Supplementary Materials for: Identifying and Reducing Gender Bias in Word-Level Language Models

A Defining sets

The gender pair list for each corpus is designed separately. We consider only those gender pairs that occur in the training corpus. Below are the gender lists corresponding to each corpus:

A.1 Penn Treebank

Male Words: "actor" "boy" "father" "he" "him " "his" "male" "man" "men" "son" "sons" "spokesman" "wife" "king" "brother"
Female Words: "actress" "girl " "mother" "she" "her " "her" "female" "woman" "women" "daughter" "daughters" "spokeswoman" "husband" "queen" "sister"

A.2 WikiText-2

Male Words: "actor" "Actor" "boy" "Boy" "boyfriend" "Boys" "boys" "father" "Father" "Fathers" "fathers" "Gentleman" "gentleman" "gentlemen" "Gentlemen" "grandson" "he" "He" "hero" "him" "Him" "his" "His" "Husband" "husbands" "King" "kings" "Kings" "male" "Male" "males" "Males" "man" "Man" "men" "Men" "Mr." "Prince" "prince" "son" "sons" "spokesman" "stepfather" "uncle" "husband" "king"
Female Words: "actress" "Actress" "girl" "Girl" "girlfriend" "Girls" "girls" "mother" "Mother" "Mothers" "mothers" "Lady" "lady" "ladies" "Ladies" "granddaughter" "she" "She" "heroine" "her" "Her" "her" "Her" "Wife" "wives" "Queen" "queens" "Queens" "female" "Female" "females" "Females" "woman" "Woman" "women" "Women" "Mrs." "Princess" "princess" "daughter" "daughters" "spokeswoman" "stepmother" "aunt" "wife" "queen"

A.3 CNN/Daily Mail

Male Words: "actor" "boy" "boyfriend" "boys" "father" "fathers" "gentleman" "gentlemen" "grandson" "he" "him" "his" "husbands" "kings" "male" "males"

“man” “men” “prince” “son” “sons” “spokesman” “stepfather” “uncle” “wife” “king” “brother” “brothers”
Female Words: “actress” “girl” “girlfriend” “girls” “mother” “mothers” “lady” “ladies” “granddaughter” “she” “her” “her” “wives” “queens” “female” “females” “woman” “women” “princess” “daughter” “daughters” “spokeswoman” “stepmother” “aunt” “husband” “queen” “sister” “sisters”

B Word Level Bias Examples

Tables 1 and 2 show the bias scores at individual word level for selected words for Wikitext-2. The tables show how the scores vary for the training text and the generated text for different values of λ
Tables 3 and 4 show the bias scores at individual word level for selected words for CNN/Daily Mail. The tables show how the scores vary for the training text and the generated text for different values of λ

Target Words	$training$	λ=0.0	λ=0.01	λ=0.1	λ=0.5	λ=0.8	λ=1.0
Arts	-0.76	-1.20	-0.87	-0.32	-0.17	0.13	-1.48
Boston	-0.95	-1.06	-0.23	-1.06	-0.13	-0.37	-0.94
Edward	-0.68	-1.06	0.09	-0.56	-0.14	-0.44	-0.23
George	-0.52	-0.91	-0.26	-0.22	-0.48	-0.26	0.01
Henry	-0.59	-1.06	0.11	-0.34	-0.84	-0.92	-0.61
Peter	-0.69	-2.06	-0.09	-0.14	-0.32	0.08	0.53
Royal	-0.01	-1.89	-0.39	-0.61	-0.64	-1.14	-0.56
Sir	-0.01	-1.76	-0.99	-0.86	-0.64	-0.16	0.07
Stephen	-0.35	-1.20	-0.18	-1.01	-0.84	-0.11	0.36
Taylor	-0.84	-0.91	0.57	0.00	-0.01	-0.39	-0.83
ambassador	-0.76	-1.20	-0.23	-0.63	-0.74	0.43	-0.81
failed	-0.46	-2.06	0.03	-0.36	-1.00	0.17	
focused	-0.22	-0.91	-0.12	-0.41	-0.40	-0.57	-0.53
idea	-0.20	-1.06	-0.36	-0.16	-0.27	-0.06	-0.42
manager	-1.58	-1.60	-0.04	-0.30	-1.08	-0.30	-1.06
students	-0.60	-0.79	-0.31	-0.29	-0.32	-0.51	-0.50
university	-0.12	-1.06	0.17	-1.01	-0.79	-0.95	-0.70
wife	-0.92	-1.29	-0.81	-1.02	-0.57	-0.67	-1.03
work	-0.24	-0.88	-0.48	-0.23	-0.49	-0.52	-0.13
youth	-0.39	-1.20	0.54	-0.16	-0.68	0.58	

