Online Appendix to "Gendered Language on the Economics Job Market Rumors Forum"

Alice H. Wu

Model I. Lasso-regularized Logistic Model

Letting $\mathbf{w}_i$ denote a vector of counts for each of the most common words (excluding the female or male classifiers) that are present in gendered post i, I assume the posterior probabilities are:

$$\begin{aligned} P(Female_i = 1|\mathbf{w}_i) &= \frac{exp(\theta_0 + \mathbf{w}_i'\theta)}{1 + exp(\theta_0 + \mathbf{w}_i'\theta)} \\ P(Female_i = 0|\mathbf{w}_i) &= \frac{1}{1 + exp(\theta_0 + \mathbf{w}_i'\theta)} \end{aligned} \tag{1}$$

Write the likelihood of each observation as:

$$P(Female_i|\mathbf{w}_i) = P(Female_i = 1|\mathbf{w}_i)^{Female_i} \times P(Female_i = 0|\mathbf{w}_i)^{(1-Female_i)} \tag{2}$$

Assume the observations are independent, the log likelihood for N observations is

$$\begin{aligned} l_N(\theta) &= log(\Pi_{i=1}^N P(Female_i|\mathbf{w}_i)) \\ &= \sum_{i=1}^{N} [Female_i * (\theta_0 + \mathbf{w}_i'\theta) - log(1 + exp(\theta_0 + \mathbf{w}_i'\theta))] \end{aligned} \tag{3}$$

I estimate θ on the counts for words through the following objective function[1]:

$$\begin{aligned} \hat{\theta}_\lambda &= argmin_\theta \, (-l_N(\theta)) + \lambda\|\theta\|_1 \\ &= argmin_\theta \, \Sigma_i[log(1 + exp(\theta_0 + \mathbf{w}_i'\theta)) - Female_i(\theta_0 + \mathbf{w}_i'\theta)] + \lambda\|\theta\|_1 \end{aligned} \tag{4}$$

where $\|\theta\|_1 = \sum_{j\geq 1} |\theta^j|$.

Given a word k, we have

$$\frac{\partial P(Female_i = 1|\mathbf{w}_i)}{\partial w_i^k} = P(Female_i = 1|\mathbf{w}_i) * P(Female_i = 0|\mathbf{w}_i) * \theta_\lambda^k \tag{5}$$

where θ_λ^k is the coefficient on w_i^k - the count for word k in post i. Therefore, I estimate the average marginal effect of word k by

$$\frac{1}{N}\Sigma_i P(Female_i = 1|\mathbf{w}_i) * P(Female_i = 0|\mathbf{w}_i) * \hat{\theta_\lambda^k} \tag{6}$$

[1] See Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. Springer. Second Edition. for a detailed discussion of penalized logistic regressions.

Model II. Lasso-regularized Linear Probability Model

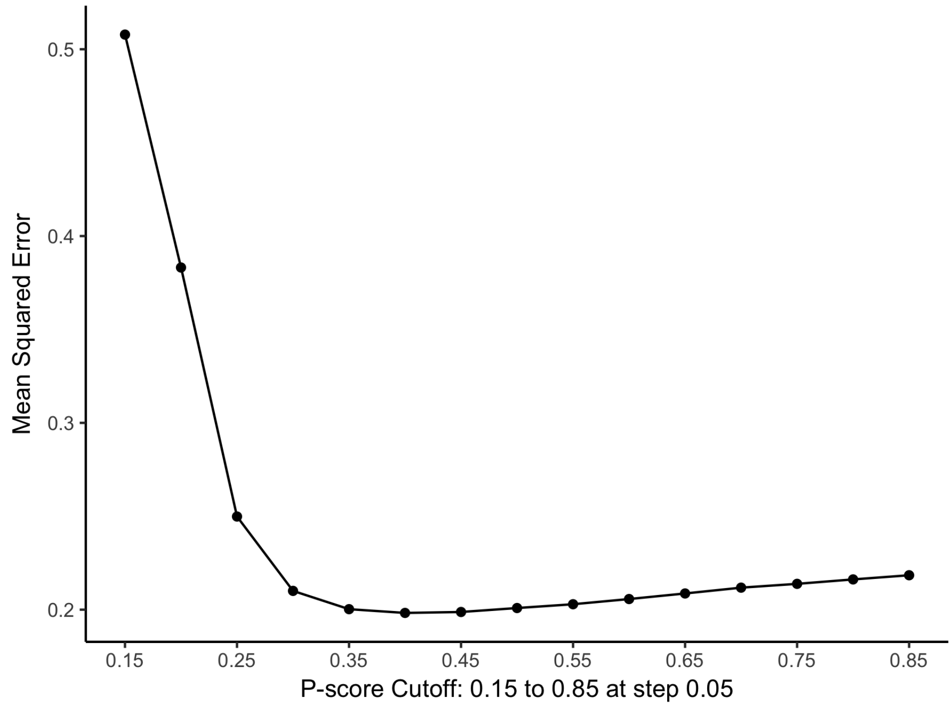
Using the same notations as above, I estimate an regularized linear probability model as follows:

