Text Analytics, HW4
Louis-Charles Généreux

Github link: https://github.com/MSIA/lgo4950_msia_text_analytics_2021/tree/homework4

Note to grader:

I used 2 different approaches to detect gender bias in Yelp reviews:

1. Generate word2vec embeddings using my text corpus, then identify bias by analyzing in nearest neighbor suggestions to gendered-words
2. Train 2 predictive models taking in text reviews to predict star ratings (one for reviews mentioning males, one for reviews not mentioning males); then interpret parameters
    a. Identify text entries that make gender references, split them in male or female gendered reviews
    b. Train logistic classification models (one for male reviews, another for female reviews)
    c. Extract regression parameters across both models
    d. Identify parameters that are unique to men/women in predicting high or low scores
    e. Identify whether parameters that are unique to men/women are indicative of bias

1. **Using the word2vec embeddings that you have created (earlier assignments), attempt to detect gender bias.**

Approach 1: using word2vec embeddings

Based on prior experiment results experimenting with this dataset, I have decided to use a skip-gram model to generate embeddings, with the following parameters:

- Embedding size: 100
- Window size used in training: 5
- Minimum observations of a word in corpus to be considered: 5

After preprocessing of data and training, I have generated various combinations of nearest neighbor suggestions, which allow me to identify words which are most interchangeable in the context of this text corpus. Overall, the model seems well tuned given the general quality of nearest neighbors. As can be seen below, it captures complex concepts like geographical locations, and can catch on to informal expressions (e.g., happy camper).

country	happy	disaster
[hawaii, 0.668, philadelphia, 0.651, city, 0.645, socal, 0.638, world, 0.637, southeast, 0.631, africa, 0.621, northeast, 0.621, houston, 0.62, austin, 0.616]	[pleased, 0.798, thrilled, 0.791, delighted, 0.737, satisfied, 0.73, camper, 0.682, impressed, 0.645, ecstatic, 0.631, stoked, 0.624, disappointed, 0.622, oblige, 0.618]	[nightmare, 0.739, travesty, 0.63, disastrous, 0.619, disappointment, 0.616, fiasco, 0.602, letdown, 0.59, botched, 0.59, horrible, 0.587, dud, 0.585, terrible, 0.581]

However, below are some blatant examples of gender bias in nearest neighbor suggestions:

- Some employment positions are so implicitly tied to the female gender that they appear within the top 10 nearest neighbors for the word **'lady'**. This is the example with 'cashier' below

    - woman, 0.939
    - girl, 0.903
    - gal, 0.822
    - gentleman, 0.788
    - guy, 0.725
    - gent, 0.711
    - fella, 0.691
    - man, 0.676
    - cashier, 0.673
    - person, 0.665

- On the flip side, other higher-paying positions (which could also be considered more prestigious) are associated with the male gender. For example, see the top top nearest neighbors for the word **'sales':**

- finance, 0.803
- salesperson, 0.735
- salesman, 0.711
- patrick, 0.7
- rep, 0.667
- financing, 0.664
- leasing, 0.648
- cody, 0.642
- vito, 0.639
- consultants, 0.636

- There is evidence from these embeddings that higher prestige jobs are tied with references to the male gender while less prestigious jobs are associate to the female gender
    - **Owner** has 'Trevor' (a masculine name) as a top 10 close neighbor
    - **Employee** (which one would assume to be less prestigious than owner) has the gendered 'saleswoman' word as a top 10 nearest neighbor

- Also, some jobs are implied to belong to one gender versus the other
    - **Cashier** has 'waitress' and 'hostess' (feminine words) amongst its top 10 nearest neighbors
    - **Driver** has 'bellman' (masculine word) among its top nearest neighbors

- Even more concerning, comparing the male and female equivalents of the same word, reveals that there are some negative behaviors only associated to the female gender
    - **Saleswoman** has the negative nearest neighbors: 'coldly'
    - However, for **salesman**, there are no such negative words