Table 1: WikiText-2 bias scores for the words biased towards male gender for different λ values

Target Words	$training$	λ=0.0	λ=0.01	λ=0.1	λ=0.5	λ=0.8	λ=1.0
Katherine	1.78	2.27	1.38	0.69	0.95	0.75	0.70
Zenobia	0.05	0.88	1.84	0.47	0.65	1.24	
childhood	0.48	1.80	0.12	1.10	0.37	0.38	0.34
cousin	0.13	0.88	0.67	0.13	0.09	0.67	0.71
humor	0.34	1.29		0.69	0.61	0.34	
invitation	0.19	1.80	-0.87	0.69	0.57	-0.44	-0.25
parents	0.51	0.76	0.77	0.08	0.45	0.57	1.11
partners	0.85	2.27	-0.28	0.98	0.87	-0.17	3.22
performances	0.79	1.02	-0.20	0.16	0.03	0.10	-1.80
producers	1.04	1.58	0.33	0.78	1.35	-1.45	0.18
readers	0.22	0.88	0.28	0.29	0.36	-0.32	-1.29
stars	0.85	1.58	0.16	0.90	0.46	-0.28	-0.08
talent	0.02	0.88	-0.75	0.10	0.31	-0.86	
wore	0.09	0.88	0.48	0.29	0.65	0.16	-0.69

Table 2: WikiText-2 bias scores for the words biased towards female gender for different λ values

Target Words	$training$	λ=0.0	λ=0.1	λ=0.5	λ=0.8	λ=1.0
abusers	-0.66	-1.17	-0.56	-0.77	-0.16	-1.93
acting	-0.23	-0.81	-0.59	-0.35	-0.54	0.60
actions	-0.27	-0.51	-0.06	-0.07	-0.53	-0.45
barrister	-1.35	-2.00	-0.64	-0.76	-0.08	-0.69
battle	-0.27	-0.53	-0.10	-0.32	-0.16	0.21
beneficiary	-1.64	-1.87	-1.06	-0.22	0.63	
bills	-0.32	-0.53	-0.18	-0.50	0.23	0.69
businessman	-0.19	-1.81	-0.71	-0.45	-0.53	-1.93
cars	-0.43	-0.55	-0.32	-0.11	-0.24	-0.27
citizen	-0.03	-0.30	-0.03	-0.22	-0.01	0.04
cocaine	-0.59	-1.00	-0.84	-0.44	-0.42	-0.32
conspiracy	-0.57	-0.73	-0.66	-0.39	-0.83	-0.43
controversial	-0.21	-0.39	-0.39	-0.02	-0.17	-0.43
cooking	-0.48	-0.53	-0.24	-0.22	0.07	-0.52
cop	-1.30	-1.42	-0.77	-0.72	0.00	0.26
drug	-0.76	-0.82	-0.53	-0.42	-0.54	-0.63
executive	-0.04	-0.34	-0.22	-0.04	-0.48	-0.36
fighter	-0.59	-0.90	-0.48	-0.36	-0.89	-0.11
fraud	-0.17	-0.30	-0.16	-0.19	0.10	-0.63
friendly	-0.48	-0.53	-0.30	-0.23	0.36	-0.21
heroin	-0.57	-0.67	-0.28	-0.26	-0.66	0.49
journalists	-0.25	-1.08	-0.55	-0.76	-0.44	
lawyer	-0.39	-0.47	-0.14	-0.10	-0.20	-0.50
lead	-0.47	-0.50	-0.40	-0.09	-0.07	-0.32
leadership	-0.25	-0.74	-0.28	-0.68	-0.57	-0.99
notorious	-0.18	-0.64	-0.36	-0.22	-0.12	-1.49
offensive	-0.17	-0.39	-0.28	-0.17	-0.52	-0.11
officer	-0.25	-0.29	-0.21	-0.13	-0.17	-0.19
outstanding	-0.25	-1.55	-0.98	-0.50	0.03	0.04
parole	-0.54	-0.86	0.00	-0.08	0.07	-1.30
pensioners	-0.48	-0.86	-0.77	-0.07	0.64	0.40
prisoners	-0.52	-0.99	-0.18	-0.29	-0.17	-1.59
religion	-0.41	-0.97	-0.15	-0.48	0.18	-1.68
reporters	-0.60	-0.93	-0.26	-0.05	-0.52	-0.97
representatives	-0.07	-0.48	-0.40	-0.18	-0.46	-0.83
research	-0.34	-0.46	-0.05	-0.33	0.03	-0.58
resignation	-0.95	-1.67	-0.61	-0.58	-0.40	-1.12
sacrifice	-0.03	-1.08	-0.38	-0.17	-1.29	
supervisor	-0.66	-0.92	-0.44	-0.25	-0.17	0.48
violent	-0.17	-0.54	-0.07	-0.22	-0.19	-0.19

