1. In this project, I’d like to explore the relationship between misogyny displayed on social media and gender-based discrimination in the labor market. So why might these two be related? Well, an economist named Gary Becker established that if an employer doesn’t like women, …
2. he would hire more men than in the optimal decision. Since there’s more demand for men, their wages will be higher than equally skilled women. Based on this model, if there’s a prevailing misogynistic culture, the gender wage gap would be high.
3. So in my project, I want to construct a measurement of misogynistic culture with Twitter. I will first introduce how I construct this measurement and show some observational results. Then I’ll proceed with the data of the wages, and other variables related with labor outcomes and present my findings.
4. The measurement is quite intuitive, for ideally each county in the US, I’ll calculate how often misogynistic tweets occurred, divided by how often people tweets in that county. And for the denominator, I figured out that it’s unnecessary to scrape all the tweets in a county…
5. to calculate people’s frequency of tweeting, instead, I can just count the number of tweets returned by a very common word, like “the”. And to identify the misogynistic tweets, I want to construct a lexicon. I found a lexicon from Github that contains thirteen hundred terms.
6. But I only want terms that are most related to misogyny. So I want to go through a primary data collection step, and find the terms that returns me at least one misogynistic tweet in a limited sample. So, the first step is to get an annotated dataset…
7. and use it to train a classifier that can recognize misogynistic tweets.I found the dataset I want from GitHub, which is a sample of around sixty-five hundred Reddit posts, with 700 of them being misogynistic, quite small, to be honest.
8. I feed it to the pretrained bert-base-uncased model and got an accuracy score of point nine for the test set. I think it’s impressive, though it’s related with the small size of the data and is under the risk of overfitting.
9. And then, I collected tweets by searching for each of the terms in that thirteen hundred words’ lexicon at five randomly selected dates, and got around 4 million tweets. I used the BERT model I just trained before to classify each of them. And it turns out only 16 terms in that set returns at least one misogynistic tweet.
10. And the word “bitch” is the most frequently occurred misogynistic term, surprise surprise. So have constructed the lexicon, I went through a second round of tweets scrapping. And in the end I got one and a half million tweets. And I wrote a function to identify the users’ location…
11. from the location string the users wrote in their profiles. And finally, I got only around sixty thousand tweets sent by users who chose to reveal their location publicly at county level, and wrote their location in a way not very innovative.
12. And around thirty hundred thousand users who have revealed the states they are located at. Then I scrapped tweets by searching for the word “the” on March 10, just a random date, and I calculated the Twitter Misogyny Index by the equation I mentioned before. And next we’ll look at some primary results.
13. The map on the left hand side shows the state-level results. And it turns out some states in the south are “bluest”, which stands for a higher level of misogyny. And on the right side is a graph I found from US news, which shows the female median earnings…
14. as a percentage of males. The “bluest”, the larger gap. So it turns out the misogyny index I calculated didn’t make much sense. So, if the theory was right, then there’s something wrong with my data and empirical strategies.
15. One possible answer is that the within-state variation is large, which seems to be true if we look at the calculated results at the county-level. But it seems that the data size is too small, and the colored areas are very sparse.
16. Another potential problem here is that different terms in my lexicon are unequally efficient in returning misogynistic tweets. So I want to calculate a coefficient matrix between the frequency of each of the terms and the labor outcomes.
17. So I collected some data from American Community Survey, and got some variables related with earnings, labor market participation, and education attainment. And for each of them, I calculated the ratio between the female and male population.
18. So I calculated the correlation matrix between all of these variables, and drew the heatmap. Here are the takeaways. First, the occurrences of the terms themselves are highly correlated with each other. This implies that perhaps it’s enough for me to just use two of them, or simply calculate a PCA and use the first component.
19. And second, the occurrences of misogynistic terms seem to be correlated with measurements of labor market participation. But the correlation is small or even negative for other variables.
20. So, to summarize,