Exhaustive results

	lady	sales	owner	employee	cashier	driver	saleswoman	salesman
model_sg_100_5	[woman, 0.939, girl, 0.898, gal, 0.821, gentleman, 0.798, guy, 0.744, gent, 0.721, man, 0.686, person, 0.677, cashier, 0.666, cindy, 0.662]	[finance, 0.808, salesman, 0.705, salesperson, 0.704, rep, 0.677, isaac, 0.664, cody, 0.652, financing, 0.652, consultants, 0.649, ian, 0.648, leasing, 0.644]	[manager, 0.824, gm, 0.792, managerowner, 0.78, proprietor, 0.774, ownermanager, 0.757, owners, 0.732, manger, 0.729, chefowner, 0.722, trevor, 0.709, ownerchef, 0.679]	[worker, 0.801, staffer, 0.796, associate, 0.793, clerk, 0.764, pharmacist, 0.752, attendant, 0.727, managerowner, 0.725, person, 0.721, cashier, 0.712, saleswoman, 0.707]	[register, 0.869, counter, 0.803, clerk, 0.764, barista, 0.721, taker, 0.716, employee, 0.712, trainee, 0.712, waitress, 0.707, sneered, 0.706, hostess, 0.705]	[dispatcher, 0.782, drivers, 0.728, operator, 0.718, estimator, 0.713, shuttle, 0.709, bellman, 0.694, mover, 0.682, zifty, 0.681, lyft, 0.68, dispatch, 0.68]	[salesgirl, 0.762, coldly, 0.761, staffer, 0.76, saleslady, 0.747, marissa, 0.743, timeshe, 0.737, serverowner, 0.734, clerk, 0.734, hostesss, 0.733, noune, 0.733]	[salesperson, 0.804, rep, 0.719, dealership, 0.71, dealer, 0.709, mechanic, 0.705, sales, 0.705, robert, 0.699, salesmen, 0.692, salesguy, 0.683, keith, 0.678]

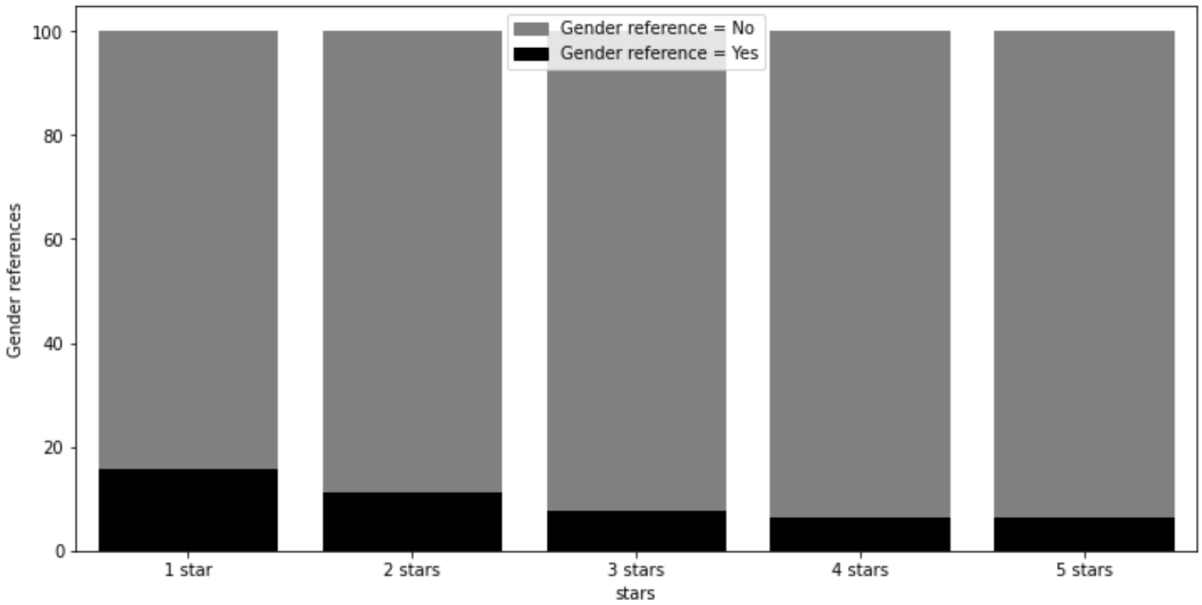
Approach 2: splitting corpus on male and female references, then training separate models

For this approach, I took over 8 million reviews and kept only which made a direct reference to gender (containing any of the following words: ['man', 'male', 'boy', 'guy', 'gentleman'], ['woman', 'female', 'girl', 'gal', 'lady']). After having selected only 700K reviews which contained the aforementioned words, I split the reviews into (a) those that made a masculine reference and (b) those that made a feminine reference.

I then trained classification models (taking text and classifying into “high” or “low” number of star reviews) specifically to the male and to the female datasets. My objective was to use explainability methods to **discover whether certain gender-specific words carried a lot of importance in classification**. My initial hypothesis was that if there were major discrepancies in important parameters between the masculine and feminine models, this could highlight some form of bias in language (different words associated with reviews speaking about women versus those in reviews speaking about men).

