Gender Classification in 19th Century Literature

Alex Kimbrell and Tori Pirtle

December 6, 2019

Abstract

Novels that were written during the 19th century were pivotal in shaping Literature today as we know it. These novels contrast from one another due to their genre, subject matter, and author. We would like to determine if differences in these novels stem from the gender of the author that wrote them. We use a variety of attributes for each author along with words that are associated with each gender in order to determine if there is a systematic difference between the literary works of male and female authors. We apply various models to our data to aid in helping us reach a conclusion.

1 Introduction

1.1 Topic

Our original problem that we were given was to determine if there were any systematic differences to distinguish novels written by male and female authors in the 19th century. Thus, we wanted to find these differences between the novels, so we looked at a variety of factors that could potentially help us. In addition, we later utilized a list of words deemed "female" and "male" to see if it could help us determine the distinguishable factor.

1.2 Theory

After speaking with our domain experts for the literary time period, we had a theory on what items could distinguish novels between male and female authors. We surmised that female authors are more likely to have longer, more descriptive sentences, more dialogue, and use adjectives more frequently. Male authors, on the other hand, are expected to have shorter sentences, less dialogue, and not use adjectives as frequently. We determined that we needed to calculate four different variables for each novel in our dataset: average sentence length, average length of quotations, number of quotations used, and the ratio of adjectives to the number of total words.

1.3 Data

Our domain experts also compiled a list of novels written in the 19th century broken down by author. We used this information to form our dataset. The data was aggregated from a variety of sources and made available to us on Project Gutenberg. It contains 215 instances (novels), with 6 variables each. Four of the variables were calculated by us while the other 2 variables were title of the novel and the gender of the author (which we had to find online). The

four variables calculated by us are: average sentence length, average number of quotations, number of quotations used, and the ratio of adjectives to the number of total words in a novel. The title of the novels stems from the information previously given to us by our domain experts. The dataset has no missing data, but is very much skewed since the majority of novels in our dataset have male authors (188) rather than female authors (27).

	Novel	Male/Female	Number Of Quotes	Average Quote Length	Adjective Ratio	Average Sentence Length
0	A Christmas Story	Male	49	5.755102041	0.05386552506	14.15539066
1	ecticut Yankee in King Arthu	Male	1251	25.54036771	0.0534505004	18.72300542
2	uet With an Occasional Cho	Male	83	11.36144578	0.06328709254	11.56114398
3	koning_A tale of Bush Life	Male	2130	21.6	0.03833040603	16.47139224
4	A Jacobite Exile	Male	1678	34.74016687	0.04017649292	19.99491008
5	White Cross_A tale of the S	Male	1543	48.88788075	0.04326394216	24.03685423
6	A Lady of Quality	Female	1499	16.50233489	0.05106634268	21.31205324
7	Laodicean_A Story of Toda	Male	47	15.40425532	0.05429400717	16.73111032
8	A Pair of Blue Eyes	Male	153	17.15686275	0.05714337594	14.22726302
9	A Search for a Secret Vol. 1	Male	434	82.34101382	0.05441615856	23.51088963
10	A Search for a Secret Vol. 2	Male	670	43.03731343	0.0461739105	21.73150582

Figure 1: Sampling of Our Data

1.4 Problem Definition

Our problem was to use an assortment of novels from the 19th century to compare and contrast works written by male and female authors to determine if there is any systemic difference to distinguish them. In order to solve this problem, we decided to train various models to see which had the highest accuracy. We trained our models on two different datasets: the complete dataset (unbalanced) and the balanced dataset where we had 50% women authors and 50% male authors (54 instances). We utilized the following linear machine learning models: Logistic Regression, Naive Bayes, Decision Tree, and SVM.

2 Initial Analysis/Preprocessing

In our initial analysis, we began by finding the average values for the four main variables that we were looking into: average sentence length, number of quotations used, average length of quotations, and the average adjective ratio. By doing this, we were able to see whose average was higher between the two groups. For the average sentence length, average number of quotations, and average adjective ratios, we found that the average for females was higher than the average for males, as expected by our domain experts.