$$\hat{\beta}_\lambda = argmin_\beta \, \Sigma_i (Female_i - \beta_0 - \mathbf{w}_i' \beta)^2 + \lambda \|\beta\|_1 \tag{7}$$

where $\|\beta\|_1 = \sum_{j \geq 1} |\beta^j|$.

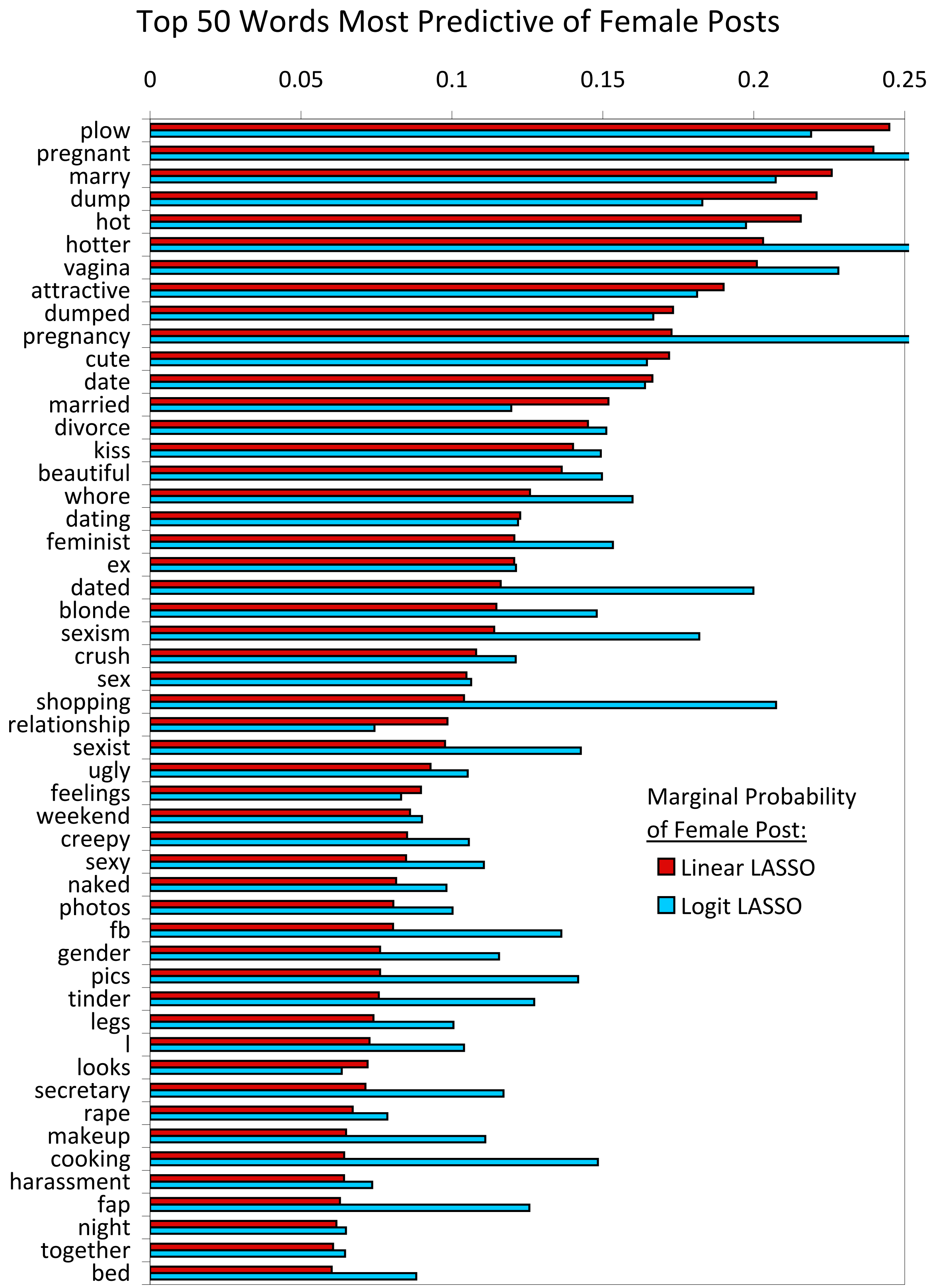
And the marginal effect of word k on the probability that a post is *Female* is estimated by $\hat{\beta}_\lambda^k$, the coefficient on the regressor w_i^k.

