Replication Exercise: Gendered Language on the Economics Job Market Rumors Forum by Alice H. Wu

Lucas Sernik

February 13, 2026

Introduction

Wu (2018) wanted to evaluate how "unwelcoming or stereotypical" is the Economics academic field towards women. To prove this she collected posts from several threads from the Economics Job Market Rumors Forum (EJMR), where users make anonymous posts regarding recent job interviews, interactions, and experiences related to the field. Her idea was that anonimity provided a "safe space" for people to reveal their true thoughts and beliefs, which would expose their prejudice against women.

Data and Methods

Wu (2018) scraped 2,217,046 posts across 223,475 threads on EJMR made between 2013 and 2017. After identifying the top 10,000 words from the raw text, she made a comprehensive list of gender classifiers, with 57 female classifiers, like "she" and "woman", and 236 male classifiers, like "he" and "man". A post with any female classifier is a *Female* post and any post with a male classifier is a *Male* post. From doing this, she filtered 444,810 gendered posts from the initial 2 million.

She described her hypothesis to be that, if such an unwelcoming environment exists, EJMR users will use terms that praises a man's professional accomplishments and status, while diminishing a women's achievements by commenting about something else.

The section of the original paper that is specially important for this replication exercise and the extension is in **Section 2**, where the **Lasso Logistic Model** is described. Let $\boldsymbol{w}_i$ be a vector with the number of appearance of the most common words (excluding gender classifiers) that occur in a given post i. The Lasso-regularized logistic model that evaluates the probability that the post is *Female* is:

$$\hat{\theta}_\lambda = \text{arg min}_\theta \left[-\log(\Pi_{i=1}^{N} P(\mathit{Female}|\boldsymbol{w}_i)) + \lambda||\theta||_1 \right] \text{, where}$$
$$||\theta||_1 = \Sigma_{j\geq 1}|\theta^j|$$

To train the model, she used a 75 percent random sample and selected the optimal tuning parameter λ through 5-fold cross validation.

Replication experience

This was one of the papers with a really good replication package among the ones that we could choose. My experience with this replication package was smooth.

I first searched the author's website, looking for the paper itself. After finding it, I read and understood the whole paper and took some time to read the online appendix available in that version.

Then, I looked for the replication package that would allow me to recreate the model. While not available on the author's website, after getting institution acces through UW-Madison to the American Economic Association database, I was able to find the replication package that was submitted with the paper.

First, I read the README file, which was a nice PDF saying what each variable meant, what was in each of the excel files, and contained instructions on which files to run to get specific results from each part of the paper.

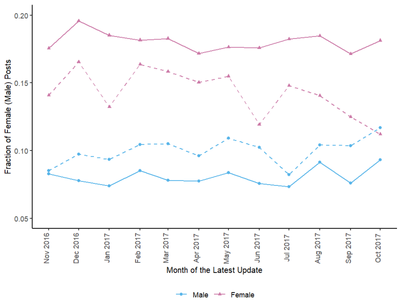
In order to replicate the images, I simply ran the .R code in the package, which creates the data presented in the tables and the graphs Figure 1. Here, I had to figure out some extra libraries that I had to download to my R, and an extra package to create a LaTeX table.

Table 1: Top 10 Words Most Predictive of Female/Male

Most *female*		Most *male*	
word	ME	word	ME
hotter	0.42	homo	-0.30
pregnant	0.32	testosterone	-0.20
plow	0.28	chapters	-0.19
marry	0.28	satisfaction	-0.19
hot	0.27	fieckers	-0.18
marrying	0.26	macroeconomics	-0.18
pregnancy	0.25	cuny	-0.18
attractive	0.25	thrust	-0.17
beautiful	0.24	nk	-0.17
breast	0.23	macro	-0.16

Table 2: Top 10 Words Most Predictive of Female/Male (*Pronoun sample*)

Most *female*		Most *male*	
Word	ME	Word	ME
pregnancy	0.29	knocking	-0.33
hotter	0.29	testosterone	-0.20
pregnant	0.26	blog	-0.18
hp	0.24	hateukbro	-0.18
vagina	0.23	adviser	-0.17
breast	0.22	hero	-0.17
plow	0.22	cuny	-0.17
shopping	0.21	handsome	-0.17
marry	0.21	mod	-0.17
gorgeous	0.20	homo	-0.16



