Author Gender Identification of English Novels

Joseph Baena and Catherine Chen

December 13, 2013

1 Introduction

Machine learning algorithms have long been used in studies of authorship, particularly in order to identify the author of a disputed work out of a group of possibilities. In this study, our aim was to classify the authors of a large selection of 19th century American and British novels based on gender. Since the goal was to classify the texts into two large groups, rather than to perform a more granular authorship test, this problem closely resembles that of spam email classification and other similar binary classification problems than the canonical studies of literature authorship. In attempting to classify novels by gender overall, we assume that there exists a significant difference between male and female writing in 19th century novels that is persistent despite variations in individual style and novel genre. We believe that this question is interesting and relevant because differences between male writing and female writing can reveal more fundamental facts about the role of these sexes within these societies. The 19th century Western literary world is an interesting example due to the societal pressures preventing women from publishing and the prevalence of female authors writing under male pseudonyms. The same methods may be extended to investigate other societies and time periods, in addition to other types of writing outside of novel.

2 Data Collection and Methodology

We began with a set of text files of selected complete novels from the 19th century English corpus and relevant metadata including author gender, provided to us by the Stanford University Literary Lab. The Literary Lab had used Optical Character Recognition (OCR) to scan the novels and save as plain text files, and as a result some of the words were not read in properly. Our first step of data processing was filtering out the files with OCR accuracy below 90%.

In order to perform our statistical analysis on the novels, we first wrote a Java program that calculated the frequencies of words within each text file and created a sparse matrix in which each row represented a novel, each column represented a word, and each entry represented the word's frequency in the novel. In obtaining word frequencies, we ignored words that appear fewer than three times in the novel, as done in Glenn Fung's 2003 paper "The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization." With these frequencies, we were able to select words on which to focus our machine learning algorithms.

Lexicon of 70,000 English Words

For our first attempt of feature selection, we found the frequencies of words appearing in a comprehensive lexicon containing over 70,000 English words of at least 3 letters in length. For the analysis, we created four training sets of size 500, 800, 1200, and 2600, each of which consisted of male-authored works and female-authored works evenly. We then ran the training and testing sets through a Naive Bayes classifier with +1 Laplace smoothing using a multi-variate Bernoulli event model.

Our attempt to classify the novels by gender based on the large lexicon was relatively unfruitful. The Naive Bayes classifier achieved an accuracy of less than 50% when trained on the training set of 500 novels. Accuracy did not scale with larger training sets; after learning from our largest training set, the Naive Bayes algorithm mis-classified over half of the works. This result suggests that when considering the entire lexicon of over 70,000 English words, machine learning algorithms are unable to notice a notable difference between male and female writing. We believe this is a reasonable conclusion since there is such a large variance in style and diction among different authors, regardless of gender. Further investigation revealed that the algorithm was overfitting to the training set, since many of the words appeared only in one or two of the training examples and not at all in the testing set. For instance, the words that were most indicative of male writing, as defined by the probability of the work being labeled male given that the word was used, were words that appeared in a single work that was written by a man. The sparseness of useful features in this model led us to believe that we needed to modify our lexicon, following the example of prior work in authorship studies, to consist of a smaller set of more common words.

Lexicon of 12 Gender-Specific Words

Given the dearth of productive learning with our previous choice of features, we set out to improve the accuracy of our models by limiting the feature space to include words that would be used more frequently by multiple authors. We were inspired to create a small lexicon by the research of Bosch and Smith, who used a set of 70 function words - in particular, widely-used pronouns, articles, and prepositions - as their feature space to classify the Federalist Papers by their authors. Since our study was focused on gender, we believed that using explicitly “gendered” words might be more indicative than ordinary function words.

For our second analysis, we used a lexicon that contained only twelve words: *he, his, him, man, male, men, she, her, hers, woman, female*, and *women*. We again classified the works using the multi-variate Bernoulli Naive Bayes classifier, as well as the Support Vector Machine (SVM) with a linear kernel.