APPENDIX FIGURE 1: Selection of Optimal P-score Cutoff by Mean Squared Error (Lasso-Logistic model on gendered posts identified by the comprehensive list of classifiers.)



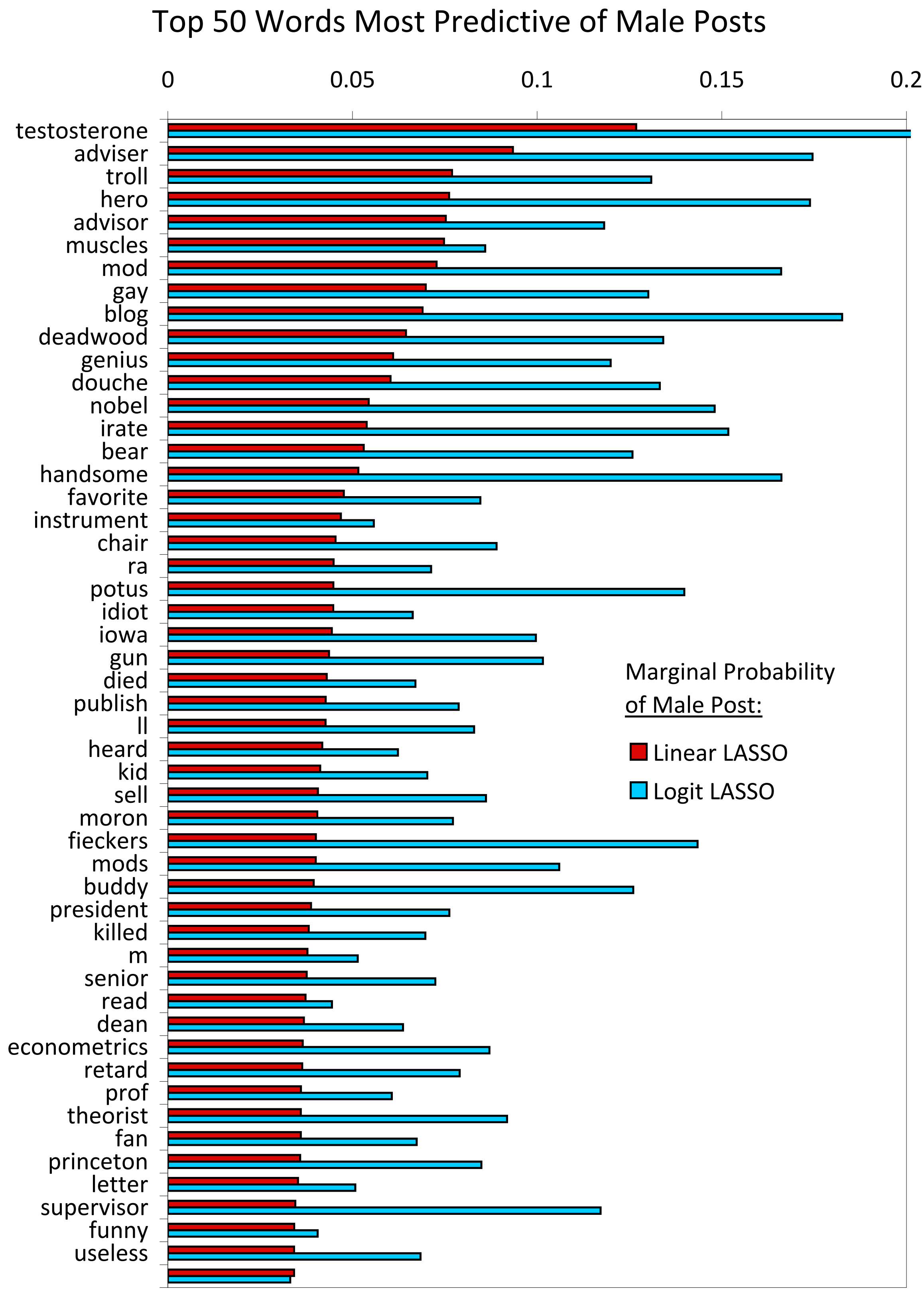
Note: This figure shows the mean squared error (MSE) for predicting gender on the test set of $99,941$ gendered posts (a left-out 25% sample) that include only female or only male classifiers from the comprehensive list, at each p-score threshold for assigning a post to *Female* that range from 0.15 to 0.85 with a step size of 0.05. The MSE is minimized at $p = 0.40$. Therefore, I use 0.40 as the threshold to assign genders for $44,081$ posts that include both female and male classifiers in the comprehensive list. As a result, $14,028$ (31.82%) posts are re-classified to *Female*, and the rest to *Male*.

