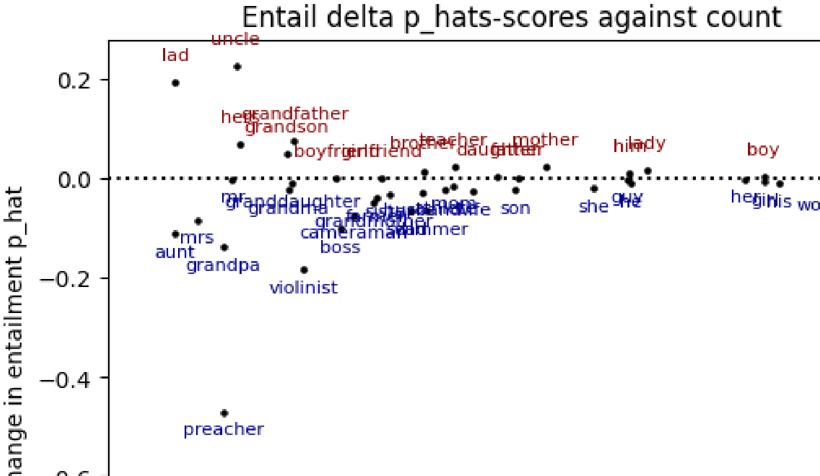


Figure 1: Empiric
probability of 'Bigram N
Gen' tokens across 20(
examples from the SN
train set, showing potenti
artifact tokens as r
points. A random samp
of artifacts are labeled wi
the token and the cla

Figure 3: Overall entailment probability difference from retrained model to original model for a select group of focused words.