Determining the Gender of Shakespeare's Characters

Matt Olson
maolson@stanford.edu
Stanford University, CS 229

Abstract

The purpose of this project was to determine the gender of characters in Shakespeare's plays by applying machine learning algorithms to the characters' word usage. In particular, a naive Bayes model and support vector machine model were implemented on a corpus consisting of the words spoken by 198 characters – 99 male and 99 female. Both models achieved reasonably good classification accuracy, with SVM unsurprisingly outperforming naive Bayes, 76.3% to 69.2%. Interestingly, characters from tragedies were classified with much greater accuracy than were characters from either comedies or histories.

I. Introduction

Work on texts of several different types, including email and nonfiction articles, has shown that men and women use language differently [1, 2]. Moreover, these distinguishing language features appear to be consistent across studies. Corney et al. [1] suggest that men tend to engage in "report talk" that is assertive and proactive, whereas women more often favor "rapport talk," which is marked by agreeing, understanding, and supporting the attitudes of other speakers. Argamon et al. [2] similarly posit that male speech is generally in a more "informational" style, and female speech in a more "involved" style.

The main question posed by this project was whether differences in language use could be observed in literary characters of different genders, in particular in the plays of William Shakespeare. Shakespeare's characters are often held up as examples of authentic humanity, so the goal of this project was to determine whether his female characters could be distinguished from his male characters on the basis of language, in the same way such distinctions have previously been made among real people. The problem is slightly different in this case because the goal is to examine text written by a single person meant to express the thoughts of people of different genders, rather than examining texts that are actually written by people of different genders. Previous work by Hota et al. [3] on a Shakespearean corpus suggests this is feasible, as they were able to predict a character's gender with accuracy in the range 56.3% to 74.3% using a variety of models and feature sets. The current project was able to offer a modest improvement on these results.

Reasonable classification accuracy makes it possible to examine the distinguishing features of male and female speech in Shakespeare's characters, and to consider whether these differences are similar in quality to the ones found in previous analyses of modern writing. The most discriminating features of male and female speech can also be used to gain some insight into how Shakespeare generally represented gender differences in his plays.

A secondary objective of this project was to evaluate the suitability of naive Bayes and SVM models for the task of text classification according to gender. Both have been shown to be effective in distinguishing spam emails from genuine ones, but the features distinguishing male speech from female speech seem likely to be more subtle. Both models proved reasonably adept at predicting

characters' genders, a result that is perhaps surprising given the relatively straightforward way in which they made their predictions, relying only on word counts and without any sort of semantic understanding.

II. Construction of the Corpus

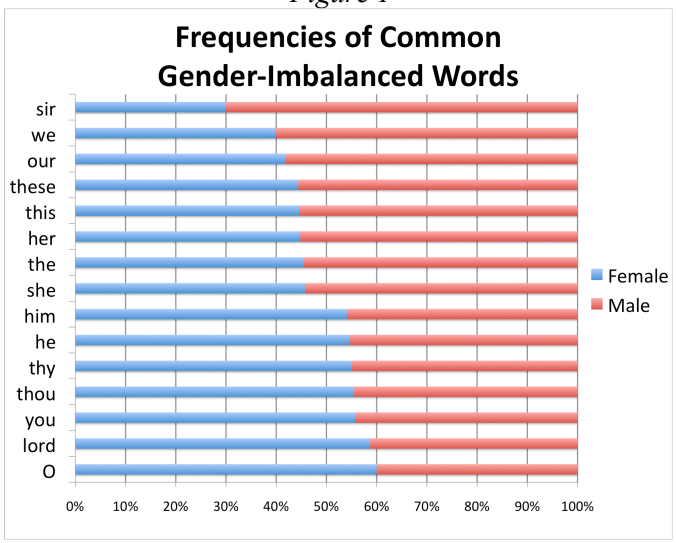
The corpus was assembled in the following manner. From each play, I selected every female character with at least 30 speaking lines, in an attempt to limit training examples to those with a reasonably large number of features. This resulted in a total of 99 female characters to be added to the data set. I then selected an equal number of male characters from each play as there were qualifying female characters. The male characters were chosen so that the number of lines they spoke was as close as possible to the number of lines spoken by their female counterparts from the same play. Since males vastly outnumber females in Shakespeare's plays and generally have larger roles, this ensured that the corpus was balanced relatively equally between the genders rather than being tilted heavily toward the male side. I hand-labeled each character as either male or female to furnish the training example labels for the learning algorithms.

The texts of the plays were acquired from Open Source Shakespeare [4], which uses the well-regarded and public domain 1864 Globe Edition of Shakespeare's complete works. For each character, I created a text file consisting of all their speeches concatenated together and cleaned up the file by removing stage directions. Using the JFreq application [5], I created a word frequency matrix from these text files, in which the *j*th cell of the *i*th row represented the number of times word *j* was spoken by character *i*. I ran JFreq both with a built-in stemmer and without. Both naive Bayes and SVM wound up performing 1-2% better with stemming; for conciseness, I omit the non-stemming results in this paper. The word frequency matrix and manually entered labels for each character were used as the inputs to the machine learning algorithms.

