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

Alejandro De La Torre [*]

February 13, 2026

1 Overview

Wu (2018) studies gendered language in anonymous posts on the Economics Job Market Rumors (EJMR) forum. The paper frames the exercise as predictive: given a post's word usage, which terms are most informative about whether the post refers to a woman or a man? The headline takeaway is qualitative rather than purely statistical—female-referenced posts are disproportionately associated with words about appearance and personal traits, while male-referenced posts are more associated with professional roles and status.

This report meets the assignment requirements in three steps. First, I locate and run the author-provided replication package and reproduce Table 1, Table 2, and Figure 1 from the published article. Second, I document the engineering work required to run the archived code on a modern Python/R stack, without altering the original analysis logic. Third, I implement a modest extension by re-estimating the core predictive exercise using alternative models and assessing how sensitive the ranked word lists are to modeling choices.

2 Replication

2.1 Data and replication package

I use the OpenICPSR replication package linked from the paper and its online appendix (Wu, 2019, Wu). The key inputs are (i) `gendered_posts.csv` with post-level labels and split indicators, (ii) `X_word_count.npz`, a sparse word-count matrix for the top 10,000 words, (iii) `vocab10K.csv` with the vocabulary and excluded terms, and (iv) `trend_stats.csv` for Figure 1. These files define the same design matrix, labels, and train/test split used by the original scripts.

2.2 Replication procedure

I built a single-command pipeline that runs the upstream Python Lasso scripts and the upstream R script for tables and figures:

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

The pipeline executes the two Lasso-logit programs in the raw package directory so relative paths resolve, then runs `tables-figures.R` to produce Table 1, Table 2, and Figure 1. All runs were

[*] Replication repository: `github.com/adelatorre2/wu2018-gendered-language-replication`.

executed on macOS arm64 in a conda environment (`wu2018`) with Python 3.11 and R + ggplot2. Logs are saved in `output/logs/`.

Conceptually, this mirrors the original workflow: estimate a Lasso-logit model of the female indicator on a high-dimensional word-count matrix, select the penalty via cross-validation, and rank words by their marginal effect on predicted probabilities. I do not modify the estimation routine, feature construction, or split logic; the aim is computational fidelity.

2.3 Obstacles and fixes

Two compatibility issues required minor, non-substantive patches. First, deprecated pandas `as_matrix()` calls were replaced with `to_numpy()`. Second, NumPy now blocks loading pickled objects in `np.load` by default, so I added `allow_pickle=True` for the sparse word-count matrix. These changes restore compatibility but do not alter analysis logic.

3 Results

Table 1, Table 2, and Figure 1 reproduce the original results in Wu (2018). The word lists and trends match the published paper up to rounding, which is expected because I use the same data, preprocessing, splits, and hyperparameters. Because the replication package fixes the design matrix, split indicators, and estimation routine, exact agreement up to rounding is expected when running the archived code on the same dataset. The qualitative pattern is identical: female-associated words emphasize appearance or personal traits, while male-associated words emphasize professional roles.

The mechanical equivalence of these results reflects the fact that the same design matrix, split, and random seeds are used. The value of the replication lies not in reinterpreting the estimates, but in verifying that the archived materials regenerate the published objects in a modern computing environment.

Most female		Most male	
Word	ME	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).

Most female		Most male	
Word	ME	Word	ME
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).