Lexicon of 13 Words Chosen by Forward Search

Hoping to improve the classification accuracy of our models, we developed another lexicon by studying the 149 most common English words, as presented by the Oxford English Corpus. These consisted largely of function words, along with select nouns, verbs, and adjectives. We implemented forward search feature selection using 10-fold cross validation on the Naive Bayes algorithm to determine the most-predictive features out of the 149 most common words in the English language.

With 10-fold cross validation, the data is equally divided into 10 sets. For each of the subsets of features, we train on 9 of the sets and test on the remaining data set. We repeat this process a total of 10 times and report the mean classification error. Through this process, we are able to determine which subset of words in the lexicon have the lowest empirical error and provide the most information for the classification algorithm. Using forward search feature selection, we obtained the following set of 13 words: *when, can, no, because, case, bad, her, so, which, a, that, not*, and *he.*

3 Results and Findings

Using the lexicon consisting of 12 strongly gender-specific nouns, pronouns, and adjectives, we obtained the following results with the Naive Bayes classifier and the SVM with a linear kernel:

Training Set Size	Naive Bayes Accuracy	SVM Accuracy
500	77.2%	79.4%
800	77.3%	81.0%
1200	77.8%	80.6%
2600	78.0%	80.1%

As the above table illustrates, there is a slight increase in accuracy as the training set size increases for both Naive Bayes and SVM, but the marginal accuracy gain is not very significant, and in some cases, there is actually a drop in accuracy when the size of the training set increases. We believe these to be idiosyncratic results due to some aspect of the specific works chosen for the training sets. Nevertheless, the accuracies obtained by the two learning algorithms were relatively consistent. The SVM consistently outperformed Naive Bayes which was not surprising given the relatively aggressive optimization in the SVM process.

As an additional data point, we filtered our novel data to include only those written by female authors under male pseudonyms, which we were able to do from the metadata that provided us the name used for publication associated with each novel. Our largest training set yielded a 79.6% accuracy on this set of works; when we tested set with the same number of works from a random combination of male and female authors, we obtained an accuracy of 77.3% Although the sample was relatively small at 44 novels, the fact that the Naive Bayes classifier was able to detect 36 of the works' authors as female suggests that these women, who attempted to hide behind male identities, still wrote in a fashion that was somehow inherently female.