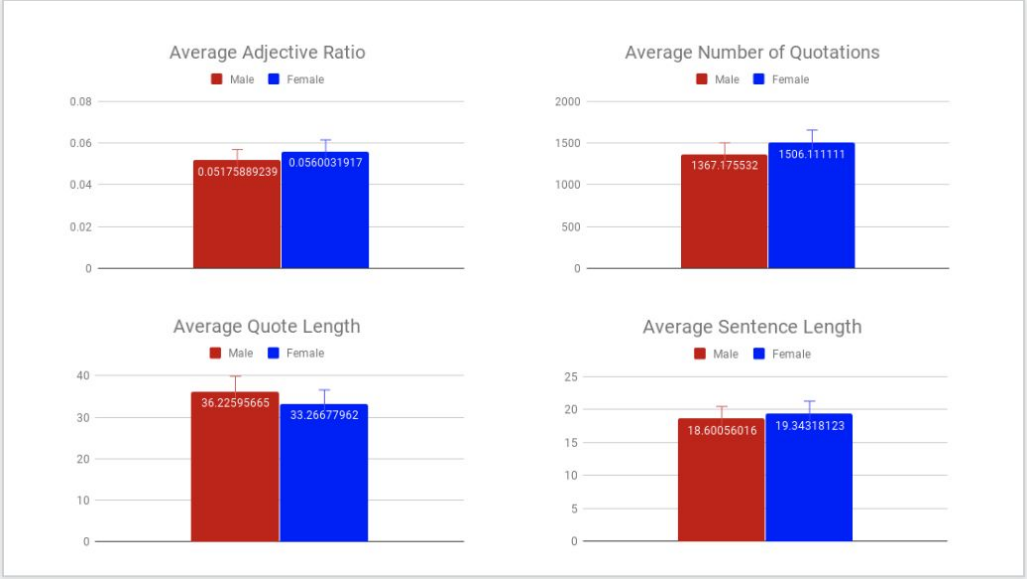


Figure 2: Average Values for Variables

However, the results for the average quote length did not have a higher average for females than males as our domain experts thought it would. After calculating these averages, we ran the data on four different models to figure out which model had the highest accuracy.

3 Results

3.1 Balanced vs. Unbalanced Dataset

Our original dataset of 215 novels only contained 27 that were authored by female writers. Therefore, a model that guessed every instance as a male would achieve an accuracy of approximately 87.5%. In order to deem any of our models 'good', they would need to score *at least* as high. Only Logistic Regression and Support Vector Machine were able to achieve accuracies higher than 87.5%, however, neither model was demonstrably better.

Train/Test	Logistic Regression	Naive Bayes	Decision Tree	SVM
30/70	86.0927	85.4305	77.4834	86.0927
50/50	90.7407	75.0000	84.2593	89.8148
70/30	92.3077	80.000	75.3846	92.3077

We also tested each model on a balanced dataset which aggregated the 27 female novels with 27 novels randomly sampled from the 188 male novels. Testing the models on a "balanced" dataset provided variable results depending on the sample of male novels. Our results from the balanced dataset indicate that Naive Bayes and Decision Tree provided the most significant results.

Train/Test	Logistic Regression	Naive Bayes	Decision Tree	SVM
30/70	50.0000	63.1575	65.7895	60.5263
50/50	51.8519	59.2593	62.9630	40.7407
70/30	47.0588	70.5882	70.5882	35.2941

