The Language of Discrimination: Using Experimental versus Observational Data

By Aislinn Bohren, Alex Imas, and Michael Rosenberg*

Social scientists have studied the incidence of discrimination across a variety of settings (Bertrand and Duflo 2016). Discrimination has been documented in labor markets (Bertrand and Mullainathan 2004), law enforcement practices (Knowles, Persico and Todd 2001) and housing applications (Ewens, Tomlin and Wang 2014), to name a few. The majority of studies focus on discrimination in observable, easy to quantify behavior, such as call back rates for job applications or initial offers for products and services. However, discrimination can also occur along dimensions that are harder to quantify, such as the language used when engaging with and evaluating members of a targeted group.

Social scientists have increasingly employed text and language data as inputs into their analysis because of its explanatory power for economically important outcomes (Gentzkow, Kelly and Taddy 2017). Language analysis has been used to capture media slant (Gentzkow and Shapiro 2010), measure policy uncertainty (Baker, Bloom and Davis 2016), and capture racial animus by geographic region (Stephens-Davidowitz 2014). The language content of central bank communication has been found to be a more important determinant of interest rates than policy rate decisions (Lucca and Trebbi 2009). Language can have profound downstream consequences for savings and health behavior (Chen 2013).

In this paper, we examine whether people respond differently to questions posed by women versus men. We use two data sources to answer this question. The first source is comprised of data from an experiment, where similar questions are randomly assigned to be posted on a large mathematics forum from accounts with male or female usernames. We analyze the language used in the responses to these question posts. Using techniques from machine learning, we document a significant difference in the distribution of language used in response to questions from male versus female usernames. Next, we employ sentiment analysis to explore the drivers of this difference. Sentiment analysis captures the valence of language, positive and negative, by measuring the usage of opinion words such as 'flawed' and 'awful' in the case of negative sentiment, or 'great' and 'stellar' in the case of positive sentiment.[1] We find a significant difference in the sentiment of answers to questions from male versus female questions. Answers to questions posted by female accounts score significantly higher on both negative *and* positive sentiment; responses to female users contain more opinion words, both positive and negative, than responses to male users.

The second source uses observational data from the forum – question posts and the associated responses. We use an algorithm to infer the gender of the username for each question post. As in the case of the experimental data, we find a significant difference in the distribution of language used in the responses to questions posted from female versus male usernames. The sentiment of these responses also differs significantly by the inferred gender of the question poster, but in the *opposite* direction to the experimental data: responses to questions posted from male usernames score higher for both positive and negative sentiment, compared to responses to questions posted

* Bohren: University of Pennsylvania, abohren@sas.upenn.edu. Imas: Carnegie Mellon University, aimas@andrew.cmu.edu. Rosenberg: Wayfair Inc., mrosenberg@wayfair.com.

[1] For example, sentiment analysis has been used to predict stock market outcomes from measures of mood in Twitter messages (Bollen, Mao and Zeng 2011).

from female usernames. One possibility to explain the difference is that in the observational data, male and female question posters use different language, whereas in the experimental data, this issue is controlled for due to random assignment. Indeed, we find that questions posted by male users score higher on both positive and negative sentiment than those posted by female users. Therefore, the sentiment of the responses may reflect this difference in sentiment of the questions.

Overall, our results document significant differences in the language used to respond to males versus females. Importantly, they also highlight the importance of experiments to establish whether there is a causal relationship between gender and language differences, as gender may be confounded with other factors in observational data.

I. Analysis

A. *Description of data*

We conduct our analysis using data from a large mathematics Q&A forum. Users on the forum post mathematics questions, answer other users' questions, and comment on both answers and questions. Users vote on other user's posts to evaluate their quality – high in the case of an upvote or low in the case of a downvote. An upvote serves a dual purpose: it highlights a quality post, and also rewards the poster for producing high quality. A poster earns reputation points for each upvote, and loses reputation points for each downvote. Reputation points give the user additional privileges on the forum, such as the ability to edit or flag other posts. They can also be used as currency on the forum – users can put 'bounties' on questions they would like answered; this bounty of reputation points is transferred from the question poster to the user who generates the best answer. For each question or answer, the number of reputation points and the username of the poster is publicly displayed in the bottom corner of the post.

To generate the experimental data, we wrote 140 original college-level mathematics questions. We randomly assigned these questions to post on 140 new accounts created for the experiment. Half of the accounts were given female usernames and the other half of the accounts were given male usernames (names were taken from the list of "Top names of the 2000s" created by the Social Security Administration). Questions were posted between 5 and 10 PM EST on Monday through Thursday, which were predetermined to be the most active times on the forum.[2] We tracked the comments and answers that were posted in response. We received 163 comments and 161 answers in total on these questions.

