Replication of Wu (2018): Gendered Language on the Economics Job Market Rumors Forum

Alejandro De La Torre

February 11, 2026

1 Introduction and motivation

Wu (2018) analyzes how women and men are described on the Economics Job Market Rumors (EJMR) forum using a Lasso logistic model to identify words most predictive of female- versus male-referenced posts [Wu, 2018a]. The paper documents a pattern of female-associated language that is disproportionately about appearance and personal information, while male-associated language is more professional in tone. Replicating this result matters because it tests whether the pattern is robust in a large, noisy, anonymous forum. Replication materials and code for this project are public, and I make my full pipeline and outputs available in a GitHub repository.[1]

2 Data and replication package

The OpenICPSR replication package provides the raw EJMR post-level dataset, a 10,000-word vocabulary, a sparse word-count matrix, and the upstream Python and R scripts used to generate the tables and figure in the paper [Wu, 2018b]. The paper reports 2,217,046 scraped posts from October 2013 to October 2017 and 444,810 gendered posts identified by gender classifiers [Wu, 2018a]. The package includes:

- `gendered_posts.csv` with post text, gender labels, and training/test splits.
- `X_word_count.npz`, a sparse matrix of word counts for the top 10,000 words.
- `vocab10K.csv` with the vocabulary and model outputs.
- `trend_stats.csv` used for Figure 1.

3 Replication procedure

I built a single-command pipeline that runs the upstream Python Lasso scripts and then the upstream R script to recreate the tables and figure. The pipeline is executed with:

```
python src/run_all.py
```

The script runs the two Lasso logit programs in the raw package directory so the relative paths resolve correctly, then captures intermediates and logs. The Lasso scripts print the matrix shapes; for the full sample, the logged shapes are (300,788 × 9,540) for training, (99,941 × 9,540) for the

[1] `https://github.com/adelatorre2/wu2018-gendered-language-replication`

held-out test set, and (44,081 × 9,540) for posts containing both genders. The pronoun sample logs analogous sizes of (139,299 × 9,540), (46,502 × 9,540), and (9,574 × 9,540). The R step uses `tables-figures.R` to produce Table 1, Table 2, and Figure 1 as CSV/PDF outputs.

All runs were executed on macOS arm64 under the conda environment `wu2018` with Python 3.11 and R + ggplot2. Detailed logs are saved in `output/logs/`.

4 Results

Table 1 and Table 2 reproduce the top 10 words most predictive of female- and male-referenced posts in the full and pronoun samples, respectively. The qualitative pattern matches the paper: female-associated words emphasize appearance or personal attributes, while male-associated words emphasize professional status or roles. Figure 1 reproduces the monthly fraction of posts containing top gendered words, mirroring the paper's trend plot.

Female word	ME	Male word	ME
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

Table 1: Top 10 words most predictive of female- and male-referenced posts (full sample).

Female word	ME (pronoun)	Male word	ME (pronoun)
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

Table 2: Top 10 words most predictive of female- and male-referenced posts (pronoun sample).