III. Gender Classification

To get an initial idea about the feasibility of differentiating male characters from female ones, I sorted the word frequency matrix to get a list of all words that appeared at least 800 times in the corpus. I then looked to see whether any of these common words were used disproportionately by one gender. The reasoning was that with so many examples of these words, even a 55-45 split from one gender to the other would testify to the existence of features that a learning algorithm could use to get a robust idea as to the gender of a character who used those words more or less frequently. *Figure 1* contains the 15 most gender-imbalanced words from this collection. Even in the absence of machine learning, there are some interesting things to notice in these data. For example, Shakespeare's males are more likely to use the determiners "this" and "these," which may be seen as indicators of their more assertive, informational language. Female characters, on the other hand, are more likely to use the second person pronouns "you" and "thou," which could be interpreted as a sign of more involved interaction with other speakers. It should also be noted that both males and females are more likely to use the pronouns of the opposite gender.

Figure 1



Having established the existence of discriminating features, I implemented a naive Bayes classifier using the multinomial event model and Laplace smoothing as a first attempt at classifying the data. When trained on only half of the data and tested on the other half, it achieved a classification accuracy of 66.7%. Since the number of examples in my data set was fairly small compared to many data sets used in machine learning applications, I decided to maximize the number of examples available for training by running leave-one-out cross validation on my naive Bayes model. The increase in training examples increased classification accuracy across the data set to 69.2%.

The naive Bayes model can also be used to get an informal sense of how indicative particular words are of either a male or female speaker. By running a classifier on the entire data set and taking

$$\log\left(\frac{P(word \mid female)}{P(word \mid male}\right)$$

for each word in the corpus, one can find the words most associated with female and male speakers. *Table 1* lists the 15 words that were most indicative of each gender on this metric. Proper names were omitted because they are almost always exclusive to a single play. It may be interesting to note, however, that proper names actually occupied the first 10 spots for males, compared to only six of the first 10 for females. Together with the common words from *Figure 1*, the following gender-imbalanced words offer a bit of insight into the character of Shakespeare's men and women.

Table 1

Female	**Male**
willow	sore
dress	consul
gather	bull
trow	apprehend
bore	singular
handkerchief	file
babe	legion
folk	mayor
maidenhead	chat
ambassador	varlet
vial	fourth
rot	harmless
meek	coronation
petticoat	cheese
pinch	supper-time

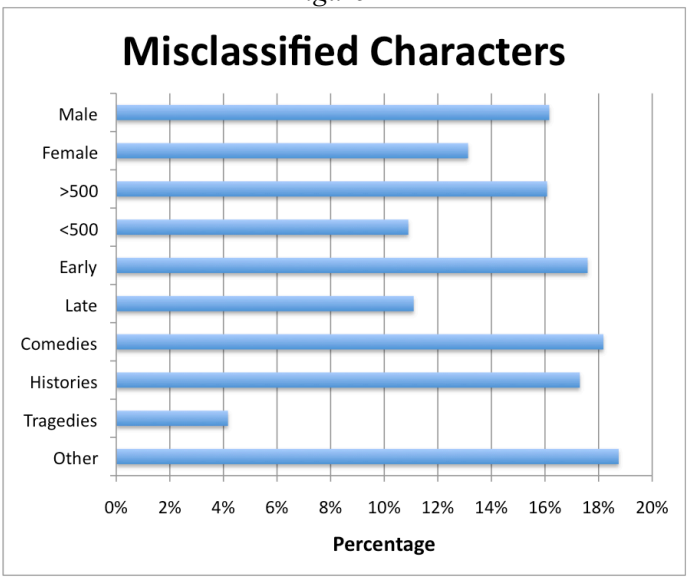
In addition to naive Bayes, I ran a support vector machine with a linear kernel using the LIBLINEAR library function [6]. When trained and tested on the same halves of the data as the initial naive Bayes model, it achieved a classification accuracy of 74.2%. On leave-one-out cross validation, its accuracy increased to 76.3%. In both cases, SVM did noticeably better than naive Bayes. Both models, however, are near or above the highest accuracy reported by Hota et al. [3].

IV. Discussion

It seems clear from the success of the learning algorithms that Shakespeare did indeed write his male and female characters to be both distinct from the opposite gender and consistent with other members of their own gender. Naive Bayes and SVM are not able to offer any sort of semantic or literary analysis, but they can expose some suggestive lexical features. Many of the words in *Table 1* that were identified by the model as especially symptomatic of one gender or the other are strongly associated with that gender in the real world, or least in the world of Shakespeare's time. "Babe" and "petticoat" are examples on the female side, along with "consul" and "mayor" on the male side. There are also words whose associated gender seems counterintuitive, such as "ambassador" for females and "chat" for males. It seems clear, however, that the learning algorithms under consideration are detecting real and significant differences in the language use of male and female characters.

One interesting question to ask is whether certain types of characters are more or less likely to be classified incorrectly by these models. Out of the 198 total characters, 29 were misclassified by both naive Bayes and SVM. *Figure 2* shows the percentage of characters from various categories who were misclassified by both models.