APPENDIX FIGURE 2: Word Selection by Lasso-Logistic vs. Lasso-Linear (Pronoun Sample)



Notes: Each model was trained on 35,850 Female posts and 103,449 Male posts identified by gender pronouns (pronoun sample). The top 50 words above are sorted by the marginal effect of each word estimated by the Linear LASSO model.

APPENDIX FIGURE 3: Word Selection by Lasso-Logistic vs. Lasso-Linear (Pronoun Sample)



Notes: Each model was trained on 35,850 Female posts and 103,449 Male posts identified by gender pronouns (pronoun sample). The top 50 words above are sorted by the marginal effect of each word estimated by the Linear LASSO model.

APPENDIX TABLE 1: Top 50 *female*(*male*) Words Selected by Lasso-Logistic (Gendered posts are identified by the comprehensive list of classifiers)

Most *female*		Most *male*	
Word	Marginal Effect	Word	Marginal Effect
hotter	0.422	homo	-0.303
pregnant	0.323	testosterone	-0.195
plow	0.277	chapters	-0.189
marry	0.275	satisfaction	-0.187
hot	0.271	fieckers	-0.181
marrying	0.260	macroeconomics	-0.180
pregnancy	0.254	cuny	-0.180
attractive	0.245	thrust	-0.169
beautiful	0.240	nk	-0.165
breast	0.227	macro	-0.163
dumped	0.225	fenance	-0.162
kissed	0.224	founding	-0.160
misogynistic	0.222	blog	-0.157
feminist	0.218	mountains	-0.156
sexism	0.210	grown	-0.156
dated	0.209	frat	-0.155
whore	0.208	handsome	-0.154
sexy	0.202	nba	-0.151
raped	0.200	lyrics	-0.151
attracted	0.198	ferguson	-0.150
slept	0.195	wasn	-0.147
blonde	0.193	supervisor	-0.146
unattractive	0.193	rfs	-0.145
gorgeous	0.192	adviser	-0.141
assaulted	0.191	minnesota	-0.140
cute	0.185	hero	-0.136
vagina	0.184	gay	-0.135
date	0.181	puerto	-0.134
dating	0.181	nobel	-0.129
ugly	0.181	keynesian	-0.128
naked	0.181	sincerely	-0.126
classified	0.179	bashing	-0.126
workforce	0.175	thanks	-0.123
banging	0.175	fiekers	-0.121
impress	0.169	homosexual	-0.121
beauty	0.169	bowl	-0.121
divorce	0.164	nordic	-0.119
feminism	0.164	disability	-0.119
crush	0.163	advised	-0.119
teenage	0.162	inflation	-0.118
dig	0.161	gray	-0.117
sexist	0.160	depth	-0.117
makeup	0.159	wolf	-0.117
cleaning	0.155	curry	-0.116
dump	0.155	teenagers	-0.116
victoria	0.150	wash	-0.116
instagram	0.150	genius	-0.116
tinder	0.149	argues	-0.114
fiecking	0.149	coase	-0.113
shopping	0.149	rip	-0.113

Notes: The top 50 *female* (*male*) words are sorted in descending (ascending) order of their marginal effect - the increase in the probability that the subject of a post is *Female* given an additional occurrence of each word. The model was trained on gendered posts identified by the comprehensive list of gender classifiers.