Table 3: CNN/Daily Mail bias scores for the words biased towards male gender for different λ values

Target Words	*training*	λ=0.0	λ=0.1	λ=0.5	λ=0.8	λ=1.0
abusive	0.00	0.40	0.06	0.39	0.48	-0.65
appealing	0.44	1.22	0.23	0.30	-0.68	1.16
bags	0.34	1.42	0.48	0.05	0.16	0.64
beloved	0.17	0.35	0.27	0.15	0.52	0.36
carol	0.76	1.41	0.20	0.39	0.27	0.48
chatted	0.03	1.83	0.20	0.19	-0.14	-0.25
children	0.29	0.46	0.36	0.26	0.41	0.27
comments	0.17	0.46	0.04	0.02	-0.35	-0.14
crying	0.28	0.70	0.19	0.57	0.17	0.87
designer	0.73	0.80	0.57	0.69	0.53	-1.53
designers	0.44	2.14	1.29	0.76	-0.11	1.11
distressed	0.15	0.53	0.23	0.26	-0.56	1.36
divorced	0.68	0.70	0.18	0.10	0.31	0.88
dollar	0.44	1.63	0.65	0.59	-0.24	
donated	0.52	0.57	0.06	0.15	0.68	0.26
donating	1.29	1.38	0.27	0.80	-0.03	-0.21
embracing	1.13	1.78	0.74	0.55	1.48	-0.94
encouragement	0.85	0.94	0.22	0.50	0.37	0.55
endure	0.85	0.94	0.26	0.29	1.02	
expecting	1.01	1.07	0.26	0.12	0.53	0.06
feeling	0.21	0.84	0.16	0.25	0.16	0.29
festive	0.15	0.53	0.52	0.14	0.21	0.26
fragile	0.44	0.94	0.20	0.45	-0.20	
happy	0.32	0.66	0.10	0.11	0.11	0.25
healthy	0.52	0.64	0.26	0.45	0.24	0.25
hooked	0.78	1.38	0.12	0.12	-0.11	-0.09
hurting	0.75	1.13	0.33	0.34	0.44	0.26
indian	0.18	0.28	0.15	0.02	-0.02	-0.26
kissed	0.31	1.03	0.17	0.19	0.28	-0.22
kissing	0.26	1.14	0.54	0.61	0.44	-0.14
loving	0.41	0.73	0.43	0.18	0.15	-0.34
luxurious	0.59	0.82	0.17	0.44	-0.03	-0.83
makeup	1.60	1.63	0.07	0.22	1.09	
mannequin	0.95	1.92	0.70	0.04	1.42	
married	0.29	0.37	0.34	0.09	0.30	0.42
models	0.35	1.22	0.28	0.38	0.90	0.08
pictures	0.08	0.50	0.10	0.04	-0.06	0.59
pray	0.62	1.58	0.25	0.35	-0.25	0.96
relationship	0.53	0.62	0.39	0.32	0.58	0.43
scholarship	0.80	1.16	0.80	0.70	0.53	0.45
sharing	0.58	0.73	0.33	0.67	0.42	0.17
sleeping	0.18	0.71	0.27	0.35	0.56	0.58
stealing	0.10	0.48	0.32	0.18	0.06	-0.53
tears	0.50	0.58	0.44	0.12	0.45	0.35
thanksgiving	0.85	2.14	1.14	1.08	0.90	
waist	1.33	1.45	0.68	0.02	0.31	0.96

Table 4: CNN/Daily Mail bias scores for the words biased towards female gender for different λ values