conventions is exhibited: a character nears death and expires in a room usually full of flowers and mourners who often "swoon." The training set for this experiment includes two other texts that were scored on the same sentimentality scale, Susanna Rowson's 1794 novel Charlotte: A Tale of Truth and Harriet Jacobs's Incidents in the Life of a Slave Girl.

Since these texts were considered sentimental, most chapters were scored in the medium or high range, so the categories were changed to "highly sentimental" and "not highly sentimental." With D2K, the Naive Bayes method was used to extract features from these texts, which we might call markers of sentimentality. Looking at the top 100 of these features, some interesting patterns have emerged, including the privileging of proper names of minor characters in chapters that ranked as highly sentimental. Also interesting are blocks of markers that appear equally prevalent, or equally sentimental, we might say: numbers 70-74 are "wet," lamentations," "cheerfulness," "slave-trade," and "author." The line of critical argument that goes that the sentimental works focus on motherhood is borne out by "mother" at number 16 and "father" not in the top 100.

As we move into the next three phases of the project, we will include stemming as an area of interest in classifying the results. Phase two will use two more novels by the same authors as those in the training set; phrase three may include ephemera, broadsides, and other materials collected in the EAF collection at the UVa Etext Center. Phase four will run the software on texts considered non-sentimental in the nineteenth century and other phases might include twentieth and twenty-first century novels that are or are not considered sentimental. We hope to discover markers that can identify elements of the sentimental in any text.

# Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters

**Sobhan HOTA**
**Shlomo ARGAMON**
*Department of Computer Science, Illinois Institute of Technology*

**Moshe KOPPEL**
**Iris ZIGDON**
*Department of Computer Science, Bar-Ilan University*

## 1. Introduction

A recent development in the study of language and gender is the use of automated text classification methods to examine how men and women might use language differently. Such work on classifying texts by gender has achieved accuracy rates of 70-80% for texts of different types (e-mail, novels, non-fiction articles), indicating that noticeable differences exist (de Vel et al. 2002; Argamon et al. 2003).

More to the point, though, is the fact that the distinguishing language features that emerge from these studies are consistent, both with each other, as well as with other studies on language and gender. De Vel et al. (2002) point out that men prefer 'report talk', which signifies more independence and proactivity, while women tend to prefer 'rapport talk' which means agreeing, understanding and supporting attitudes in situations. Work on more formal texts from the British National Corpus (Argamon et al. 03) similarly shows that the male indicators are mainly noun specifiers (determiners, numbers, adjectives, prepositions, and post-modifiers) indicating an 'informational style', while female indicators are a variety of features indicating an 'involved' style (explicit negation, first- and second-person pronouns, present tense verbs, and the prepositions "for" and "with").

Our goal is to extend this research for analyzing the relation of language use and gender for literary characters. To the best of our knowledge, there has been little work on understanding how novelists and playwrights

portray (if they do) differential language use by literary characters of different genders. To apply automated analysis techniques, we need a clean separation of the speech of different characters in a literary work. In novels, such speech is integrated into the text and difficult to extract automatically. To carry out such research, we prefer source texts which give easy access to such structural information; hence, we focus on analyzing characters in plays. The natural choice for a starting point is the corpus of Shakespeare's plays.

We thus ask the following questions. Can the gender of Shakespeare's characters be determined from their word usage? If we are able to find such word use, can we glean any insight into how Shakespeare portrays maleness and femaleness? Are the differences (if any) between male and female language in Shakespeare's characters similar to those found in modern texts by male and female authors? Can we expect the same kind of analysis in understanding Shakespeare's characters' gender, to the ones we discussed above? Keep in mind that here we examine text written by one individual (Shakespeare) meant to express words of different individuals with differing genders, as opposed to texts actually by individuals of different genders.

To address these questions, we applied text classification methods using machine learning. High classification accuracy, if achieved, will show that Shakespeare used different language for his male and for his female characters. If this is the case, then examination of the most important discriminating features should give some insight into such differences and to relate them to previous work on male/female language. The general approach of our work is to achieve a reasonable accuracy using different lexical features of the characters' speeches as input to machine learning and then to study those features that are most important for discriminating character gender.