To get the intermediate results, I inspected the lasso folder within the replication package. Each one of the three Python files in there is responsible for the training of one of the models presented in the paper. In each one of them, there were some common problems relate to different versions of numpy and pandas. In each one of them, I had to change every call from .as_mat() to .to_numpy(), abiding to my pandas version, and changed the arguments of np.load to include allowpickle=True, also to ensure it runs without errors according to my numpy version. I also added some print statements to understand at which step of the code my computer was, since they took a long time to run. After this, I ran each one of the files, which yielded the expected results. Lasso linear pronoun sample took a lot of time to run (i.e. more than 4 hours), so I was wondering if a better computer is needed to get the actual results.

Learned about sparse data structures in Python to deal with the great amount of 0s, which represent that a word is not in the post. This made me stick to scikit learn. I tried other libraries, but they would require more RAM to process the data.

Extension

For the extension, I first considered modelling a Random Forest to predict if a post is *Male* or *Female* given the occurence of the words. This seemed to be a good idea at first, as the non-linearity provided by the Random Forest algorithm could provide some new insight ffor classification. However, the first thing that made me drop this idea was the amount of time to train and tune such a model. Given the dimensionality of the problem and they way the algorithm itself works (tuning number of estimators) I figured that it would take a long time to train it. Besides that, there is no "easy" way of evaluating the Marginal Effect (ME) of a certain word in the post to classifying it as *Female*, as the trees themselves have no intuitive explanatory way of showing this. One way of doing this that I found on the internet is calculating the Average Marginal Effect, which is computationally intensive and could take too much time to evaluate the effect of all the words, because of the high dimensionality of the data. Then, I thought about the Ridge regression. I will replicate the model with a different penalty on the regressors. The **Ridge Logistic Model** changes the penalty on the θ from $|\theta|$ to θ^2. The new model can be represented as follows (Hastie et.al 2009):

$$\hat{\theta}_\lambda = \text{arg min}_\theta \left[-\log(\Pi_{i=1}^{N} P(\mathit{Female}|\boldsymbol{w}_i)) + \lambda f(\theta)_1 \right], \text{ where}$$
$$f(\theta)_1 = \Sigma_{j\geq 1}(\theta^j)^2$$

I believe the author thought about utilizing this model. I think that the model chosen for the paper was the one with L1 Regularization (LASSO) because some of the predictors are likely to become 0 under the absolute value computation (Hastie et.al 2009), contributing to a more compact relevant dimensionality of the vector $\boldsymbol{w}_i$. This is not true under L2 regularization. One advantage of L2 Regularization however, is that its implementation might be more efficient than L1 if multiple models are needed. This is because of the equivalence between the L2-regularized solution path for a convex loss function, and the solution of an ordinary differentiable equation (Zhu & Liu 2021).

This is the result of my extension, which is differs slightly from the table shown in the original paper. We can see that the word "gay" appears as predictive of *Male* posts, which confirms the discrimination towards minority male groups discussed in the paper. "Feminist" appears as a word predictive of *Female* posts, which is interesting. The coefficients are also smaller, since we didn't drop as many words like the LASSO regularization did

Table 3: Top 10 Words Most Predictive of Female/Male (Ridge Regularized Logistic Model):

Most *female*		Most *male*	
word	ME	word	ME
hotter	0.30	homo	-0.23
pregnant	0.28	macro	-0.15
hot	0.26	fieckers	-0.15
marry	0.25	blog	-0.14
attractive	0.23	testosterone	-0.14
beautiful	0.23	fenance	-0.13
plow	0.23	macroeconomics	-0.13
feminist	0.19	gay	-0.13
pregnancy	0.19	adviser	-0.13
marrying	0.18	satisfaction	-0.12

References

Hastie, T., Tibshirani, R., Friedman, J. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Second Edition. Princeton, NJ: Springer, 2009.

Wu, Alice H. 2018. "Gendered Language on the Economics Job Market Rumors Forum." *AEA Papers and Proceedings* 108:175-79.

Zhu Y., Liu R. 2021. "An algorithmic view of L2 regularization and some path-following algorithms." *Journal of Machine Learning Research* 22:1-62.