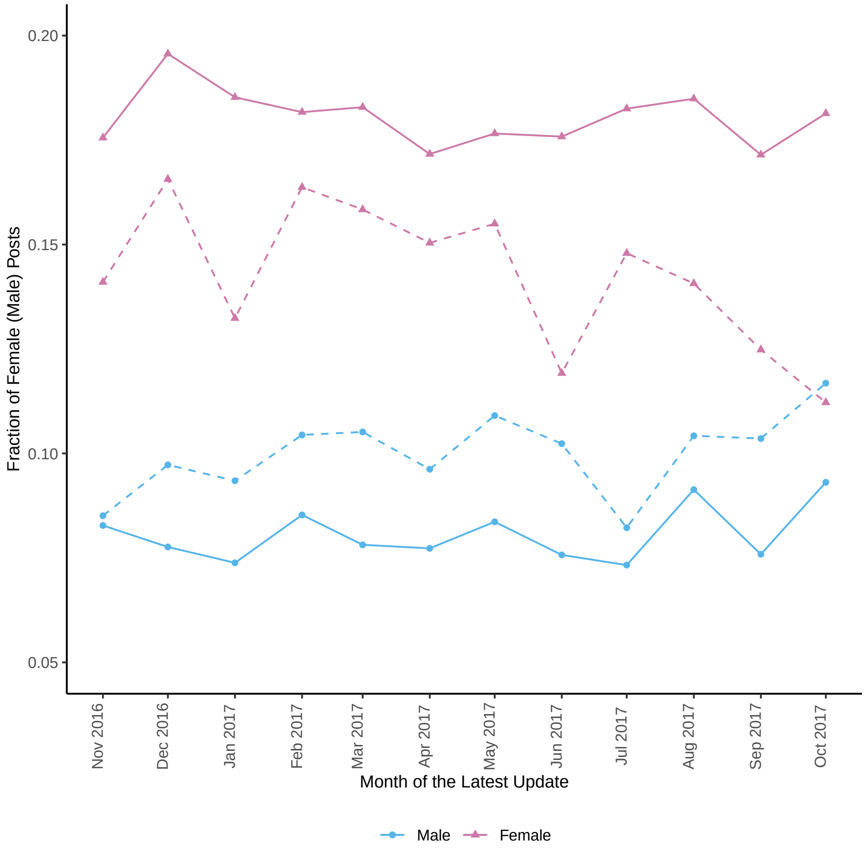

Figure 1: Fraction of female and male posts containing any of the top 50 gendered words.

Figure 1 reproduces the time-series pattern documented in the paper: posts containing the top gender-predictive words differ systematically by referenced gender, and the gap persists over time.

4 Modest extension

The assignment requires a modest extension beyond mechanical replication. Rather than altering the data or redefining the labeling strategy, I focus on the core predictive exercise underlying Table 1: how sensitive are the ranked word lists to the choice of model? Concretely, I re-estimate the female-indicator prediction problem using two alternative approaches while holding fixed the original design matrix, training/test split, and vocabulary. By keeping the data environment identical to the replication, any differences in results can be attributed to modeling choices rather than preprocessing decisions.

This extension is intentionally conservative. The goal is not to "improve" on the original paper, but to assess whether its qualitative conclusions depend heavily on the specific use of Lasso-logit.

4.1 Methodological setup

The original analysis uses a Lasso-logit model estimated on a 10,000-word sparse matrix. Lasso is well suited to high-dimensional text data: it performs regularization, selects a sparse set of predictors, and yields coefficients that can be directly ranked and interpreted. This makes it particularly aligned with the descriptive aim of the paper, which is to identify which words are most predictive of gender reference.

I implement two alternative models.

First, I estimate a Linear Probability Model (LPM) using OLS. The dependent variable is the same female indicator, and the regressors are the identical word-count features used in the Lasso specification. The appeal of the LPM is transparency: coefficients can be interpreted as linear changes in predicted probability associated with marginal changes in word usage. Although fitted values can fall outside the unit interval, this does not affect coefficient interpretation, and for classification metrics I clip predictions to $[0, 1]$ while computing ROC-AUC using raw scores. This model serves as a simple, easily interpretable benchmark.

Second, I estimate a Random Forest classifier (Breiman, 2001). The motivation is to allow nonlinearities and interactions across words that a linear model cannot capture. Because the full 10,000-word feature space is computationally intensive for tree-based methods, I restrict the feature set to the 1,000 most frequent words and subsample the training and test sets for feasibility (holding a fixed random seed for reproducibility). Variable importance is measured using Gini importance, and directionality is inferred from mean usage differences between classes.

Together, these models span a spectrum: LPM emphasizes interpretability without regularization, Random Forest emphasizes flexibility with reduced interpretability, and Lasso-logit balances sparsity and interpretability in high dimensions.

4.2 Results

The LPM delivers out-of-sample performance that is broadly comparable to the Lasso benchmark. Test-set accuracy is 0.792, with ROC-AUC of 0.734. Precision is relatively high while recall remains low, reflecting the class imbalance and the conservative thresholding common in sparse text classification. Most importantly, the highest-ranked words by absolute coefficient magnitude remain thematically similar to those in Table 1: appearance- and personal-trait-related terms are most predictive of female-referenced posts, while professional-role and status terms are most predictive of male-referenced posts.

