MS&E 231

Gender Homophily on Twitter

Team members and their contributions:

Nirali Parekh	Streaming, parsing tweets, plots, and report
Roberto Lobato López	Gender inference, plots, and report
Dinesh Moorjani	Ideas exchange (Auditing the course)

Introduction

In this assignment, we aim to offer a perspective on gender homophily by utilizing real-time data from Twitter. According to Statista [1], about 60% of Twitter users worldwide are between 25 and 49 years old i.e. working age. Hence, to analyze the differences in perception of work and the workplace between men and women [2], we chose the keyword "work." The domain we are trying to understand through this keyword is how gender interactions differ in a regular setting versus a professional setting (i.e. sample vs control). This analysis can serve as a hypothesis for gender studies research topics like gender bias, and workplace support for women.

Methodology

A. Steaming Tweets from API and Parsing Zipped Files:

We utilized the helper script provided to stream real-time tweets for our assignment. We made minor changes in the script to extract some other useful fields, e.g. user description and tweet type. To parse the zipped files obtained in the previous step, we utilized the utility “zcat” to make sure we obtained the data in a streaming fashion. This made the process less memory-intensive and faster. The output of this step was the two .csv files with the following fields: Original Poster Name, Original Poster Description, Tweet Author Name, Tweet Author Description, Date, Time, and Tweet Type.

B. Inferring Gender

To infer the gender of the users, we use a 3-step hierarchical approach, each with its own strengths and weaknesses. Before any step is applied, we tokenized the users’ names and descriptions and converted the tokens into a set so we could match them against other sets of gendered words faster. If a match was found in a step, the gender was assigned. If not, the process continues to the next step.

For the first step, we noticed that many English-speaking users offer their self-identified gender through their Twitter bios by explicitly stating their pronouns: she/her for females, he/his for males, and they/them for non-binaries. Although we acknowledge that there are more genders and pronouns that could be inferred through this method, we limited ourselves to the three most common. The main strength of this method is that self-identification supersedes any other gender-inferring method by being 100% certain, if the user is truthful. There might be cases where users troll by stating a fake gender identity, but we consider that this should be an insignificant amount of cases. Nevertheless, a constant problem in any Twitter-related analysis is the high amount of non-human accounts and trolls. We did not take any action to reduce this problem. The main weakness of the first step is that only a small number of users (less than 5% in some tests) self-identify their pronouns and these users might have liberal biases which could affect the homophily analysis..

The second step consists of using the SSA’s baby name popularity data set. Although the data set includes all the names from 1880 to 2013, we restricted our sample to the names after 1920 to eliminate names that have fallen out of fashion. We also restricted our search to names that were used at least 500 times across all the years. This step reduces the number of female names from 63287 to 9547 and the male names from 38022 to 6235, making the process faster and with higher accuracy. Then we looked at names that were used by both females and males and assigned the gender that had the higher frequency. A possible weakness of this method is that a name’s main gender could change through the years. The strength of this step is that it accounts for the majority of the data (at least 90% in some tests). But the weakness is that some people have last names that are used as names for the other gender. Since our method first checks female and then male, it might be biased to female names. This method also assumes that only 2 genders exist, which is not true. Another weakness is that a lot of people do not use their real names on Twitter. At last, this method is US-centric, and using global datasets would make a more complete analysis.

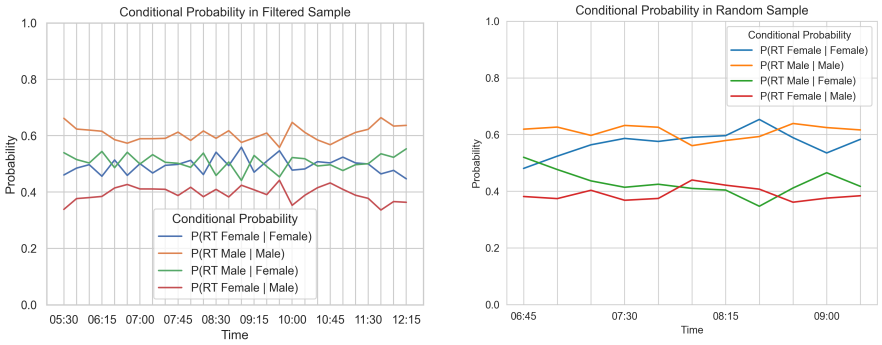
The third step is inferring the gender through keywords on their bios. This step is the noisiest and uses stronger assumptions. The idea is that some people will express their gender by using gendered words while actively describing their identity. For

example, someone's bio can say "Resident Evil Girl" or "Father of three". We can assume female in the former and male in the latter (see the Appendix for our complete set of gendered keywords). The assumption is that people won't passively identify themselves or use the words in another context, for example, "I am patriot from the Mother Land" or "Looking for the woman of my life."

Non-binary users represented less than 1% of the classified data, so we dropped them. After applying the steps to the "work"-filtered sample, we were left with 63,389 observations that had identified the gender of the original poster and the tweet author. In contrast, in the random sample, we were left with 26,585 observations.