## 2. Corpus Construction

We constructed a corpus of characters' speeches from 34 of Shakespearean plays, starting with the texts from the Moby Shakespeare[1]. The reason behind choosing this edition is that it is readily available on the web and has a convenient hierarchical form of acts and scenes for every play, while we do not expect editorial influence to unduly affect our differential analysis. The files collected from this web resource were converted into text files from hypertext media and then we cleaned the text files by removing stage directions. The gender of each character was entered manually. A text file for each character in each play was constructed by concatenating all of that character's speeches in the play. We only considered characters with 200 or more words. From that collection, all female characters were chosen. Then we took the same number of male characters as female characters from a play, restricted to those not longer than the longest female character from that particular play. In this way, we balanced the corpus for gender, giving a total of 83 female characters and 83 male characters, with equal numbers of males and females from each play. This corpus is termed the 'First Corpus'. We also built a second corpus based on the reviewer's comments, in which we equalized the number of words in male and female characters by taking every female character with more than 200 words and an equal number of the longest male characters from each play. The longest male and female characters were then matched for length by keeping a prefix of the longer part (male or female) of the same length (in words) as the shorter part. This procedure ensured that the numbers of words per play for both genders are exactly the same. This corpus is termed the 'Second Corpus'. We also split each corpus (somewhat arbitrarily) into 'early' and 'late' characters. We used the term early to those plays which were written in 16th century and late to those in 17th century. This chronology in plays as captured from Wikipedia[1]. The numbers of characters from each play for 'First Corpus' and 'Second Corpus' are shown in Table 1.

## 3. Feature Extraction

We processed the text using the ATMan system, a text processing system in Java that we have developed[3]. The text is tokenized and the system produces a sequence of tokens, each corresponds to a word in the input text file. We use two sets of words as features. A stylistic feature set (FW) is a list of more-or-less content-independent words comprising mainly function words, numbers, prepositions, and some common contractions (e.g., "you'll", "he'll"). A content-based feature set comprises all words that occur more than ten times in a corpus, termed Bag of Words (BoW).

We calculate the frequencies of these FWs and BoWs and turn them into numeric values by computing their relative frequencies, computed as follows. We first count the number of times two different features occurring together; then we divide this number to the count of the

feature in reference. In this way we calculate the relative frequency for each feature and a collection forms a feature vector, which represents a document (i.e. a character's speech). The FW set has 645 features including contractions; the BoW set has 2129 features collected from the first corpus and 2002 BoW features collected from the second corpus. The numeric vectors collected for each document is used as an input for machine learning.

## 4. Text Classification

The classification learning phase of this task is carried out by Weka's (Frank & Witten 1999) implementation of Sequential Minimal Optimization (Platt 1998) (SMO) using a linear kernel and default parameters. The output of SMO is a model linearly weighting the various text features (FW or BoW). Testing was done via 10 fold cross validation. This provides an estimation of generalization accuracy by dividing the corpus into 10 different subsets. The learning is then run ten times, each time using a different subset as a test set and combining the other nine subsets for training. In this way we ensure that each character is tested on at least once with training that does not include it. Tables 3 and 4 present the results obtained by running various experiments. It is clear that BoW has performed better than the FW in both selection criteria, as expected, since it has more features on which to operate. This shows that both style and content differ between male and female characters. As expected, the FWs have proven the stylistic evidence and not the content, which are visible from the Table 4. BoW gives a high 74.09 on over all corpuses with the equalizing on number of words selection strategy. Interestingly, FW gives highest accuracy of 74.28 in Late plays with only 63 training samples. This indicates that there is a greater stylistic difference between the genders in late Shakespeare than in early Shakespeare.

## 5. Discussion

The feature analysis phase is carried out by taking the results obtained from Weka's implementation of SMO. SMO provides weights to the features corresponding to both class labels. After sorting the features based on their weights, we collected the top twenty features from both character genders. Tables 5-8 lists the top 20 features from male and female characters and is shown with their assigned weights given by the SMO, for FWs and BoWs respectively. Tables 9-12 list the same for the Second Corpus. These tables also show the 'Average frequency of 100 words', which finds the frequency of a particular feature divided by total gender characters, and then for easy readability this figure is scaled by 100 times. To discriminate binary class labels, SMO uses positive and negative weight values in Weka's implementation. We see from the Tables 5-10, male features are designated as negative weights and female characters are given as positive weights. In top 20 male features, this can be observed that 'Average Frequency of 100 Words' value of male is more than the corresponding value for female. This hold same in the case of the top 20 female features where female 'Average Frequency of 100 Words' value is more than the male for the same feature.