The Random Forest produces slightly lower ROC-AUC (0.661) and similar accuracy (0.783), with more variation in precision and recall. Because the RF is fit on a reduced feature space (top

1,000 words) and subsamples for feasibility, its ranked list is intended for qualitative comparison rather than a one-for-one match to the 10,000-word Lasso ranking. Rankings of "important" words are somewhat more sensitive to feature frequency and sampling. Nevertheless, the broad qualitative pattern persists. Even when allowing nonlinear splits and interactions, the dominant predictors cluster around the same thematic categories documented in the original paper.

In this sense, the extension reinforces rather than overturns the main descriptive finding: the association between gender reference and the type of language used is not an artifact of a single linear-regularized specification.

4.3 Interpretation and limitations

This extension remains predictive, not causal. The dependent variable is itself generated by gender-classification rules, and any misclassification propagates through all models. The LPM ignores heteroskedasticity inherent in binary outcomes, and the Random Forest requires feature restriction and subsampling, introducing additional researcher discretion. Moreover, Gini importance is known to favor more frequent predictors, which may bias importance rankings toward common words.

From this, it makes sense why Wu (2018) relied on Lasso-logit. In high-dimensional sparse settings, Lasso provides a disciplined and transparent method for variable selection while preserving interpretability of coefficient rankings. More flexible models introduce additional complexity and tuning decisions without materially changing the qualitative conclusions.

Accordingly, this goes to show the robustness of the paper's central descriptive claim while illustrating the tradeoffs between interpretability, flexibility, and regularization in text-based prediction exercises. I will also note that his extension does not pursue a structural interpretation of bias, it relates indirectly to subsequent work such as Wu (2020), which develops an identity-based interpretation of gender bias among professionals. The credibility of such interpretive advances rests in part on the stability of the underlying descriptive patterns. By showing that the core gender-language associations are not highly sensitive to modeling choice, this exercise reinforces the empirical foundation upon which later theoretical and interpretive contributions are built.

5 Conclusion

This replication exercise underscores how much of empirical credibility rests not only on statistical technique, but on transparency of process. I was fortunate to work with a recent paper accompanied by a thorough and carefully documented replication package. Because the data, scripts, and intermediate objects were clearly organized, reproducing Table 1, Table 2, and Figure 1 required only minor updates to accommodate software deprecations. The fact that these updates were small—and did not alter the analytical logic—speaks to the durability of well-documented computational research.

The experience also highlights why Wu (2018) has had such influence. The paper itself is concise, but every empirical claim is tightly linked to transparent modeling choices, and the accompanying appendix and code anticipate many natural questions a reader might raise. This combination of clarity, reproducibility, and methodological discipline strengthens the persuasive force of the results. It also helps explain why subsequent work in this literature builds directly on Wu's findings: other researchers can verify, extend, and interrogate the analysis rather than treating it as a black box.

More broadly, this project reinforces a professional lesson. Replication packages are not peripheral add-ons; they are integral to modern empirical economics. They allow others to "show the work," audit decisions, and distinguish substantive findings from artifacts of implementation. As I develop my own research, I aim to document design choices, preprocessing steps, and modeling decisions with similar care. Doing so not only improves external credibility but sharpens one's own thinking about identification, measurement, and robustness.

At the same time, the relative ease of replicating a recent paper makes me more aware of the fragility of older empirical work that lacks accessible data or code. Revisiting, updating, and stress-testing such papers may be a productive avenue for early-career researchers. Replication, in this sense, is not merely a pedagogical exercise but a mechanism for cumulative knowledge. This project therefore serves both as confirmation of Wu's central descriptive finding and as a reminder that rigorous, well-documented empirical practice is itself a substantive contribution to the discipline. In short, when I grow up, I want to be like Professor Wu.

References

Wu, A. H. Online Appendix to "Gendered Language on the Economics Job Market Rumors Forum".

Wu, A. H. (2018, May). Gendered Language on the Economics Job Market Rumors Forum. *AEA Papers and Proceedings 108*, 175–179.

Wu, A. H. (2019, October). Replication data for: Gendered Language on the Economics Job Market Rumors Forum.

Wu, A. H. (2020, December). Gender Bias among Professionals: An Identity-Based Interpretation. *The Review of Economics and Statistics 102*(5), 867–880.