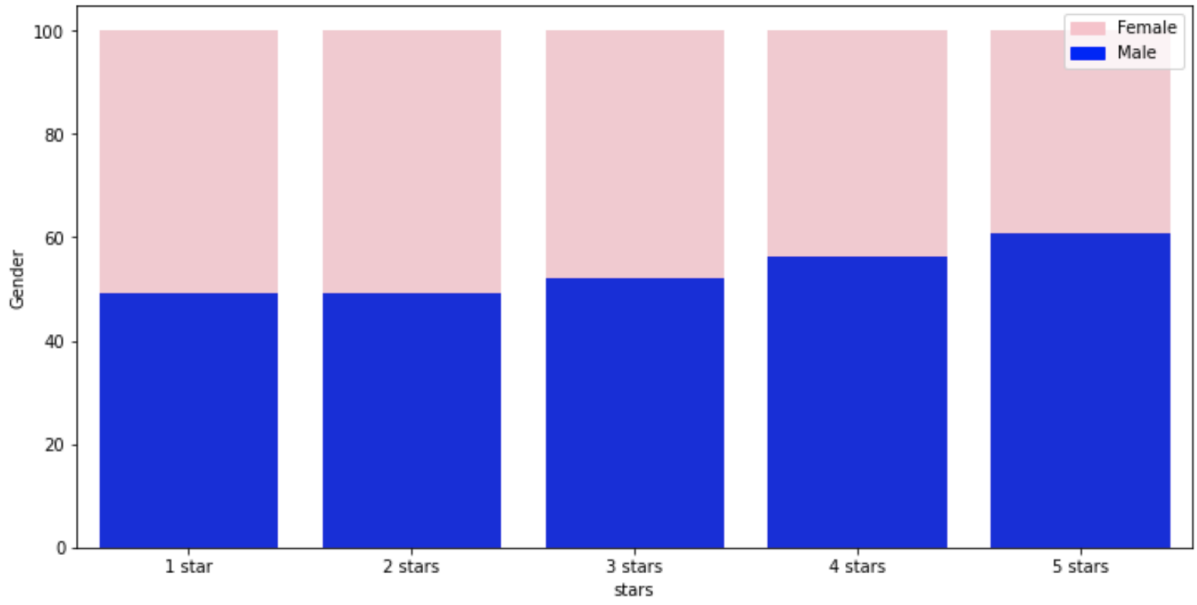
Below are some of my findings:

1. **References to gender are more common in bad reviews**



16% of 1 star reviews have a reference to gender. For 5 star reviews, this proportion is only 6%.

2. **For low reviews, there is a roughly 50-50 split of male to female reference, however male references are 1.5X more likely than female in 5-star reviews**



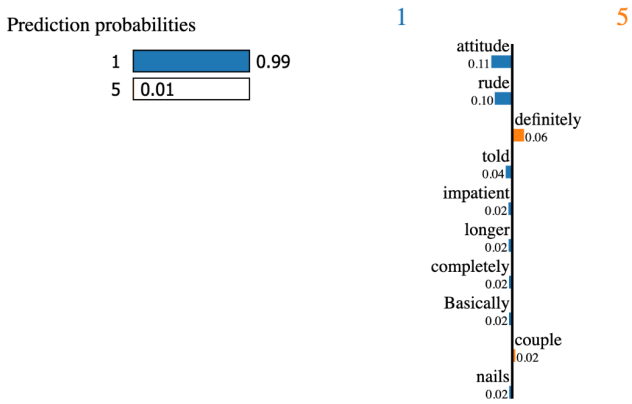
48% of gender references in 1 star reviews are 'male' references. For 5 star reviews, 60% of gender references are male

3. **Finally, analysis of significant classification model parameters highlights further gender bias. More specifically, some negative words like 'attitude' or positive words like 'beautiful' and 'sweetest' are only strongly significant (top 50 parameters) for women**

Words that are POSITIVE predictors of strong ratings when MEN are mentioned only	Words that are POSITIVE predictors of strong ratings when WOMEN are mentioned only	Words that are NEGATIVE predictors of strong ratings when MEN are mentioned only	Words that are NEGATIVE predictors of strong ratings when WOMEN are mentioned only
['satisfied', 'notch', 'did charge', 'good', 'nicest', 'happier', 'hesitate', 'hesitation', 'exceptional', 'saved', 'pleasantly', 'enjoyed']	['delightful', 'tasty', 'hooked', 'exactly', 'thankful', 'sweetest', 'loved', 'beautiful', 'accommodating', 'efficient', 'fabulous', 'gorgeous']	['condescending', 'wtf', 'excuse', 'shame', 'incompetent', 'tasteless', 'negative stars', 'response', 'ignored', 'wasted', 'unfriendly', 'flavorless']	['lack', 'ridiculous', 'poisoning', 'unhelpful', 'wanted love', 'charged', 'attitude', 'ruined', 'left', 'asked', 'zero', 'acted']

