Data for "Gender Bias in Rumors among Professionals: An Identity-based Interpretation"

Alice Wu

This file lists all data sets provided for replication in the "data" folder. Please also refer to "README_code.pdf" for programs that generate and clean the datasets.

- **Final data sets:**
  - **"full_sample2019_stata_final.dta"**: all threads in the gender sample that include at least one Female/Male post. Variables include measures of gender and topics at the posts level. Please see "codebook_full_sample2019_dta.log" for variable descriptions. All results in this paper are run on this dataset, through "static_analysis_2019.do" and "dynamic_analysis_2019.do"

  - **"full_sample2019_stata.csv"**: csv format for the dta file above

  - **"full_sample2019.csv"**: all threads in gender sample. The column "raw_cat" contains the content of original posts. Please see "codebook_datasets.xlsx" for variable descriptions. The step from this dataset to the final dataset "full_sample2019_stat.csv" is completed by "gen_full_sample.R".

Now I start from the raw data scraped from the Economics Job Market Rumors forum and explain how I generate the final datasets above.

- **Raw data sets:**

  (a) *Directory: "..data/raw"*
  - **"scraped_posts.txt"**: raw text for threads scraped from the EJMR forum as of October 2017

  - **"EJR0_raw_text.csv"**: scraped data above converted into a .csv file

  - **"EJR0_raw_text_cleaned.csv"**: raw data with unique IDs for threads, and posts under each thread, cleaned from "EJR0_raw_text.csv" through the program "clean_raw_text.R"

  - **"main_stats_cleaned.csv"**: thread-level statistics as shown on the main listings on the EJMR websites, including information such as thread title, total number of posts and views, etc.

  - **"raw_time_stamp.csv"**: time stamp (e.g., "Posted on: 6 years ago") for the first post in each thread

  - **"vocab_Nov2017.pkl"**: a dictionary for the most frequent 10,000 words, generated through the program "word_count.py"

- **"counts_Nov2017.npz"**: 2.2 million – by – 10,000 matrix of word counts saved in compressed format, also generated through "word_count.py".

- **"all_cat_2019.csv"**: counts of words under each category in each post. The category assignment for each of the most frequent 10,000 words can be found at "../vocab/cleaned_vocab.csv".

- **"EJR_ALL_categories_2019.csv"**: same as above but include IDs for threads and posts, see "prep_for_analysis.py".

- **"all_classifiers_2019.csv"**: counts of female and male classifiers in each post.

- **"EJR_ALL_gender_classifiers_2019.csv"**: same as above but include IDs for threads and posts, see "prep_for_analysis.py".

- **"full_sample_job_rank.csv"**: counts of keywords indicating each job rank in each post in the full gender sample (threads that include at least one Female or Male post), generated by "job-rank.R".

(b) *Directory: "..data/vocab"*

- **"cleaned_vocab.csv"**: the most frequent 10,000 words as in "vocab_Nov2017.pkl", indexed from 1 to 10,000, and category classification

- **"gender_classifiers.csv"**: 53 female and 204 male classifiers, e.g., gender pronouns, woman/man.

- **"category"** folder: each .txt file here lists the indices under each category

(c) *Directory: "../data/gender_sample_by_classifiers"*

- "**EJR_gender_dataset_Jan2018.csv"**: threads that include at least one Female or Male post identified through gender classifiers alone (not names of economists as of January 2018). This sample includes 1.7 million posts in total. The column "female0_pred" indicates if it is a Female versus Male post.

(d) *Directory: "..data/NBER"*

- **"author-history-ID-by-full.csv"**: thread-post IDs for posts that mention an NBER author's full name, generated by "NBER-post-history-ID (by full).R"

- **"author-history-ID-by-part.csv"**: IDs for posts that mention first name/last name of an author, within threads that include at least one post mentioning his or her full name; generated by "NBER-post-history-ID (by part).R"

- **“author-history-merged.csv”**: merge the datasets above with 2.2 million posts in “EJR0_raw_text_cleaned.csv”, and preserve all posts under threads with at least one occurrence of an author’s full name.

- **“nber-author-info.csv”**: information about authors of NBER working papers, scraped from the NBER websites.
- **“complete-nber-author-info.csv”**: information about NBER authors as above with manually assigned job rank (when one is not affiliated with NBER).

(e) *Directory: “..data/JMC”*

- **“jmc_data_gender_nonmissing.csv”**: a directory for 4,750 economics Ph.D. who graduated between 2011 and 2018. The dataset includes their names, gender, school and year of graduation.

- **“JMC-history-ID.csv”**: thread-post IDs for posts that include a person’s full name, or first name/last name/initials under threads where full name has occurred at least once. The dataset is generated by “JMC-post-history-ID.R”.

- **“JMC-history-ID-merged.csv”**: merge the two datasets above.

(c) (d) (e) provide three sources of posts that discuss female or male subjects. I merge them together and generate the full gender sample in “full_sample2019.csv” that preserve all posts under threads that include at least one Female or Male post.