For the observational data, we used publicly available data on all posts on the forum between July 2010 and September 2017. To make the analysis comparable, we focused on "first questions" on the forum, i.e. each user's first post to the forum, if the post was a question. These questions were posted when the user had no reputation points. We used a gender inference tool to infer the gender of the usernames.[3] We restricted the analysis to questions posted by accounts identified as either male or female. There were 87,133 questions that met this criteria, and these questions received 205,077 comments and 125,933 answers.

B. *Methods*

We employ techniques found in the literature on statistical natural language processing.[4] The main challenge in the analysis of language is the high dimensionality of the data. The dimensionality of language representations quickly explodes: the unique representation of n words drawn from a set of possible words V has dimension $|V|^n$ (Gentzkow, Kelly and Taddy 2017). Therefore, we employ methods from machine learning that are used to analyze high dimensional data. We are interested in whether the language contained in responses to questions differs by the gender of

[2] Of the 140 questions, 5 were posted incorrectly and 14 were closed by moderators on the site.

[3] The gender inference tool and accompanying documentation can be found at https://github.com/tue-mdse/genderComputer.

[4] See Manning and Schütze (1999) for an overview.

the question poster. To address this question, we first need to quantify the respective language distributions.

We use a unigram language model to create a probabilistic representation of language (Jurafsky and Martin 2014). Let vocabulary V_G denote the set of possible words in a text corpus G (a collection of documents in a given language). A unigram model represents a language as a probability distribution over V_G. The fundamental assumption of a unigram model is that the probability of a word is independent of previous words, i.e. for $w \in V_G$,

$$p(w) \equiv C_G(w)/W_G, \tag{1}$$

where C_G is the count function over individual words in corpus G and W_G is the total number of words in corpus G.

In order to measure the difference between two unigram language models, we define a measure of distance between the probability distributions that represent these languages. Let p and q be probability distributions that represent unigram language models. The Kullback-Liebler (KL) Divergence is defined as

$$D_{KL}(p \parallel q) \equiv \sum_{w \in V} p(w) \log \left(\frac{p(w)}{q(w)} \right). \tag{2}$$

This measures the divergence of p from q. It is often used in machine learning as a measure of "surprise" of data being generated by p, conditional on the hypothesis that the data is generated by q. If there are no differences between the two distributions, $D_{KL}(p \parallel q) = 0$. If $D_{KL}(p \parallel q)$ is large, the likelihood of seeing data generated under p, conditional on q being the true distribution, approaches 0.

If language data is relatively sparse, one issue that arises is calculating D_{KL} when the two language models have different supports (i.e. there exists a word w such that $p(w) = 0$ and $q(w) > 0$, or vice versa). We use two methods to address this issue. The first is Lidstone smoothing (Manning and Schütze 1999), which assigns small probability to unseen and rarely seen words by shaving off small probability from commonly seen words.[5] The second is to restrict attention to the set of words that have positive measure in both distributions, i.e. $V_p \cap V_q$, where V_p and V_q are the vocabularies of the corpora for measures p and q, respectively. We refer to the first method as "Smoothing" and the second method as "Shared Vocabulary".

We use nonparametric bootstrapping (Wasserman 2013) to test for a difference between the distributions p_M and p_F representing the language of the response posts to male versus female question posters, respectively. Let R be the set of response posts to questions, with n_F posts in response to females and n_M posts in response to males. The nonparametric bootstrap calculates the null distribution for a test statistic $T(p_F, p_M)$ in the following way. For each simulation $s = 1, ..., 1000$, without replacement, sample n_M posts from R to create male corpus G_s^M and sample n_F posts from R to create female corpus G_s^F. Using Equation (1), estimate sample distribution $\hat{p}_s^M$ from G_s^M and $\hat{p}_s^F$ from G_s^F. In the case of the KL Divergence test statistic, estimate the sample distributions using either smoothing or shared vocabulary, as defined above. Next, use the sample distributions to calculate $\hat{T}_s = T(\hat{p}_s^F, \hat{p}_s^M)$. The set $\{\hat{T}_1, ...\hat{T}_{1000}\}$ creates the null distribution. If the male and female corpuses in each simulation are sampled from the same distribution of posts R, then the null hypothesis that the true distributions p_M and p_F do not differ holds when estimating each sample distribution $\hat{p}_M$ and $\hat{p}_F$. We use the distribution of the test statistic under the null to calculate p-values for whether the true language models p_F and p_M differ.