**Feature Analysis: BoW**

We can see cardinal number usage is found in male characters. Plural and mass nouns ('swords', 'dogs', 'water') are used more in males than females. On the other hand, there is strong evidence for singular noun ('woman', 'mother', 'heart') usage in females. The use of 'prithee' as an interjection is found in female character. This may represent a politeness aspect in their attitude. The past participle form is generally found in females ('gone', 'named', 'known'). Present tense verb forms ('pour', 'praise', 'pray', 'love', 'dispatch', 'despair') are used in female characters. In the case of male characters, Shakespeare used these verb forms ('avoid', 'fight', 'wrought'). Male characters seem to be aggressive while female characters seem to be projected as supporters of relationships.

**Feature Analysis: FW**

We observed that Shakespeare's female characters used more adverbs and adjectives, as well as auxiliary verbs and pronouns. On the other hand, cardinal numbers, determiners, and some prepositions are generally indicative of male characters. These observations are in line with previous work (Argamon et al. 2003) on discriminating author gender in modern texts, supporting the idea that the playwright projects characters' gender in a manner consistent with authorial gender projection. We did observe some contrasting results in the FW features from the second corpus. Number (i.e. twice) is found in female characters. Certain prepositions are used for females, while negation only appears distinctive for early females. Determiner 'the' which is a strong male character indicator in first corpus is found only in early part of second corpus. Some negation ('cannot') is found in late males as well. Clearly, more and deeper analysis is needed.

## 6. Conclusion

This is the first work, to our knowledge, in analyzing literary character's gender from plays. It seems clear that male and female language in Shakespeare's characters is similar to that found in modern texts by male and female authors (Argamon et.al 2003), but more work is needed in understanding character gender. We have also observed possible differences between early and late Shakespeare in gender character classification. In particular, the later Shakespeare plays appear to show a greater stylistic discrimination between male and female characters than the earlier plays. We are particularly interested in collaborating with literary scholars on this research to explore these issues further.

| Play Name | Gender Count |
|---|---|
| All's Well That Ends Well | 8 |
| Antony and Cleopatra | 4 |
| As You Like It | 6 |
| Cymbeline | 4 |
| King Lear | 4 |
| Loves Labours Lost | 8 |
| Measure for Measure | 4 |
| Midsummer Nights Dream | 6 |
| Much Ado About Nothing | 8 |
| Othello The Moore of Venice | 6 |
| Pericles Prince of Tyre | 8 |
| Romeo and Juliet | 6 |
| The Comedy of Errors | 8 |
| The First part of King Henry The Fourth | 2 |
| The First part of King Henry The Sixth | 4 |
| The Life and Death of Julies Caesar | 2 |
| The Life and Death of Richard The Second | 4 |
| The Life and Death of Richard The Third | 8 |
| The Life of King Henry The Eighth | 4 |
| The Life of King Henry The Fifth | 4 |
| The Merchant of Venice | 6 |
| The Merry Wives of Windsor | 6 |
| The Second part of King Henry The Fourth | 2 |
| The Second part of King Henry The Sixth | 2 |
| The Taming of the Shrew | 4 |
| The Tempest | 2 |
| The Third part of King Henry The Sixth | 6 |
| The Tragedy of Coriolanus | 4 |
| The Tragedy of Hamlet | 2 |
| Titus Andronicus | 4 |
| Troilus and Cressida | 2 |
| Twelfth Night | 6 |
| Two Gentlemen of Verona | 6 |
| Winter's Tale | 6 |

*Table 1: Corpus Composition*

| | Male | Female |
|---|---|---|
| **All** | 83 | 83 |
| **Early** | 48 | 48 |
| **Late** | 35 | 35 |

*Table 2: Overall Corpus Statistics*

| Feature Set | Accuracy |
|---|---|
| **All** | |
| Function Words | **66.26** |
| Bag-of-Words | **73.49** |
| **Early** | |
| Function Words | **63.54** |
| Bag-of-Words | **62.50** |
| **Late** | |
| Function Words | **62.85** |
| Bag-of-Words | **60.00** |

*Table 3: Accuracy is expressed in percentage for First Corpus Selection*

| Feature Set | Accuracy |
|---|---|
| **All** | |
| Function Words | 65.66 |
| Bag-of-Words | 74.09 |
| **Early** | |
| Function Words | 56.25 |
| Bag-of-Words | 58.33 |
| **Late** | |
| Function Words | 74.28 |
| Bag-of-Words | 64.28 |