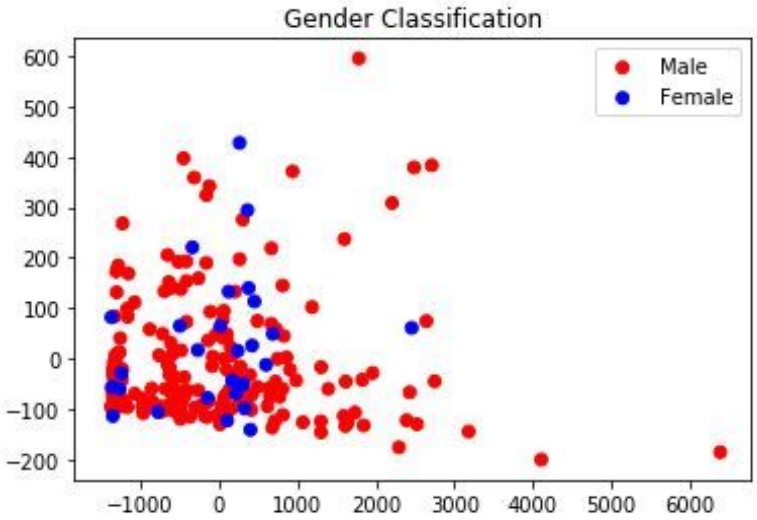
3.2 Additional Features

Our data was initially represented through four features (average sentence length, average length of quotations, number of quotations, and the adjective ratio), however, we decided to add two features: masculine word frequency and feminine word frequency in order to further define the structure of each novel. Masculine words included words like “king”, “civil”, and “absolutely” while Feminine words contained words such as “china”, “dress”, and “lace”.

Train/Test	Logistic Regression	Naive Bayes	Decision Tree	SVM
30/70	65.7895	68.4211	63.1579	76.3158
50/50	70.3704	70.3704	62.9630	44.4444
70/30	82.3529	76.4706	70.5882	47.0588

3.3 Data Shape

In an attempt to learn more about the structure of our data, we applied Incremental PCA to our dataset of 215 Novels to shrink the dimensionality from six to two dimensions.



4 Analysis

The previous scatter plot indicates that the male and female clusters contain a considerable amount of overlap since neither group is statistically separable using our six features. The inclusion of masculine and feminine words provided an approximate 20 percent increase in accuracy for Logistic Regression and SVM with Logistic Regression scoring the highest overall accuracy of any model at 82.3%. The weights of the Logistic Regression model we used supported that the most significant features were Average Sentence Length, Feminine Words, and Masculine Words (in that order) with Average Sentence Length being nearly an order of magnitude more significant than any of the other features. More specifically, a greater presence of feminine words was indicative of a female writer, and a greater presence of masculine words was indicative of a male writer. The Number of Quotations, Average Length of Quotation, and Adjective Ratio had relatively little impact in the decision making of our Logistic Regression model.

What is more interesting is our assumption that female authors tend to have longer sentences on average holds for the data we've collected. A longer average sentence length *was* symptomatic of female writing. This likely indicates that 19th century female writers are indeed more descriptive and less concise than their male counterparts. It could also indicate that feminine words are generally more difficult to describe in fewer words. For example, a masculine word such as "king" paints a cognitive picture of leadership, respect, kingdom, loyal subjects, knights, etc. while a feminine word such as "china" may require the author to give more information to make the mental image of the china more clear for the reader. What kind of china is it? What color is the china? What distinguishing patterns/designs are present? These extra

details could provide a hypothesis for why women authors tend to write longer sentences than male authors do on average.

5 Conclusion

Our research shows that the most distinguishable aspects of male and female literature in the 19th century are word choice and sentence length. Therefore, we can conclude that there are systematic differences that distinguish novels written by female and male authors in the 19th century for the specific novels in our dataset. Our conclusion can not be applied to all novels written in the 19th century because our models would need to be retrained to account for the new data.

We did encounter some limitations as we completed our analysis. Our models would have been more accurate if we had more data. 215 data values is not a very large dataset, and with more data, we would have had more data to train the models on. In addition, another limitation was that the data was very much skewed towards male authors which could have affected our results. If our project had not been limited to only Project Gutenberg, we might have been able to account for this skewness and more successfully balance the entire dataset.

References

DigiUGA. *Project Gutenberg*. 2019. DigiUGA/Gutenberg_Text. Retrieved from https://github.com/DigiUGA/Gutenberg_Text.

Ha, T.-H. (2017, March 27). The words that give away a writer's gender, in classic works of literature. Retrieved November 2019, from https://qz.com/934268/words-that-give-away-a-writers-gender-from-ben-blatts-book-nabokovs-favorite-word-is-mauve/.

Sci-Kit Learn. 2019. Retrieved from https://scikit-learn.org/stable/.