APPENDIX TABLE 2: Number of Posts that Contain Each of the Top 50 *female*(*male*) Words Selected by Lasso-Logistic (Gendered posts are identified by the comprehensive list of classifiers)

Most *female*			Most *male*		
Word	No. *Female*	No. *Male*	Word	No. *Female*	No. *Male*
hotter	307	31	homo	48	715
pregnant	564	120	testosterone	51	102
plow	274	83	chapters	9	361
marry	1,287	258	satisfaction	59	145
hot	3,613	1,053	fieckers	49	604
marrying	262	49	macroeconomics	19	850
pregnancy	202	61	cuny	8	248
attractive	1,578	417	thrust	6	47
beautiful	1,419	610	nk	3	260
breast	134	48	macro	178	4,282
dumped	361	100	fenance	46	640
kissed	218	50	founding	6	186
misogynistic	66	48	blog	109	1,839
feminist	422	234	mountains	14	90
sexism	269	171	grown	69	394
dated	362	148	frat	59	290
whore	239	148	handsome	103	323
sexy	430	207	nba	16	301
raped	297	155	lyrics	17	111
attracted	415	182	ferguson	10	221
slept	368	85	wasn	32	171
blonde	292	79	supervisor	40	273
unattractive	172	32	rfs	7	284
gorgeous	213	78	adviser	78	712
assaulted	98	52	minnesota	35	703
cute	912	488	hero	47	579
vagina	199	68	gay	406	1,755
date	1,729	835	puerto	7	101
dating	1,423	399	nobel	204	3,379
ugly	1,046	404	keynesian	8	567
naked	376	213	sincerely	55	520
classified	47	96	bashing	15	199
workforce	78	92	thanks	655	4,999
banging	306	109	fiekers	44	406
impress	160	164	homosexual	33	169
beauty	330	193	bowl	24	203
divorce	673	192	nordic	103	537
feminism	264	127	disability	27	117
crush	320	207	advised	33	227
teenage	168	116	inflation	41	1,000
dig	152	176	gray	34	108
sexist	469	358	depth	17	257
makeup	174	66	wolf	19	144
cleaning	175	169	curry	12	143
dump	503	339	teenagers	36	113
victoria	40	49	wash	74	204
instagram	100	63	genius	92	1,007
tinder	301	110	argues	23	313
fiecking	377	226	coase	7	200
shopping	165	129	rip	56	484

Notes: This table shows the number of *Female* posts and the number of *Male* posts that contain each of the top 50 *female* or *male* terms selected by Lasso, in the same order as in Appendix Table 1. Using the comprehensive list of gender classifiers, I identified 103,584 *Female* posts and 341,226 *Male* posts.

APPENDIX TABLE 3: Most Frequent Words in *Female* (*Male*) posts, identified by the comprehensive list of classifiers

Most common in *Female*			Most common in *Male*		
Word	No. *Female*	No. *Male*	Word	No. *Female*	No. *Male*
life	4,034	7,644	work	3,800	13,989
work	3,800	13,989	paper	1,503	11,727
hot	3,613	1,053	job	3,091	10,313
love	3,297	4,274	economics	1,120	9,808
sex	3,103	1,535	great	2,323	9,181
job	3,091	10,313	best	2,558	8,552
feel	2,574	5,167	research	1,407	8,238
best	2,558	8,552	school	2,446	8,228
school	2,446	8,228	market	1,750	7,954
kids	2,441	2,200	life	4,034	7,644
great	2,323	9,181	phd	1,751	7,295
married	2,231	1,207	papers	854	7,177
friends	2,048	2,504	econ	1,133	6,950
nice	1,978	4,590	students	1,474	6,889
money	1,951	6,011	theory	415	6,347
home	1,778	2,734	money	1,951	6,011
phd	1,751	7,295	data	729	5,648
market	1,750	7,954	student	1,560	5,607
date	1,729	835	economist	855	5,539
family	1,653	2,685	wrong	1,344	5,487
attractive	1,578	417	economists	697	5,461
student	1,560	5,607	course	1,320	5,416
relationship	1,506	1,169	question	1,109	5,257
paper	1,503	11,727	idea	1,158	5,184
students	1,474	6,889	feel	2,574	5,167
happy	1,452	2,536	economic	466	5,152
dating	1,423	399	department	935	4,985
beautiful	1,419	610	university	955	4,970
friend	1,412	2,423	r	682	4,774
research	1,407	8,238	nice	1,978	4,590
single	1,373	2,578	finance	357	4,469
wrong	1,344	5,487	working	1,282	4,465
children	1,337	1,449	field	547	4,339
course	1,320	5,416	policy	504	4,330
young	1,315	2,751	macro	178	4,282
marry	1,287	258	love	3,297	4,274
working	1,282	4,465	model	463	4,210
social	1,257	3,590	tenure	930	3,891
fat	1,237	1,170	public	820	3,877
aspie	1,235	1,412	journal	324	3,787
idea	1,158	5,184	professor	679	3,781
marriage	1,150	614	class	1,115	3,614
age	1,142	1,881	social	1,257	3,590
econ	1,133	6,950	harvard	418	3,533
economics	1,120	9,808	business	546	3,478
class	1,115	3,614	math	394	3,421
question	1,109	5,257	offer	777	3,401
college	1,095	2,651	nobel	204	3,379
ugly	1,046	404	able	979	3,320
experience	1,043	2,876	academic	654	3,280