*Table 4: Accuracy is expressed in percentage for Second Corpus Selection*

## Features from Various Experiments Using First Corpus

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| this | -0.7154 | 1195.18 | 845.78 | look | 0.7082 | 174.69 | 159.03 |
| follows | -0.658 | 12.04 | 3.61 | such | 0.6609 | 308.43 | 240.96 |
| allow | -0.6545 | 4.81 | 1.20 | thorou -ghly | 0.6248 | 3.61 | 0.0 |
| in | -0.6309 | 2045.78 | 1679.51 | comes | 0.6134 | 102.40 | 78.31 |
| well | -0.6051 | 213.25 | 172.28 | gone | 0.6095 | 31.32 | 19.27 |
| three | -0.5525 | 85.54 | 18.07 | he's | 0.5935 | 67.46 | 54.21 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| there | -0.5513 | 296.38 | 231.32 | never | 0.5841 | 228.91 | 163.85 |
| allows | -0.5279 | 1.20 | 1.20 | only | 0.5805 | 56.62 | 44.57 |
| the | -0.5184 | 5408.43 | 3906.02 | am | 0.5404 | 474.69 | 395.18 |
| toward | -0.5184 | 14.45 | 4.81 | he | 0.5353 | 1198.79 | 1103.61 |
| one | -0.5084 | 333.73 | 263.85 | there -fore | 0.5212 | 104.81 | 72.28 |
| immed-iate | -0.4997 | 6.02 | 0.0 | might | 0.4557 | 102.40 | 73.49 |
| here's | -0.4789 | 39.75 | 24.09 | you | 0.4375 | 2283.13 | 2116.86 |
| appear | -0.4752 | 34.93 | 12.04 | further-more | 0.4349 | 1.20 | 0.0 |
| himself | -0.4713 | 51.80 | 27.71 | outside | 0.4342 | 2.40 | 1.20 |
| we'll | -0.4573 | 43.37 | 37.34 | take | 0.4298 | 237.34 | 228.91 |
| another | -0.4517 | 66.26 | 40.96 | brief | 0.4242 | 8.43 | 4.81 |
| five | -0.4204 | 39.75 | 8.43 | you'll | 0.4201 | 46.98 | 31.32 |
| thus | -0.4201 | 103.61 | 77.10 | wish | 0.418 | 44.57 | 25.30 |
| thank | -0.4122 | 79.51 | 59.03 | consid-ering | 0.3975 | 2.40 | 0.0 |

*Table 5: Statistics of Top 20 FW Features from Gender Char*

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| three | -0.1774 | 85.54 | 18.07 | alas | 0.22 | 12.04 | 1.20 |
| lying | -0.1602 | 4.81 | 0.0 | gone | 0.1634 | 31.32 | 19.27 |
| friendship | -0.1583 | 7.22 | 1.20 | brow | 0.1581 | 9.63 | 4.81 |
| shortly | -0.1474 | 6.02 | 1.20 | love | 0.1525 | 343.37 | 209.63 |
| bare | -0.1449 | 12.04 | 2.40 | o | 0.1483 | 279.51 | 172.28 |
| wrought | -0.1438 | 10.84 | 1.20 | prithee | 0.1437 | 12.04 | 2.40 |
| avoid | -0.1369 | 9.63 | 1.20 | he | 0.1337 | 1198.79 | 1103.61 |
| this | -0.1333 | 1195.18 | 845.78 | pray | 0.1289 | 189.15 | 139.75 |
| answer | -0.1302 | 63.85 | 33.73 | dispatch | 0.1289 | 7.22 | 2.40 |
| very | -0.1255 | 222.89 | 131.32 | sick | 0.1275 | 24.09 | 3.61 |
| purse | -0.1255 | 13.25 | 2.40 | such | 0.1273 | 308.43 | 240.96 |
| served | -0.1217 | 8.43 | 3.61 | woman | 0.1268 | 54.21 | 24.09 |
| savage | -0.1211 | 8.43 | 3.61 | am | 0.1257 | 474.69 | 395.18 |
| thrice | -0.1182 | 12.04 | 7.22 | glass | 0.1253 | 7.22 | 3.61 |
| whom | -0.1167 | 87.95 | 37.34 | mother | 0.1193 | 46.98 | 16.86 |
| fresh | -0.1165 | 19.27 | 8.43 | warrant | 0.1164 | 45.78 | 24.09 |
| her | -0.1162 | 736.14 | 536.14 | colour | 0.1147 | 12.04 | 6.02 |
| fears | -0.114 | 8.43 | 1.20 | me | 0.1144 | 1046.98 | 1032.53 |
| dogs | -0.1126 | 7.22 | 1.20 | poor | 0.1141 | 160.24 | 97.59 |
| hundred | -0.1117 | 27.71 | 8.43 | woo | 0.1134 | 16.86 | 7.22 |