Figure 2



>500 and <500 indicate characters who speak more and less than 500 total words, respectively. "Early" characters are from the first half of Shakespeare's career, and "late" characters are from the second. The genres are self-explanatory, but it should be noted that for a good number of Shakespeare's plays, there is no consensus on categorization. If there is not general consensus that a play is a comedy, history, or tragedy, I am considering it under "other."

As seen from *Figure 2*, characters who speak less than 500 words, despite having fewer features for the algorithm to use, are actually less likely to be classified incorrectly than characters who speak more than 500. One possible explanation for this is that Shakespeare's minor

characters are more likely to conform to predictable linguistic patterns given their gender, since they are filling relatively small and mechanical roles in the plays. Another interesting thing to note from *Figure 2* is the relative success of the models on characters from tragedies as compared to other genres. Only two out of 48 tragic characters were misclassified by both naive Bayes and SVM. There is no obvious explanation for this, but perhaps the greater emotional weight of tragedies means that characters use words that mark their positions in the play's society more clearly. In contrast, Shakespeare's comedies often involve cross-dressing and other confusion and playfulness around the issue of gender, so it is possible that his comedic characters are more likely to adopt the speech patterns of the opposite gender. Further investigation in this line of inquiry could prove fruitful for literary scholars working in conjunction with machine learning analysis. However, it bears remembering that the number of examples in this data set is small enough that these patterns may simply be due to chance.

A final thing to consider is the level of confidence with which the gender of each character was predicted. *Table 2* lists the 20 characters classified most confidently by the naive Bayes classifier. Other than Orlando, who was mistakenly classified as female, all these predictions were correct.

Table 2

Character	**Play**	**Gender**
Bianca	*Othello*	Female
Chatillon	*King John*	Male
Octavia	*Antony and Cleopatra*	Female
Phebe	*As You Like It*	Female
Desdemona	*Othello*	Female
Virgilia	*Coriolanus*	Female
Dull	*Love's Labor's Lost*	Male
Second Merchant	*The Comedy of Errors*	Male
Mariana	*Measure for Measure*	Female
Duchess of York	*Richard II*	Female
Margaret	*Much Ado About Nothing*	Female
Bishop of Carlisle	*Richard II*	Male
Lavinia	*Titus Andronicus*	Female
Charmian	*Antony and Cleopatra*	Female
Orlando	*As You Like It*	~~Female~~
Gadshill	*Henry IV, Part 1*	Male
Valeria	*Coriolanus*	Female
Emilia	*Othello*	Female
Duchess of York	*Richard III*	Female

The model appears to be more confident classifying characters as female than as male, with 15 of its 20 most confident predictions doing so. In terms of characters from tragedies, *Table 2* makes sense alongside *Figure 1*. Eight of the 20 characters listed above are from tragedies, despite the fact that tragic characters make up less than one-quarter of all the characters in the data set. It appears as though the gender of tragic characters is indeed easier to determine than the gender of characters from other genres.

V. Conclusion

This project has shown that the relatively straightforward approaches of using naive Bayes and support vector machines can be effective for differentiating between literary characters of different genders. It has shown that Shakespeare's male and female characters can be recognized on the basis of their language, and it has lent support to previous work on linguistic differences between males and females in general.

It would be interesting to apply the models trained on data from Shakespeare to the male and female characters of other contemporary playwrights, such as Christopher Marlowe and Ben Jonson. I think it is likely that these models would achieve accuracy well over the 50% expected with random chance, although they would probably be somewhat less accurate than they are on the Shakespearean characters on which they learned. A more difficult test would be to use the models on modern writings – perhaps emails, blog posts, or newspaper columns – and attempt to predict the gender of the author. There does seem to be some overlap between the characteristic linguistic markers of gender in Shakespeare and those put forward by recent studies on today's texts, but whether that would translate into success for the models developed here is unclear.

VI. Acknowledgements

I would like to acknowledge Professor Andrew Ng and the CS 229 teaching assistants from whom I learned the techniques that made this project possible. Thanks also to Mario Villaplana for helping to shape the idea for this project.

VII. References

[1] M. Corney, O. Vel, A. Anderson, G. Mohay. "Gender Preferential Text Mining of E-mail Discourse." *Proceedings of 18th Annual Computer Security Applications Conference ACSAC*.

[2] S. Argamon, M. Koppel, J. Fine, A. Shimoni. "Gender, Genre and Writing Style in Formal Written Texts." *Text* 23(3), 2003, pp. 321-346.

[3] S. Hota. S. Argamon, M. Koppel, I. Zigdon. "Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters." *Digital Humanities*, 2006, pp. 82-88.

[4] Open Source Shakespeare: An Experiment in Literary Technology. © 2003-2013 George Mason University. http://www.opensourceshakespeare.org/

[5] W. Lowe (2011). JFreq: Count words quickly. Java software version 0.5.4. http://www.conjugateprior.org/software/jfreq/

[6] C. Lin (2013). LIBLINEAR. Version 1.93. http://www.csie.ntu.edu.tw/~cjlin/liblinear/