Notes: The words that are most common in *Female* (*Male*) are sorted by the number of *Female* (*Male*) posts they appear in. Using the comprehensive list of gender classifiers, I identified 103,584 *Female* posts and 341,226 *Male* posts.

APPENDIX TABLE 4: Top 50 *female*(*male*) Words Selected by Lasso-Logistic (Gendered posts are identified by pronouns only)

Most *female*		Most *male*	
Word	Marginal Effect	Word	Marginal Effect
pregnancy	0.292	knocking	-0.329
hotter	0.289	testosterone	-0.204
pregnant	0.258	blog	-0.183
hp	0.238	hateukbro	-0.176
vagina	0.228	adviser	-0.175
breast	0.220	hero	-0.174
plow	0.219	cuny	-0.173
shopping	0.207	handsome	-0.166
marry	0.207	mod	-0.166
gorgeous	0.201	homo	-0.160
dated	0.200	rfs	-0.154
marrying	0.198	irate	-0.152
hot	0.197	nobel	-0.148
dump	0.183	dictator	-0.144
sexism	0.182	fieckers	-0.143
attractive	0.181	spell	-0.143
sperm	0.171	potus	-0.140
dumped	0.167	nk	-0.137
intimate	0.167	repec	-0.137
cute	0.165	minnesota	-0.135
date	0.164	advising	-0.135
whore	0.160	deadwood	-0.134
commonly	0.159	ego	-0.133
commodities	0.159	douche	-0.133
consent	0.153	punch	-0.131
feminist	0.153	troll	-0.131
classified	0.152	gay	-0.130
divorce	0.151	gays	-0.129
beautiful	0.150	beard	-0.127
kiss	0.149	writings	-0.127
victoria	0.149	blanket	-0.127
cooking	0.148	bowl	-0.127
blonde	0.148	buddy	-0.126
yoga	0.147	bear	-0.126
oct	0.144	ferguson	-0.125
sexist	0.143	legend	-0.124
pics	0.142	assumes	-0.123
university's	0.140	westerners	-0.123
improvements	0.140	rip	-0.121
fb	0.136	sins	-0.120
aej	0.136	genius	-0.120
yahoo	0.134	evolution	-0.119
cum	0.133	advisor	-0.118
rct	0.133	supervisor	-0.117
activist	0.133	calculus	-0.117
flirting	0.132	goals	-0.116
feminism	0.129	decency	-0.116
tinder	0.127	penalty	-0.116
flowers	0.126	injured	-0.113
instagram	0.126	depth	-0.113

Notes: The top 50 *female* (*male*) words are sorted in descending (ascending) order of their marginal effect - the increase in the probability that the subject of a post is *Female* given an additional occurrence of each word. The model was trained on gendered posts identified by feminine or masculine pronouns only.

APPENDIX TABLE 5: Number of Posts that Contain Each of the Top 50 *female*(*male*) Words Selected by Lasso-Logistic (Gendered posts are identified by pronouns only)