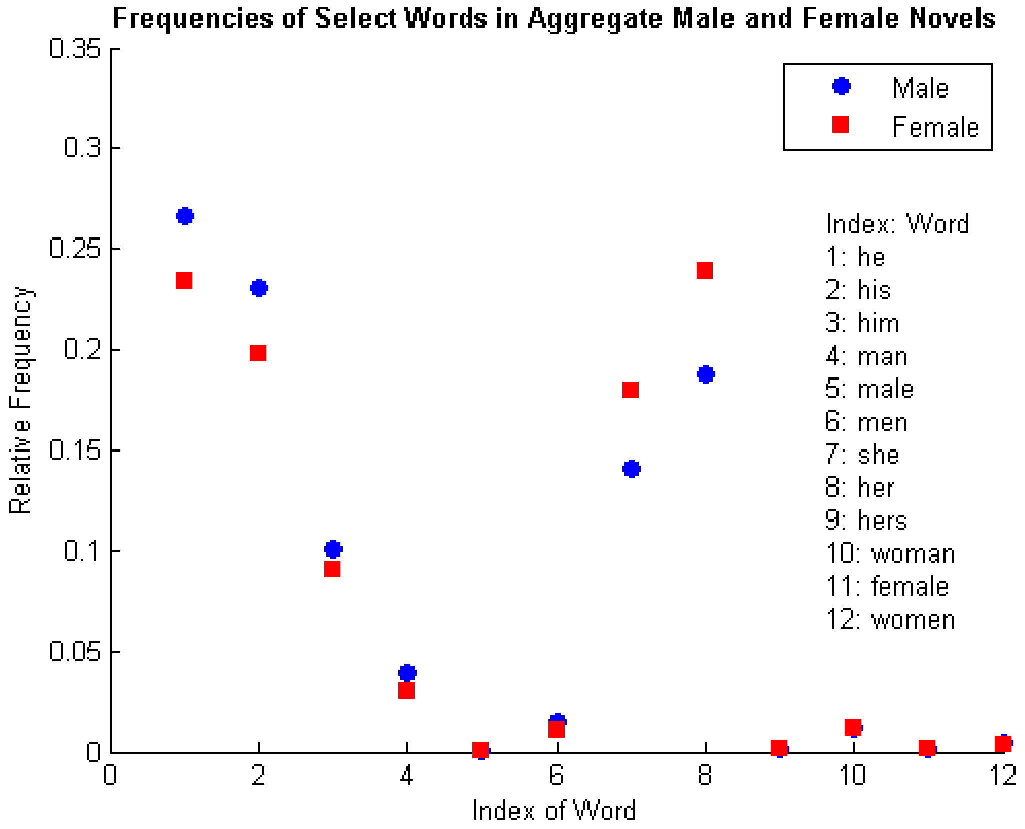

Figure 1: Relative usage frequencies of the 12 gender-specific words among male and female authors of the 19th century.

From the aggregate frequencies, we observe the relative frequencies of usage of the 12 gendered words, illustrated in Figure 1. As the plot shows, male authors use “he” and “his” much more often than female authors, and the reverse is true for words like “she” and “her.”

Finally, using the lexicon of 13 words chosen by forward search feature selection, we obtained results that represent a notable improvement from the earlier Naive Bayes results, more closely resembling the SVM accuracies. The classifier with only the 13 features selected from forward search achieved the following accuracies:

Training Set Size	Naive Bayes Accuracy
500	81.2%
800	81.3%
1200	81.2%
2600	81.6%

4 Conclusions and Further Work

The promising results of our study suggest that machine learning can indeed be used to determine the gender of a novel's author. We found that the choice of words to analyze is a very important factor in the accuracy of the Naive Bayes and SVM classification algorithms. Smaller lexicons were much more effective in predicting in author gender than the large, comprehensive lexicon. Additionally, the choice of specific words included in the model had a large impact on the classification accuracy. The results from our hand-picked lexicon of gender-related words indicate that male authors are more likely to use male pronouns (such as "he" and "his") than female authors. In a similar fashion, female authors are more likely to use female pronouns (such as "she" and "her") than male authors. This conclusion is further supported by the fact that the set of 13 words out of the 149 most common English words found via the forward search feature selection algorithm include the words "he" and "her."

Our research in automatic author gender classification through machine learning techniques is a stepping stone for further research in the field. One avenue of research is further refining the choice of words to analyze in the texts. We believe that experimenting with other methods of feature selection, for instance, can further refine the accuracy of gender prediction in a variety of classification algorithms. Another direction for research is to investigate other properties of writing style aside from word frequency, such as phrasing, sentence structure, and the names of main characters (in the case of fiction novels). It would also be potentially interesting to see whether training on a data set from the 19th century has the same level of performance when testing on more modern works of literature, or to conduct the same study on literature in other languages with explicitly gendered nouns.

5 Acknowledgements

We would like to thank Professor Andrew Ng and the entire CS 229 course staff for their guidance during the course. We are also very grateful to Ryan Heuser, Amir Tevel, and the Stanford Literary Lab for providing the 19th century English novel corpus for our research.

6 References

Bosch, Robert A. and Smith, Jason A."Separating Hyperplanes and the Authorship of the Disputed Federalist Papers." The American Mathematical Monthly , Vol. 105, No. 7 (Aug. - Sep., 1998), pp. 601-608.

Fung, Glenn. "The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization." 25 Oct 2003.