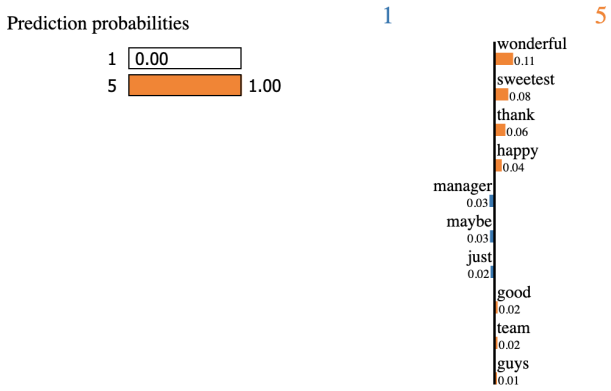
It is common for strong reviews to contain references to women's appearance. It is also common for bad reviews to contain negative words like 'attitude' or 'ridiculous'.
To validate some of these findings (on variable importance), I have used the LIME framework to show word importance at the review level. The below examples show situations where:

A) The word **attitude** carries strong negative weight in reviews that mention women



Text with highlighted words
I was here with my bridesmaids couple days before my wedding. The girl who did my nails had major attitude issue and was very impatient and rude doing mine. I had small rhinestones on some my nails but most of it came off after 3 days when the mail technician told it it would last at least 7 days. Then my thumb started chipping. A day later, my pointing figure which I had extension on completely came off. Basically I needed to redo my nails within a week and half. Ive been to cheaper salons and they all last WAY longer than 1.5 weeks. I will definitely not be coming back o this salon, especially not for the price they charge and the snobby attitude this nail technician has.

B) The potentially sexist word **sweetest** carries strong positive weight in reviews that mention women



Text with highlighted words
I don't know what it is but the coffee I get just tastes so good from here. I usually go for large coffee with almond milk.

However, the 5 starts are for the morning crew that are there around 8:30Am. You guys are absolutely wonderful. Especially the lady with the blue shirt on (manager maybe) she's just the sweetest and happiest person ever. It's not fake happy either. I was a Starbucks fan before but I've just turned to team Dunkin!!!

Looking forward to my next coffee and thank you for the wonderful service!!

2. **List the shortcomings of the method you used and suggestions for improvement.**

   We are currently in a world where we are aware of bias and have the tools to create fair synthetic datasets, but we just need stakeholders to take the lead (Papakyriakopoulos et al, 2020). Indeed, I have used 2 methods to highlight bias in less than a week (both using word2vec generated nearest neighbors and then interpreting classification model parameters); these are simple methods and are within reach for any serious business which uses NLP.

   However, acting upon these findings to remove bias from training data or from models requires a lot of resources (e.g., manually adjusting embeddings, synthetically adapting datasets). I believe that an authority should require that businesses perform bias "stress tests" similar to those that banks undergo for solvency following the 2008 crisis (https://www.investopedia.com/terms/b/bank-stress-test.asp). All major US banks must undergo such tests, with no exceptions to ensure that their financial position falls within an acceptable range. I believe that at a minimum, some independent entities should publish the following results for transparency purposes:
   - Description of algorithm inputs and outputs (explaining the model's usage)
   - Bias report score

   Reference

Papakyriakopoulos et al. - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency - 2020

   Some other observations:
   - Practitioners often use pre-trained models. In such situations, it is impossible to know exactly which text elements led to bias (nor is it practical to share the full corpus that led to training)
   - Creating synthetic datasets may lead to lost information (as noise is added)

3. **Give a hypothetical example of how gender bias in embeddings can lead to 1) an allocational and 2) a representational harm.**

   Allocation harm (when an automated system allocates resources or opportunities unfairly to different social groups) may occur in a situation where:
   - A possible example of allocation harm could occur in credit origination for restaurants and small businesses.