*Table 6: Statistics of Top 20 BoW Features from Gender Character*

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| three | -0.4626 | 85.54 | 18.07 | theref-ore | 0.5684 | 104.81 | 72.28 |
| in | -0.4327 | 2045.78 | 1679.51 | only | 0.4555 | 56.62 | 44.57 |
| this | -0.4065 | 1195.18 | 845.78 | just | 0.4092 | 22.89 | 18.07 |
| the | -0.3971 | 5408.43 | 3906.02 | thorou-ghly | 0.4057 | 3.61 | 0.0 |
| five | -0.3785 | 39.75 | 8.43 | brief | 0.4043 | 8.43 | 4.81 |
| there | -0.3772 | 296.38 | 231.32 | he's | 0.3834 | 67.46 | 54.21 |
| certain | -0.3691 | 36.14 | 7.22 | gone | 0.3657 | 31.32 | 19.27 |
| together | -0.3684 | 20.48 | 10.84 | below | 0.3251 | 6.02 | 2.40 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| allow | -0.3609 | 4.81 | 1.20 | wish | 0.3177 | 44.57 | 25.30 |
| on't | -0.359 | 3.61 | 1.20 | never | 0.3012 | 228.91 | 163.85 |
| here's | -0.3537 | 39.75 | 24.09 | outside | 0.2931 | 2.40 | 1.20 |
| we | -0.3247 | 595.18 | 381.92 | in't | 0.2931 | 6.02 | 4.81 |
| whence | -0.3195 | 21.68 | 4.81 | known | 0.2784 | 31.32 | 28.91 |
| appear | -0.3095 | 34.93 | 12.04 | still | 0.2719 | 84.81 | 63.85 |
| necess-ary | -0.2927 | 2.40 | 0.0 | value | 0.2683 | 6.02 | 4.81 |
| seem | -0.2716 | 26.50 | 25.30 | else | 0.2622 | 63.85 | 50.60 |
| beyond | -0.2634 | 10.84 | 4.81 | keep | 0.2586 | 97.59 | 75.90 |
| indeed | -0.2607 | 33.73 | 19.27 | help | 0.2575 | 40.96 | 39.75 |
| wonder | -0.25 | 15.66 | 12.04 | along | 0.2571 | 22.89 | 14.45 |
| upon't | -0.2497 | 1.20 | 0.0 | me | 0.2566 | 1046.98 | 1032.53 |

*Table 7: Statistics of Top 20 FW Features from Early Gender Character*

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| another | -0.298 | 66.26 | 40.96 | such | 0.4355 | 308.43 | 240.96 |
| follows | -0.2849 | 12.04 | 3.61 | gone | 0.3838 | 31.32 | 19.27 |
| of | -0.2801 | 3115.66 | 2381.92 | am | 0.3198 | 474.69 | 395.18 |
| immediate | -0.2704 | 6.02 | 0.0 | you | 0.2784 | 2283.13 | 2116.86 |
| toward | -0.27 | 14.45 | 4.81 | on's | 0.2774 | 4.81 | 2.40 |
| three | -0.2624 | 85.54 | 18.07 | you're | 0.2684 | 21.68 | 8.43 |
| cannot | -0.2459 | 157.83 | 127.71 | might | 0.2522 | 102.40 | 73.49 |
| doing | -0.2422 | 16.86 | 4.81 | hence | 0.2518 | 21.68 | 18.07 |
| consider | -0.2367 | 6.02 | 2.40 | apart | 0.2349 | 3.61 | 2.40 |
| where | -0.235 | 221.68 | 186.74 | among | 0.2305 | 31.32 | 24.09 |
| this | -0.2314 | 1195.18 | 845.78 | use | 0.2254 | 51.80 | 46.98 |
| example | -0.2249 | 7.22 | 1.20 | sure | 0.221 | 46.98 | 36.14 |
| very | -0.2241 | 222.89 | 131.32 | seem | 0.1876 | 31.32 | 26.50 |
| whom | -0.22 | 87.95 | 37.34 | he | 0.1848 | 1198.79 | 1103.61 |
| already | -0.2135 | 10.84 | 4.81 | comes | 0.1806 | 102.40 | 78.31 |
| every | -0.2124 | 110.84 | 96.38 | where't | 0.179 | 1.20 | 0.0 |
| own | -0.2074 | 143.37 | 109.63 | almost | 0.1729 | 36.14 | 24.09 |
| be | -0.205 | 1265.06 | 1261.44 | lately | 0.1713 | 4.81 | 3.61 |
| we | -0.2018 | 595.18 | 381.92 | near | 0.1706 | 32.53 | 16.86 |
| rather | -0.1974 | 87.95 | 65.06 | little | 0.1662 | 86.74 | 77.10 |