We also test whether the sentiment of responses differ depending on the gender of the question poster. We focus on positive and negative sentiment, which measure the incidence of positive and negative opinion words, respectively. To calculate the sen-

[5] Consider a unigram language model on G with vocabulary V_G. Let V' represent a larger vocabulary, $V' \supseteq V_G$. Given $\lambda \in [0, 1]$, define the unigram language model for $w \in V'$ as $p(w) \equiv \frac{C_G(w)+\lambda}{W_G+\lambda|V'|}$. Following Manning and Schütze (1999), we use $\lambda = 0.5$ in our analysis.

timent of a post $(w_1, ..., w_n)$, we use the NLTK package to tag the part-of-speech $t \in T$ of each word in the post, where T denotes the set of parts-of-speech (Bird, Klein and Loper 2009). This yields tagged post $((w_1, t_1), ..., (w_n, t_n))$. Given a sentiment (i.e. positive or negative), we measure the sentiment of each word in a post using the SentiWordNet corpus (Baccianella, Esuli and Sebastiani 2010) to define a sentiment function $\sigma : V \times T \to [0, 1]$, which maps each word and part-of-speech pair to a sentiment score. The sentiment score is increasing in how indicative the word is of the given sentiment. We assign word and part-of-speech pairs that are not in the SentiWordNet corpus a sentiment score of zero. The overall sentiment score for a post is the average score for each word in the post.

C. Results

We first test for language differences in the responses to male and female question posts. We calculate the test statistic defined in (2) for the KL-Divergence of the estimated language distribution of responses to female question posts, with respect to the estimated language distribution of responses to male question posts. These results are presented in Table 1. In the observational data, we observe a significant difference in the distributions of language in response to female versus male question posts for both comments and answers, using both the smoothed and the shared vocabulary distributions. In the experimental data, we observe a significant difference in the smoothed language distributions for both answers and comments. The difference in the shared vocabulary distributions is not significant for either answers or comments, though the p-values approach conventional levels of significance (.124 and .132, respectively). Given the high-dimensionality of data used in the distributional analysis, the experiment is likely underpowered to detect differences in shared vocabulary.

Next, we test for sentiment differences in the responses to male and female question posts. We calculate the difference (female minus male) in the average positive and negative sentiment score. A positive difference indicates that responses to females posts display more of the sentiment, while a negative difference indicates that responses to male posts display more of the sentiment. These results are presented in Table 2.[6] In the experimental data, we find that answers to female posts are significantly more sentimental, both on the positive and negative dimension, than answers to male posts. We find no significant differences in the sentiment of comments. In the observational data, we also find differences in the sentiment of answers to female versus male question posts. However, the differences are smaller and in the opposite direction, compared to the experiment: answers in response to male posts are more sentimental, both on the positive and negative dimension, than those in response to female posts. We also find a significant difference in the sentiment of comments, in the same direction as the effect for answers.

One important factor that differs between the experimental and observational data is that questions in the former were randomly assigned to gendered accounts, whereas questions in the latter were not. If male and female users write questions that differ in language and sentiment, then one cannot establish the causal role of gender in generating different responses in the observational data. Specifically, the difference in language and sentiment of responses could be caused by the differences in the language and sentiment of questions from female and male users. Indeed, in the observational data, we observe a significant difference in the language distribution of questions posted by females relative to males – suggesting that male and female users do use different language to ask questions (Table 1). We observe no such difference in the experimental data, as expected – the p-values are .916 and .806 for smoothed and shared vocabulary, respectively. Further,

[6] The dimensionality of sentiment data is substantially lower than the dimensionality of language distribution data, so a given dataset has more power to detect differences in sentiment, relative to differences in the language distributions.

Table 1—Language differences by gender.

		Smoothed		Shared Vocabulary		
		$D_{KL}(\hat{p}_F \| \hat{p}_M)$	p-value	$D_{KL}(\hat{p}_F \| \hat{p}_M)$	p-value	# Obs.
Experiment	Answers	.645	(.002)	.271	(.124)	161
	Comments	.561	(.007)	.227	(.132)	163
	Questions	.459	(.916)	.195	(.806)	121
Obs. Data	Answers	.097	(.000)	.042	(.000)	125,933
	Comments	.068	(.000)	.034	(.000)	205,077
	Questions	.090	(.000)	.042	(.000)	87,133

Note: p-values are bootstrapped on null distribution with 1000 simulations each.