- I could imagine a world in which filings with references to female ownership might have lower success rates (given that lower proportions of women have historically submitted such files)
- Given the unbalanced nature of the dataset, there are high chances that the model would reject applications that do not closely resemble past successes (who were mostly male owners)

Representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether. Examples with this dataset are:

- Businesses with higher Yelp reviews appear higher in recommendations (e.g., in Google search), while those with lower ratings get lower exposure
- Such a situations might happen if a restaurant gets poorer ratings because of a female employee's perceived 'attitude' (see question 1)
- This lower star rating would in turn lead to lower likelihood of a restaurant to appear high in search results (hurting traffic)

Running py scripts

preprocess.py

```
(base) [14:21:45] louisgenereux:HW4 git:(main) $ python preprocess.py
- JSON format review data has been read
- Data converted to pd format
- Gender references identified in reviews
- Gendered df subset created
- Text pre-processed for gendered df
- DF saved to csv
```

embeddings.py

```
(base) [14:35:29] louisgenereux:HW4 git:(main) $ python embeddings.py
- Gendered CSV read, entries referencing both genders are removed
- All words of corpus added to list of list
- READY FOR WORD2VEC MODELING!
+ Created skipgram models with 100 embeddings and window size 5
NEAREST NEIGHBORS FOR: lady
woman, 0.943
girl, 0.902
gal, 0.821
gentleman, 0.796
guy, 0.736
gent, 0.705
fella, 0.68
person, 0.675
associate, 0.671
man, 0.67
---   ---   ---   ---   ---   ---   ---   ---
NEAREST NEIGHBORS FOR: sales
finance, 0.798
salesperson, 0.712
salesman, 0.7
```

gendered_prediction.py

```
(base) [14:24:13] louisgenereux:HW4 git:(main) $ python gendered_prediction.py
- JSON format review data has been read
- Data converted to pd format
- Gendered CSV read, entries referencing both genders are removed
- Summary of corpus:
      male_present female_present pct_male pct_female
                sum            sum
stars
1.0          96876         100370    7.672      7.948
2.0          39107          40362    5.497      5.674
3.0          37504          34401    4.047      3.712
4.0          68114          52955    3.548      2.758
5.0         144697          93389    3.793      2.448
- Male only and female only gendered corpa created
- Running LOGISTIC REGRESSION
```

```
- Testing result for men:
Acc:  0.96608  Prec:  [0.959 0.971]  Rec:  [0.956 0.973]  f1:  [0.958 0.972]
- Testing result for women:
Acc:  0.96756  Prec:  [0.969 0.966]  Rec:  [0.969 0.966]  f1:  [0.969 0.966]
- Top predictors of high ratings male
   Weight? Feature  feature_number      feature_name weight_num
0  +29.460   x1512            1512           amazing      29.46
1  +28.737  x20183           20183         delicious     28.737
2  +23.498  x35392           35392             great     23.498
3  +22.875  x26310           26310         excellent     22.875
4  +22.729   x4558            4558           awesome     22.729
5  +22.598   x6391            6391              best     22.598
6  +21.915  x63282           63282           perfect     21.915
7  +20.004  x27823           27823         fantastic     20.004
8  +17.652  x63826           63826        phenomenal     17.652
9  +17.307  x39550           39550  highly recommend     17.307
```

- Top predictors of low ratings female

	Weight?	Feature	feature_number	feature_name	weight_num
92	-15.513	x7518	7518	bland	-15.513
93	-17.162	x92557	92557	used love	-17.162
94	-17.427	x21952	21952	disgusting	-17.427
95	-19.363	x4944	4944	awful	-19.363
96	-19.525	x86045	86045	terrible	-19.525
97	-21.000	x21844	21844	disappointing	-21.0
98	-21.163	x39449	39449	horrible	-21.163
99	-21.374	x92134	92134	unprofessional	-21.374
100	-28.044	x73048	73048	rude	-28.044
101	-33.147	x98786	98786	worst	-33.147

- Words that are POSITIVE predictors of strong ratings when MEN are mentioned only:
['exceptional', 'pleasantly', 'hesitate', 'happier', 'enjoyed', 'saved', 'hesitation', 'good', 'notch', 'nicest', 'satisfied', 'did charge']
- Words that are POSITIVE predictors of strong ratings when WOMEN are mentioned only:
['fabulous', 'hooked', 'beautiful', 'delightful', 'tasty', 'exactly', 'gorgeous', 'loved', 'thankful', 'efficient', 'accommodating', 'sweetest']
- Words that are NEGATIVE predictors of strong ratings when MEN are mentioned only:
['ignored', 'tasteless', 'unfriendly', 'excuse', 'condescending', 'negative stars', 'wasted', 'shame', 'flavorless', 'wtf', 'response', 'incompetent']
- Words that are NEGATIVE predictors of strong ratings when WOMEN are mentioned only:
['ridiculous', 'zero', 'left', 'ruined', 'lack', 'unhelpful', 'poisoning', 'charged', 'wanted love', 'asked', 'acted', 'attitude']