*Table 8: Statistics of Top 20 FW Features from Late Gender Character*

## Features from Various Experiments Using Second Corpus

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| in | -0.85 | 1262.65 | 1037.95 | he | 0.78 | 1623.49 | 1454.81 |
| three | -0.71 | 47.59 | 13.85 | whoe'er | 0.72 | 1.80 | 0.0 |
| itself | -0.70 | 29.51 | 13.85 | him | 0.64 | 512.65 | 360.24 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| toward | -0.68 | 15.66 | 7.83 | below | 0.58 | 4.81 | 2.40 |
| follows | -0.67 | 9.63 | 1.80 | wish | 0.57 | 34.93 | 26.50 |
| thus | -0.63 | 69.87 | 54.21 | did | 0.51 | 158.43 | 103.01 |
| allow | -0.61 | 8.43 | 3.01 | he's | 0.48 | 33.13 | 19.27 |
| will | -0.57 | 501.20 | 440.36 | thorough | 0.47 | 3.01 | 0.60 |
| necess-ary | -0.51 | 2.40 | 0.60 | thou'dst | 0.47 | 0.60 | 0.0 |
| being | -0.49 | 75.90 | 44.57 | every | 0.47 | 46.98 | 39.75 |
| beyond | -0.49 | 6.62 | 3.01 | whet'st | 0.47 | 0.60 | 0.0 |
| gives | -0.47 | 15.06 | 9.03 | outside | 0.45 | 1.20 | 0.0 |
| these | -0.47 | 109.63 | 87.95 | gone | 0.45 | 59.03 | 35.54 |
| another | -0.46 | 37.34 | 27.71 | else | 0.44 | 40.96 | 31.92 |
| may | -0.45 | 156.02 | 135.54 | me | 0.43 | 1157.83 | 1045.18 |
| whence | -0.45 | 9.63 | 2.40 | to's | 0.43 | 1.80 | 0.60 |
| whom | -0.45 | 39.15 | 21.68 | down | 0.43 | 51.20 | 39.15 |
| allows | -0.44 | 0.60 | 0.0 | help | 0.43 | 34.93 | 24.09 |
| of | -0.44 | 1418.67 | 1225.30 | twice | 0.42 | 8.43 | 3.61 |
| after | -0.43 | 42.16 | 25.90 | does | 0.42 | 39.15 | 25.30 |

*Table 9: Statistics of Top 20 FW Features from All Gender Character*

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| in | -0.18 | 1262.65 | 1037.95 | prithee | 0.25 | 27.10 | 5.42 |
| three | -0.16 | 47.59 | 13.85 | him | 0.23 | 512.65 | 360.24 |
| her | -0.15 | 611.44 | 481.32 | alas | 0.22 | 17.46 | 2.40 |
| will | -0.14 | 501.20 | 440.36 | he | 0.22 | 1623.49 | 1454.81 |
| answer | -0.14 | 47.59 | 28.31 | heart | 0.21 | 134.93 | 88.55 |
| seest | -0.14 | 7.83 | 1.20 | o | 0.19 | 2519.87 | 2391.56 |
| appetite | -0.14 | 5.42 | 1.20 | mother | 0.17 | 59.03 | 24.09 |
| being | -0.14 | 75.90 | 44.57 | fortune | 0.17 | 42.77 | 27.71 |
| prepare | -0.14 | 9.03 | 3.61 | maiden | 0.16 | 14.45 | 3.01 |
| to | -0.13 | 1985.54 | 1822.28 | warrant | 0.15 | 30.72 | 8.43 |
| thrive | -0.13 | 9.63 | 1.20 | master's | 0.15 | 9.63 | 1.20 |
| itself | -0.13 | 29.51 | 13.85 | dispatch | 0.15 | 9.63 | 3.61 |
| hopes | -0.13 | 7.83 | 3.01 | wish | 0.15 | 34.93 | 26.50 |
| months | -0.13 | 8.43 | 1.20 | easily | 0.14 | 4.21 | 1.80 |
| another | -0.13 | 37.34 | 27.71 | did | 0.14 | 158.43 | 103.01 |
| fellow | -0.13 | 35.54 | 20.48 | named | 0.13 | 6.02 | 1.20 |
| fellows | -0.13 | 7.22 | 2.40 | lord | 0.13 | 298.19 | 137.95 |
| steel | -0.12 | 7.22 | 1.80 | does | 0.12 | 39.15 | 25.30 |
| ink | -0.12 | 6.62 | 1.80 | pray | 0.12 | 125.30 | 72.28 |
| greatn-ess | -0.12 | 10.24 | 2.40 | same | 0.12 | 22.28 | 10.84 |