Most *female*			Most *male*		
Word	No. *Female*	No. *Male*	Word	No. *Female*	No. *Male*
pregnancy	106	27	knocking	6	82
hotter	120	31	testosterone	15	31
pregnant	270	98	blog	89	1,244
hp	26	14	hateukbro	0	70
vagina	137	41	adviser	66	591
breast	62	30	hero	32	412
plow	146	60	cuny	3	104
shopping	99	69	handsome	41	170
marry	557	191	mod	30	384
gorgeous	110	41	homo	31	162
dated	194	86	rfs	5	137
marrying	117	34	irate	24	235
hot	1,309	658	nobel	125	1,944
dump	369	215	dictator	6	167
sexism	87	76	fieckers	20	201
attractive	547	246	spell	26	127
sperm	46	22	potus	20	202
dumped	240	88	nk	0	119
intimate	49	27	repec	8	176
cute	463	298	minnesota	21	282
date	902	477	advising	10	193
whore	123	112	deadwood	38	426
commonly	25	57	ego	47	245
commodities	12	28	douche	48	288
consent	83	62	punch	25	153
feminist	162	128	troll	206	1,606
classified	33	56	gay	163	737
divorce	376	147	gays	5	78
beautiful	524	346	beard	14	99
kiss	308	148	writings	3	157
victoria	18	17	blanket	9	55
cooking	72	44	bowl	14	104
blonde	155	51	buddy	50	193
yoga	53	38	bear	104	736
oct	23	143	ferguson	10	126
sexist	145	161	legend	13	117
pics	121	77	assumes	12	139
university's	30	68	westerners	5	39
improvements	14	43	rip	33	218
fb	120	84	sins	5	88
aej	34	82	genius	50	650
yahoo	23	42	evolution	15	152
cum	67	60	advisor	286	2,145
rct	15	26	supervisor	34	199
activist	41	89	calculus	10	246
flirting	103	27	goals	38	304
feminism	68	56	decency	5	64
tinder	106	34	penalty	14	170
flowers	73	42	injured	9	158
instagram	65	39	depth	11	143

Notes: This table shows the number of *Female* posts and the number of *Male* posts that contain each of the top 50 *female* or *male* terms selected by Lasso, in the same order as in Appendix Table 4. Using gender pronouns, I identified 49,993 *Female* posts and 145,382 *Male* posts.

APPENDIX TABLE 6: Most Frequent Words in *Female* (*Male*) posts, identified by pronouns only

Most common in *Female*			Most common in *Male*		
Word	No. *Female*	No. *Male*	Word	No. *Female*	No. *Male*
work	2,227	8,018	work	2,227	8,018
life	2,017	4,133	paper	1,030	6,500
love	1,762	2,055	job	1,609	5,517
job	1,609	5,517	great	1,371	4,840
feel	1,523	2,339	economics	640	4,696
sex	1,377	831	best	1,320	4,423
great	1,371	4,840	school	1,334	4,314
school	1,334	4,314	research	828	4,270
best	1,320	4,423	papers	592	4,194
hot	1,309	658	life	2,017	4,133
married	1,116	678	students	766	3,867
student	1,109	3,781	phd	968	3,837
friends	1,088	1,459	student	1,109	3,781
nice	1,055	2,412	market	702	3,706
paper	1,030	6,500	economist	542	3,345
kids	1,009	1,236	money	975	3,307
home	989	1,562	course	769	3,146
money	975	3,307	wrong	827	3,144
friend	974	1,951	idea	702	3,009
phd	968	3,837	department	620	2,926
date	902	477	econ	587	2,820
relationship	880	644	theory	258	2,787
family	863	1,601	question	619	2,717
happy	850	1,334	professor	485	2,578
research	828	4,270	university	640	2,533
wrong	827	3,144	economists	338	2,482
course	769	3,146	tenure	618	2,462
students	766	3,867	working	702	2,449
market	702	3,706	nice	1,055	2,412
working	702	2,449	economic	257	2,376
idea	702	3,009	feel	1,523	2,339
economics	640	4,696	data	341	2,278
university	640	2,533	field	324	2,225
department	620	2,926	advisor	286	2,145
question	619	2,717	class	606	2,106
tenure	618	2,462	offer	522	2,097
class	606	2,106	public	488	2,077
couple	598	1,300	policy	309	2,069
papers	592	4,194	love	1,762	2,055
econ	587	2,820	journal	219	1,987
mind	583	1,607	friend	974	1,951
marriage	580	368	able	542	1,950
dating	573	242	nobel	125	1,944
marry	557	191	r	363	1,933
young	547	1,498	published	282	1,930
attractive	547	246	smart	535	1,904
economist	542	3,345	editor	201	1,837
able	542	1,950	stupid	456	1,822
social	538	1,760	academic	378	1,801
smart	535	1,904	social	538	1,760

Notes: The words that are Most common in *Female* (*Male*) are sorted by the number of *Female* (*Male*) posts they appear in. Using gender pronouns, I identified 49,993 *Female* posts and 145,382 *Male* posts.