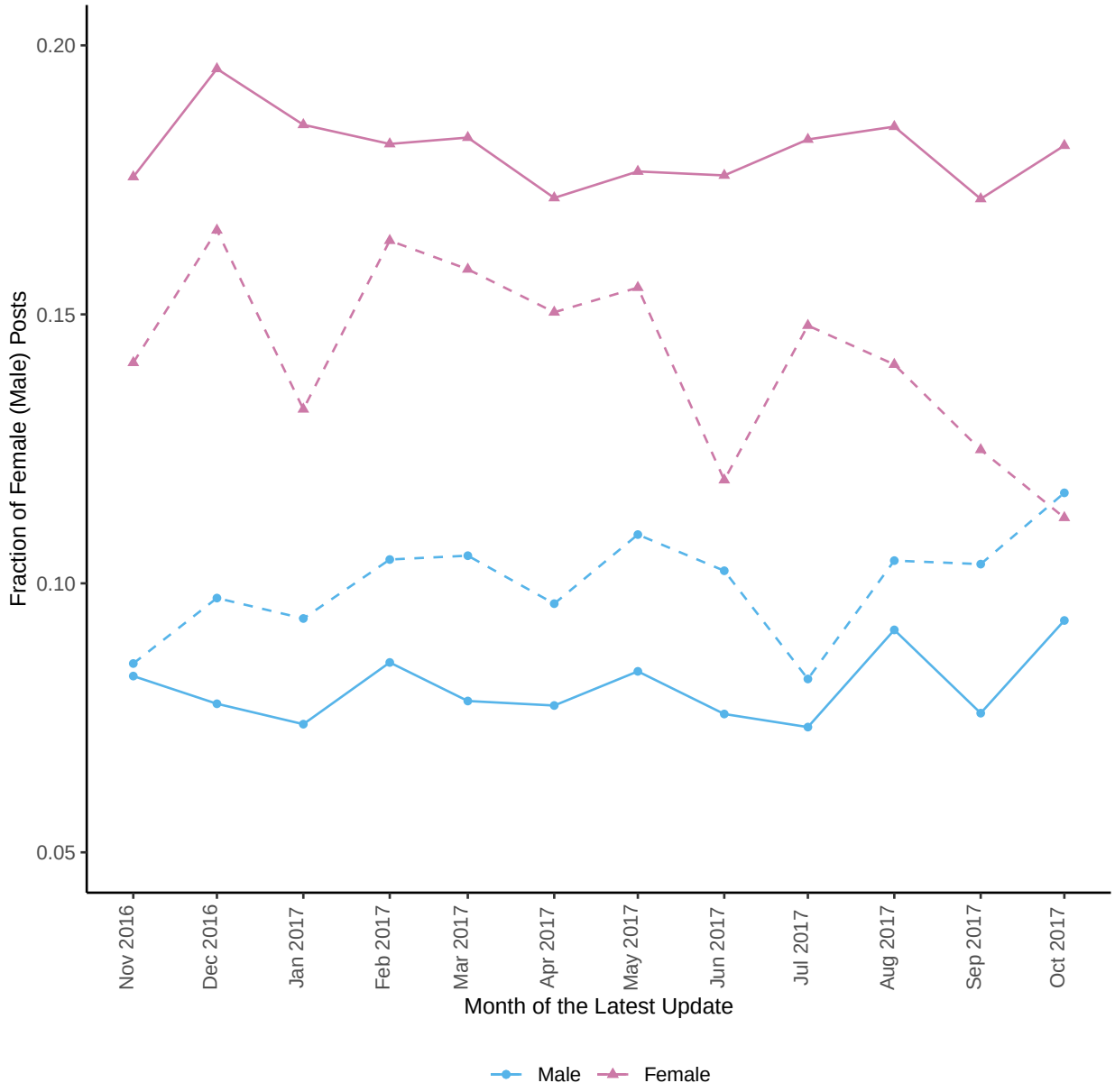

Figure 1: Fraction of female and male posts containing any of the top 50 gendered words.

5 Modest extension: overlap across samples

As a small extension, I compare overlap in the top 50 positive (female-associated) and top 50 negative (male-associated) coefficients between the full-sample and pronoun-sample Lasso models. Table 3 reports overlap counts and Jaccard similarity. The overlap is moderate: 29 of 50 female-associated words and 18 of 50 male-associated words appear in both lists, suggesting that the strongest signals are fairly stable across labeling schemes while still leaving room for sampling variation.

Group	N	Overlap	Jaccard	Example overlap
Female-associated (pos.)	50	29	0.408	attractive, beautiful, blonde, breast
Male-associated (neg.)	50	18	0.22	adviser, blog, bowl, cuny

Table 3: Overlap of top 50 words between full-sample and pronoun-sample Lasso models.

6 Obstacles and fixes

Two compatibility issues required small, non-substantive patches to the upstream code. First, deprecated pandas `as_matrix()` calls were replaced with `to_numpy()`. Second, NumPy's default behavior now blocks loading pickled objects in `np.load`, so I added `allow_pickle=True` when reading the sparse word-count matrix. These changes only restore compatibility with modern libraries; the analysis logic is unchanged.

7 Conclusion

The replication successfully reproduces Table 1, Table 2, and Figure 1 from Wu (2018) using the OpenICPSR package and a single-command pipeline. The extension indicates a nontrivial overlap in the most predictive words across labeling schemes, reinforcing the qualitative pattern of gendered language on EJMR. Future work could implement the assignment's alternative model requirement (e.g., random forest or linear probability model) within the same pipeline.

References

Alice H. Wu. Gendered language on the economics job market rumors forum. *AEA Papers and Proceedings*, 108:175–179, 2018a. doi: 10.1257/pandp.20181101.

Alice H. Wu. Replication data for: Gendered language on the economics job market rumors forum. OpenICPSR, 2018b. URL `https://www.openicpsr.org/openicpsr/project/114486/version/V1/view`.