*Table 10: Statistics of Top 20 BoW Features from All Gender Character*

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| will | -0.58 | 501.20 | 440.36 | only | 0.51 | 28.31 | 24.69 |
| in | -0.58 | 1262.65 | 1037.95 | below | 0.47 | 4.81 | 2.40 |
| three | -0.51 | 47.59 | 13.85 | there-fore | 0.42 | 53.01 | 49.39 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| gives | -0.45 | 15.06 | 9.03 | help | 0.42 | 34.93 | 24.09 |
| after | -0.44 | 42.16 | 25.90 | need | 0.41 | 37.34 | 29.51 |
| but | -0.44 | 519.27 | 551.20 | whole | 0.40 | 12.04 | 7.22 |
| thus | -0.41 | 69.87 | 54.21 | he's | 0.38 | 33.13 | 19.27 |
| allows | -0.41 | 0.60 | 0.0 | gone | 0.36 | 59.03 | 35.54 |
| necess-ary | -0.41 | 2.40 | 0.60 | wish | 0.36 | 34.93 | 26.50 |
| beyond | -0.38 | 6.62 | 3.01 | he | 0.35 | 1623.49 | 1454.81 |
| greetings | -0.35 | 1.80 | 0.60 | old | 0.35 | 40.96 | 46.38 |
| here | -0.35 | 178.31 | 166.26 | never | 0.35 | 109.63 | 80.72 |
| got | -0.35 | 9.03 | 6.02 | keep | 0.35 | 55.42 | 49.39 |
| thence | -0.33 | 11.44 | 4.81 | oh | 0.32 | 0.60 | 0.0 |
| how | -0.32 | 200.60 | 179.51 | will't | 0.31 | 1.80 | 0.0 |
| her | -0.31 | 611.44 | 481.32 | for't | 0.31 | 8.43 | 6.02 |
| follows | -0.31 | 9.63 | 1.80 | every | 0.31 | 46.98 | 39.75 |
| accord-ing | -0.31 | 7.83 | 3.01 | away | 0.31 | 60.84 | 53.61 |
| certain | -0.31 | 16.86 | 13.25 | thou'dst | 0.29 | 0.60 | 0.0 |
| ta'en | -0.29 | 7.22 | 5.42 | him | 0.29 | 512.65 | 360.24 |