Table 2—Sentiment differences by gender.

		Female Sentiment	Male Sentiment	Difference (F-M)	p-value	# Obs. Female	# Obs. Male
Positive Sentiment							
Experiment	Answers	.0423	.0334	.0089	.029	77	84
	Comments	.0618	.0523	.0095	.141	82	81
Obs. Data	Answers	.0419	.0438	-.0019	.000	26,444	99,489
	Comments	.0602	.0610	-.0008	.002	42,256	162,821
Negative Sentiment							
Experiment	Answers	.0376	.0279	.0097	.036	77	84
	Comments	.0483	.0533	-.0050	.465	82	81
Obs. Data	Answers	.0335	.0347	-.0012	.000	26,444	99,489
	Comments	.0479	.0490	-.0011	.000	42,256	162,821

Note: p-values are bootstrapped on null distribution with 1000 simulations each.

the differences in the sentiment of questions follow the same pattern as the differences in the sentiment of responses to those questions: in the observational data, male question posts contain more sentimental language, both positive and negative, than female question posts.[7] This highlights the importance of using experimental methods to isolate the causal impact of an attribute (i.e. gender) on the language of responses.

In Table 3, we regress response sentiment on question sentiment to test whether the sentiment of a question influences the sentiment of the response in the observational data. We find that the positive sentiment of responses (both answers and comments) are increasing in the positive sentiment of questions, and to a much lesser extent, the negative sentiment of questions. A similar result holds for the negative sentiment of responses: for both answers and comments), it is increasing in the negative sentiment of questions, and to a much lesser extent, the positive sentiment of questions. Therefore, questions with more sentimental language are more likely to receive responses with more sentimental language, and particularly, language expressing similar sentiment to the question.

[7]The sentiment difference between males and females is -.001 and significant at the 0.000 level for both positive and negative sentiment. In the experiment, there is no significant difference in questions, as is expected given the randomized assignment of gender.

II. Conclusion

We use experimental and observational data to examine differences in the language used in response to questions posted by users with male versus female usernames. We document significant differences in the language of responses, both in terms of the distribution of language utilized and the sentiment of this language. In the observational data, we also document gender differences in the language and sentiment of

Table 3—Regression of sentiment of response on sentiment of question (p-values in parentheses)

	Comments		Answers	
	Positivity	Negativity	Positivity	Negativity
Question positivity	.1112 (.000)	.0005 (.912)	.2140 (.000)	.0981 (.000)
Question negativity	.0164 (.000)	.1339 (.000)	.0933 (.000)	.1786 (.000)
Constant	.0544 (.000)	.0430 (.000)	.0282 (.000)	.0216 (.000)
R^2	.004	.005	.060	.052
# Obs.	205,077	205,077	125,993	125,993

questions – in other words, males and females pose questions in different ways. This highlights the importance of using experimental data to identify the causal role that an individual's gender plays in how others respond to him or her.

REFERENCES

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." Vol. 10, 2200–2204.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty*." *The Quarterly Journal of Economics*, 131(4): 1593–1636.

Bertrand, M., and E. Duflo. 2016. "Field Experiments on Discrimination." In . *Handbook of Economic Field Experiments*, , ed. A.V. Banerjee and E. Duflo, –. North-Holland.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991–1013.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter mood predicts the stock market." *Journal of Computational Science*, 2(1): 1 – 8.

Chen, M. Keith. 2013. "The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets." *American Economic Review*, 103(2): 690–731.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. 2014. "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *The Review of Economics and Statistics*, 96(March): 119–134.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica*, 78(1): 35–71.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. "Text as Data." National Bureau of Economic Research Working Paper 23276.

Jurafsky, Dan, and James H Martin. 2014. *Speech and language processing.* Vol. 3, Pearson London.

Knowles, J., N. Persico, and P. Todd. 2001. "Racial Bias in Motor Vehicle Searches: Theory and Evidence." *Journal of Political Economy*, 109(1): 203–229.

Lucca, David O., and Francesco Trebbi. 2009. "Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements." National Bureau of Economic Research Working Paper 15367.

Manning, Christopher D, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing.* MIT press.

Stephens-Davidowitz, Seth. 2014. "The cost of racial animus on a black candidate: Evidence using Google search data." *Journal of Public Economics*, 118(C): 26–40.

Wasserman, Larry. 2013. *All of statistics: a concise course in statistical inference.* Springer Science & Business Media.