C. Measuring Gender Homophily

For the purpose of our analysis, we treat retweets, quotes, and replies as the same interaction. We don't think it is a strong assumption, but it might bias it by reducing homophily, such as in cases where males only retweet males but interact with females to troll or harass them. Our filtering keyword (**work**) is non-controversial, but it might still reflect behavior disparities among genders. In our hypothesis, we expected to see men interacting more with men and women interacting more with women due to both genders having different expectations in the workplace and work-related tasks. To test it, we plotted the frequency of the interactions across genders (See Appendix). The filtered sample had 55% males vs 50% in the random sample. To resolve the unbalance, we calculated the conditional probabilities for each period.



In the filtered sample, the plot shows men retweet men with a higher probability than women: 60% vs 40%. While no difference is shown between women. In the random sample, both genders are more probable to retweet their own gender. A possible explanation is that men pay less attention to what women talk about work, while the reverse is not true. Another possible explanation is that women talk about domestic work and office work at the same frequency while men talk more about just office work, so they would interact with women talking about domestic work.

Limitations

We couldn't make Twitter's API stream for more than 6 continuous hours, and we reached our 20% limit in those trials. The stream is limited by 6 hours of data during bedtime for Americans, so the sample is biased to segments where our second step of the gender inference is less powerful. The day of the inference was the weekend, so the filtered word work could be more interesting during workdays. The sample data used as a control was taken a day later and has fewer observations, so it might not be as good as a control. The results are highly dependent on the gender inference technique, which is highly dependent on the set of words. Twitter's users are overrepresented by males (60%) in the US [3], while our data is closer to 50%. It could reflect selection bias from Twitter users (more males) and from our sample (not representative of Twitter).

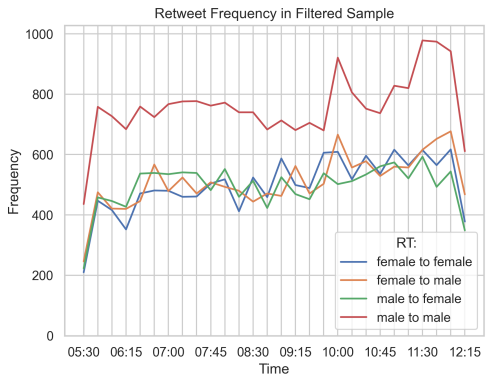
Conclusion and Future Scope

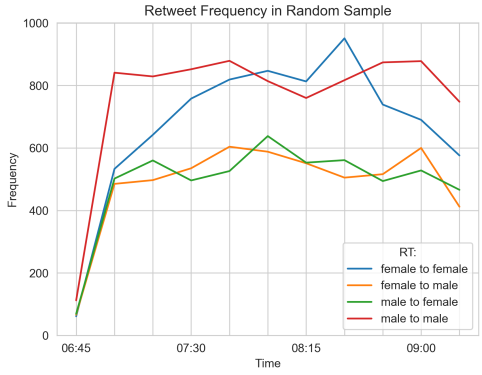
Homophily is not always symmetrical. Context can produce lower and higher homophily among different segments. In our case, homophily observed in the random sample is reduced in women when tweets are work-related.

We expect the results to be highly sensitive to changes in the samples and methodology, so more robustness checks are needed. Testing our gender inference method and analyzing a bigger sample would give us more certainty. We could run a separate analysis for each of the steps in the gender inference method. To validate our homophily explanations, we could dive deeper into the tweets' text to understand the context of the interactions. Since the amount of data makes it prohibitive to do an exhaustive manual search, we could use sentiment analysis in the tweets' text.

Appendix

Frequency Plots





Conditional probabilities across all the samples

Work-filtered sample:
p(male) = 0.554
p(female) = 0.446
p(RT female | female) = 0.494
p(RT male | male) = 0.606
p(RT male | female) = 0.506
p(RT female | male) = 0.394

Random sample
p(male) = 0.519
p(female) = 0.481
p(RT female | female) = 0.581
p(RT male | male) = 0.609
p(RT male | female) = 0.419
p(RT female | male) = 0.391

Gendered keywords

- Female keywords: "female", "mother", "daughter", "sister", "woman", "girl", "mujer", "madre", "hija".
- Male keywords: "male", "father", "son", "brother", "man", "boy", "hombre", "padre", "hijo".
- Non binary keywords: "nb", "nonbinary", "binarie".

Files

- tweets_random.gz → zipped file of a random sample of collected tweets
- output_random.csv → output of parse_tweets.py on tweets_random.gz
- tweets_keyword.gz → zipped file of tweets collected with keyword “work”
- output_keyword.csv → output of parse_tweets.py on tweets_keyword.gz
- tweet_stream.py → script to get real-time tweets from Twitter API
- parse_tweets.py → script to convert .gz files into readable .csv files
- tweet_analysis.ipynb → notebook with the gender inference and analysis code
- tweet_analysis.py → script to reproduce the graphs and analysis

To reproduce the results

```
>> python3 tweet_stream.py --keyfile creds.txt --gzip tweets_keyword.gz --filter work
>> zcat tweets_keyword.gz | python3 parse_tweets.py > output_keyword.csv
>> python3 tweets_analysis.py
```

References

[1] https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

[2]https://www.pewresearch.org/social-trends/2017/12/05/americans-see-different-expectations-for-men-and-women/

[3]https://www.oberlo.com/blog/twitter-statistics#:~:text=3.-,Twitter%20Demographics%3A%20Gender,females%20at%202.5%20to%20one.