*Table 11: Statistics of Top 20 FW Features from Early Gender Character*

| Male Features | | | | Female Features | | | |
|---|---|---|---|---|---|---|---|
| Feature | SMO Weight | Avg. Freq. (x 100) | | Feature | SMO Weight | Avg. Freq. (x 100) | |
| | | Male | Female | | | Female | Male |
| itself | -0.33 | 29.51 | 13.85 | he | 0.40 | 1623.49 | 1454.81 |
| in | -0.30 | 1262.65 | 1037.95 | who's | 0.32 | 7.83 | 3.01 |
| toward | -0.25 | 15.66 | 7.83 | such | 0.30 | 150.0 | 115.66 |
| the | -0.23 | 3341.56 | 2884.93 | does | 0.28 | 39.15 | 25.30 |
| three | -0.22 | 47.59 | 13.85 | best | 0.26 | 50.0 | 37.95 |
| cannot | -0.21 | 75.90 | 62.65 | tell | 0.24 | 118.07 | 96.98 |
| and | -0.21 | 1917.46 | 1801.80 | might | 0.23 | 60.24 | 43.37 |
| of | -0.21 | 1418.67 | 1225.30 | him | 0.23 | 512.65 | 360.24 |
| follows | -0.20 | 9.63 | 1.80 | you | 0.23 | 2054.81 | 1733.73 |
| own | -0.20 | 76.50 | 60.24 | your | 0.22 | 665.66 | 585.54 |
| ones | -0.20 | 7.22 | 3.61 | last | 0.20 | 29.51 | 20.48 |
| followed | -0.2 | 1.80 | 1.20 | goes | 0.20 | 17.46 | 7.83 |
| may | -0.19 | 156.02 | 135.54 | among | 0.20 | 10.84 | 7.83 |
| without | -0.19 | 28.91 | 19.87 | little | 0.20 | 49.39 | 42.16 |
| someb-ody | -0.19 | 0.60 | 0.0 | on's | 0.19 | 2.40 | 0.60 |
| towards | -0.19 | 8.43 | 6.62 | lately | 0.19 | 2.40 | 1.20 |
| this | -0.18 | 530.72 | 447.59 | almost | 0.19 | 15.66 | 9.63 |
| beyond | -0.18 | 6.62 | 3.01 | twice | 0.17 | 8.43 | 3.61 |
| gives | -0.18 | 15.06 | 9.03 | yes | 0.17 | 7.83 | 3.01 |
| another | -0.17 | 37.34 | 27.71 | howe'er | 0.16 | 2.40 | 0.0 |

*Table 12: Statistics of Top 20 FW Features from Late Gender Character*

# References

**Koppel M., Argamon S., Shimoni A**. (2004). Automatically

Categorizing Written Texts by Author Gender : *Literary and Linguistic Computing* 17(4).

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

**Argamon S., Whitelaw C., Chase P., Hota S., Dhawle S., Garg N., Levitan S.** (2005) *Stylistic Text Classification using Functional Lexical Features, Journal of the Association for Information Sciences and Technology*, to appear.

**Corney M., Vel O., Anderson A., Mohay G.** (2002). *Gender Preferential Text Mining of E-mail Discourse :* In *Proceedings of 18th Annual Computer Security Applications Conference ACSAC*

**Corney M., Vel O., Anderson A.** (2001). Mining E-mail Content for Author Identification Forensics *: ACM SIGMOD Record* Volume 30, Issue 4 December

**Joachims, T.** (1998). Text Categorization with Support Vector Machines: Learning with many relevant features. *ECML-98, Tenth European Conference on Machine Learning*.

**Platt, J.** (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research Technical Report MSR-TR-98-14,

**Mitchell, T.** (1997) *Machine Learning*. (McGraw-Hill)

**Witten I., Frank E.** (1999). *Weka3: Data Mining Software in Java* http://www.cs.waikato.ac.nz/ml/weka/**Tables**

# Criticism Mining: Text Mining Experiments on Book, Movie and Music Reviews

**Xiao HU**
**J. Stephen DOWNIE**
**M. Cameron JONES**

*University of Illinois at Urbana-Champaign*

## 1. INTRODUCTION

There are many networked resources which now provide critical consumer-generated reviews of humanities materials, such as online stores, review websites, and various forums including both public and private blogs, mailing lists and wikis. Many of these reviews are quite detailed, covering not only the reviewers' personal opinions but also important background and contextual information about the works under discussion. Humanities scholars should be given the ability to easily gather up and then analytically examine these reviews to determine, for example, how users are impacted and influenced by humanities materials. Because the ever-growing volume of consumer-generated review text precludes simple manual selection, the time has come to develop robust automated techniques that assist humanities scholars in the location, organization and then the analysis of critical review content. To this end, the authors have conducted a series of very promising large-scale experiments that bring to bear powerful text mining techniques to the problem of "criticism analysis". In particular, our experimental results concerning the application of the Naïve Bayes text mining technique to the "criticism analysis" domain indicate that "criticism mining" is not only feasible but also worthy of further exploration and refinement. In short, our results suggest that the formal development of a "criticism mining" paradigm would provide humanities scholars with a sophisticated analytic toolkit that will open rewarding new avenues of investigation and insight.

## 2. EXPERIMENTAL SETUP

Our principal experimental goal was to build and then evaluate a prototype criticism mining system that could automatically predict the: