Automatically Categorizing Written Texts by Author Gender

Moshe Koppel[1] **Shlomo Argamon**[2,1] **Anat Rachel Shimoni**[1]

[1]*Dept. of Computer Science, Bar-Ilan University*
Ramat Gan 52900, Israel

[2]*Dept. of Computer Science, Jerusalem College of Technology*
21 Havaad Haleumi St. Jerusalem 91102, Israel

Abstract
The problem of automatically determining the gender of a document's author would appear to be a more subtle problem than those of categorization by topic or authorship attribution. Nevertheless, it is shown that automated text categorization techniques can exploit combinations of simple lexical and syntactic features to infer the gender of the author of an unseen formal written document with approximately 80% accuracy. The same techniques can be used to determine if a document is fiction or non-fiction with approximately 98% accuracy.

1. Introduction

1.1 Text Categorization

The last ten years has seen an explosion of research in automated text categorization (Sebastiani 2002). In the text categorization problem, we are given a set of two or more categories and examples of texts in each category and are asked to correctly categorize unseen texts. The components of a text categorization system are now well understood:

1. Document Representation – Choose a large set of text features which might potentially be useful for categorizing a given text (typically words that are neither too common nor too rare) and represent each text as a vector in which entries represent (some non-decreasing function of) the frequency of each feature in the text.
2. Dimension Reduction – Optionally, use various criteria for reducing the dimension of the vectors – typically by eliminating features which don't seem to be correlated with any category (Yang & Pedersen 1997), by using latent semantic indexing (Deerwester et al 1990), or by stepwise iteration of a learning algorithm (see below).
3. Learning Method – Use some machine learning method to construct one or more models of each category. Yang (1999) compares and assesses some of the most promising algorithms, which include k-nearest-neighbor, neural nets, Winnow, SVM, etc. (If multiple models are learned, methods such as bagging and boosting (Bauer & Kohavi 1999) can be used to combine the models.)
4. Testing Protocol – Finally, use some testing protocol, such as bootstrapping or k-fold cross-validation to estimate the reliability of the system.

1.2 Stylometry

Driven by the problem of Internet search, the text categorization literature – outside of the stylometric research community – has, with a few exceptions (Argamon-Engelson et al 1998, Wolters & Kirsten 1999), concerned categorization by topic rather than categorization by writing style. The problem considered in this paper concerns categorization by style and thus is more similar to the stylometric work which has been vigorously pursued for decades, mostly in the context of authorship attribution (Holmes 1998, McEnery & Oakes 2000). Although some important crossover work between

stylometrics and text categorization has been done (Forsyth 1999), the bulk of stylometric research has differed from the more recent work in text categorization in a few important ways. While categorization by topic is typically based on keywords which reflect a document's content, categorization by author style uses precisely those features which are independent of content. Thus, stylometric models for categorization have typically been based on hand-selected sets of content-independent, lexical (Mosteller & Wallace 1964), syntactic (Baayen et al 1996, Stamatatos et al 2001), or complexity-based (Yule 1938) features. Researchers in text categorization by topic typically use much larger feature sets, often in conjunction with automated feature selection methods (Yang & Pedersen 1997); some work in the stylometric community has also considered automated methods for selecting features (Forsyth & Holmes 1996). Moreover, stylometric research has tended to use statistical methods such as multivariate analysis (Burrows 1992, Holmes & Forsyth 1995), rather than machine-learning algorithms, for categorization, although a number of researchers have applied machine learning methods to stylometric problems (Matthews & Merriam 1993, 1997, Merriam & Matthews 1994, Forsyth 1999).

1.3 Gender

The object of this paper is to explore the possibility of automatically classifying formal written texts according to author gender. This problem differs from the typical text categorization problem which focuses on categorization according to topic. It also differs from the typical stylometric problem which focuses on authorship attribution – individual authors are more likely to exhibit consistent habits of style than large classes of authors. We will see, though, that using ideas from both the stylometric community and the text categorization community, we are able to achieve surprising results.

While a substantial literature has been devoted to isolating distinguishing characteristics of male/female linguistic styles (Lakoff 1975, Holmes 1993), this paper goes beyond earlier work in two ways:

First, most of the previous work considered spoken language (Key 1972, Trudgill 1972, Labov 1990, Eckert 1997), which, unlike the formal written texts (i.e. books and articles) we consider, includes intonational, phonological and conversational cues. The relatively few studies on gender differences in writing have focused on more informal contexts – such as student essays (Mulac et al 1990, Mulac & Lundell 1994), electronic communications (Herring 1996) and correspondence (Biber et al 1998, Palander-Collin 1999) – and some authors (Berryman-Fink & Wilcox 1983, Simkins-Bullock and Wildman 1991) have asserted that no difference between male and female writing styles in more formal contexts should be expected.

Second, there has been scant evidence thus far that differences between male and female writing are pronounced enough that they could be parlayed into an algorithm for categorizing an unseen text as being authored by a male or by a female. In this paper, we employ machine learning algorithms on a genre-controlled corpus of 566 documents taken from the British National Corpus (BNC) to construct models for performing just such a task. We show that these models classify unseen texts according to author gender with accuracy of approximately 80%.

2. The Corpus

The BNC includes 920 documents in British English that are labeled both for author gender and for genre: fiction and several non-fiction genres and sub-genres as will be shown below. All the experiments reported in this paper were performed on a genre-controlled subset of the BNC constructed as follows: in each sub-genre, we use all the documents in the smaller (male or female) class and randomly select an equal number of documents from the other class, discarding the excess documents.

The resulting corpus contains 566 documents (a full listing of which can be found at http://shekel.jct.ac.il/~argamon/gender-style).

No single author wrote more than three documents in this corpus. All of the non-fiction documents and 75% of the fiction documents are from the years 1975-1993; the remaining fiction documents are from the years 1960-1974. The documents contain between 554 and 61,199 words with an average of about 34,320 words each (female=34,795; male=33,845).

3. Document Representation

Unlike some earlier studies on authorship attribution (see Holmes 1998, McEnery & Oakes 2000), we do not begin with a small hand-selected set of features deemed most likely to distinguish between categories. Rather, we begin with a very large set of lexical and quasi-syntactic features that were chosen solely on the basis of their being more-or-less topic-independent. The features include a list of 405 function words (which appear at least once in the corpus) and a list of n-grams of parts-of-speech using the BNC's tag set of 76 parts of speech (such as PRP=*preposition*, NN1=*singular noun*, and AT0=*article*) and punctuation marks. (A full listing of all these features can be found at http://shekel.jct.ac.il/~argamon/gender-style). We use the 500 most common ordered triples, 100 most common ordered pairs and all the single tags as features. For example, a common triple is PRP_AT0_NN1 as in the phrase "…*above the table*…". The use of parts-of-speech n-grams is a relatively efficient way to capture the heavier syntactic information shown in (Baayen et al 1996, Stamatatos et al 2001) to be useful for distinguishing writing styles.

Each document is thus represented as a vector of length 1081 (the total number of features), in which each entry represents the number of appearances of the feature in the document divided by document length. In order that different feature types all have values in roughly the same range, the values associated with function words and POS doubles were multiplied by 2 and those associated with POS triples were multiplied by 4.

We will use automated methods to significantly reduce the number of features actually used for classification. However, these methods make use of iterated runs of our learning algorithm, so we turn first to the details of this learning algorithm.

4. The Learning Method

Our objective is to use a set of training documents to find a linear separator between male-authored and female-authored documents. That is, we seek a weight vector $\mathbf{w}$ such that for each training document, $\mathbf{x}$, the vector dot-product $\mathbf{w}*\mathbf{x}$ exceeds a threshold T if and only if $\mathbf{x}$ was authored by a female. Our method for finding the weight vector $\mathbf{w}$ is a variant of the Exponential Gradient (EG) algorithm of (Kivinen & Warmuth 1997) which itself is a generalization of the Balanced Winnow algorithm of (Littlestone 1987). These algorithms have nice theoretical mistake-bound properties and have previously been shown to be effective for text-categorization by topic (Lewis et al 1996, Dagan et al 1997).

Briefly, our method works as follows: We initially define two component weight vectors $\mathbf{w}^+ = \{1,1,\ldots,1\}$ and $\mathbf{w}^- = \{-1,-1,\ldots,-1\}$, defining $\mathbf{w} = \mathbf{w}^+ + \mathbf{w}^-$. We then calibrate the vectors $\mathbf{w}^+$ and $\mathbf{w}^-$ using the following iterative procedure. The training examples are randomly ordered. For each training example $\mathbf{x}$, we define $c(\mathbf{x}) = 1$ if $\mathbf{x}$ is female-authored and $c(\mathbf{x}) = 0$ otherwise. Let $s(\mathbf{w},\mathbf{x}) = 1$ if $\mathbf{w}*\mathbf{x} > 0$ and $s(\mathbf{w},\mathbf{x}) = 0$ otherwise, where $\mathbf{w}$ is the weight vector at the time that example $\mathbf{x}$ is encountered, and let $\mathbf{w}_i$ and $\mathbf{x}_i$ be the i^{th} element in $\mathbf{w}$ and $\mathbf{x}$, respectively. Then we take the examples one at a time and iteratively update the weights after each example using the formulas

$$w_i^+ \leftarrow w_i^+ (1+\beta x_i)^{(c(x)-s(w,x))}$$

$$w_i^- \leftarrow w_i^- (1+\beta x_i)^{(s(w,x)-c(x))}$$

β is a learning constant greater than 0; in all our experiments we used $\beta = 3$. Thus weights that improperly reduce the dot product are increased, and vice versa. Note that as in EG, but unlike Balanced Winnow, we allow x_i to take on non-binary values. However, like Balanced Winnow, but unlike EG, we restrict $s(\mathbf{w},\mathbf{x})$ to binary values.

Once all the examples have been used for training, they are randomly reordered and another cycle of updates is run. This continues until all training examples are correctly classified or until 100 consecutive cycles fail to produce improvement in the number of training examples correctly classified. Along the way, any element of $\mathbf{w}^+$ or $\mathbf{w}^-$ that drops below some threshold (0.000001 of the sum of all the weights) is set to zero.

The intuition behind the update rule is that weights of features that appear most prominently in misclassified documents are changed most dramatically. A well-known advantage of multiplicative update rules such as we use is that the weights of irrelevant features tend to zero. As a result of this property, the learning algorithm itself can be used for feature reduction: by iteratively running the learner, eliminating low-weighted features and rerunning, we can produce models with fewer and fewer features. Experimental results using this method will be discussed below.

5. Results

Table 1 (top row) shows accuracies obtained from 10 separate runs 56-fold cross-validation (that is, ten examples in each fold) using feature sets consisting of function words only (FW), parts-of-speech only (POS) and both function words and parts-of-speech (FWPOS). For function words, 73.7% (±0.86% stderr) of the documents are correctly classified, for parts-of-speech, 70.5% (±0.90%), and for the full feature set, 77.3% (±0.79%). Clearly, the combined feature set is the best choice, despite the fact that using more features than documents (that is, more free parameters than constraints) might easily have led to over-fitting to the training set at the expense of testing accuracy.

One of the difficulties in obtaining greater accuracy overall is the difference between fiction and non-fiction. These differences are generally greater than the difference between male and female writing styles and thus training on fiction and non-fiction documents together actually harms results. When we train together, accuracy on fiction test documents is 74.5% and on non-fiction is 79.7%. When we train only on fiction documents (thus using a substantially smaller training set), results of 36-fold cross-validation (maintaining ten examples per fold) actually increase to 79.5%. Likewise, when training on non-fiction only, accuracy on non-fiction test documents increased to 82.6%. This is because, as we shall see in detail below, the frequencies of the critical distinguishing features are different in fiction

and non-fiction. In fact, when training is performed on fiction (non-fiction) only, non-fiction (fiction) documents are categorized with barely more than 50% accuracy – no better than random!

Table 1 (rows 2 and 3) shows the results of training and testing separately on fiction and non-fiction using different feature sets. What is most remarkable is the extent to which results for non-fiction improve when both feature sets are used in tandem. The reasons for this will be discussed below. The full breakdown of results for different sub-genres – respectively, training (using all features) on all documents together and training (likewise using all features) separately on fiction and non-fiction – is shown in Table 2.

Table 1: Accuracies with standard errors for 10 cross-validation runs each (see text for details) for different train/test domains and different feature sets.

Domain	**FW**	**POS**	**FWPOS**
All	73.7 ± 0.86	70.5 ± 0.90	77.3 ± 0.79
Fiction	78.8 ± 1.1	77.1 ± 0.85	79.5 ± 1.1
Nonfiction	68.5 ± 1.3	67.2 ± 1.2	82.6 ± 0.99

Table 2 Accuracy averaged over 10 cross-validation runs, broken down by genre of the test documents, with training either on all documents or on fiction or non-fiction documents only. The number of documents indicated on each line of column 2 reflects an equal number of male and female documents. See text for further details of the experimental methodology.

Testing on Genre:	**# of docs**	**Train on All**	**Train on Fiction**
Fiction	264	74.5	79.5
Fiction / Female	132	74.8	81.7
Fiction / Male	132	74.2	77.3
			Train on Non-fiction
Non-fiction	302	79.7	82.6
Non-fiction / Female	151	79.2	83.3
Non-fiction / Male	151	80.2	81.9
Arts (Non-academic)	16	76.0	76.3
Arts (Academic)	24	75.6	77.5
Belief & Thought	24	85.0	85.0
Biography	54	87.0	90.0
Commerce	10	60.0	84.0
Leisure	16	85.7	81.3
Science	26	74.2	78.5
Social Science (Non-academic)	52	77.5	83.0
Social Science (Academic)	38	82.9	78.4
World Affairs	42	79.2	82.9

Winnow is able to overcome differences between different genres by exploiting subtle dependencies between features. Less subtle learning methods are unable to deal successfully with this problem. Thus, Naïve Bayes, which ignores feature dependencies correctly classifies fiction documents with only 64.4% accuracy and non-fiction documents with 60.9% accuracy. Ripper, which is a decision-tree learner that greedily selects feature thresholds for optimal category seperation, correctly classifies fiction documents with 64.9% accuracy and non-fiction documents with 73.7% accuracy. (Results are for the full feature set; results for FW and POS vary only slightly.)

6. Feature Reduction

How many features actually contribute to categorization and which ones contribute most? In order to test this we ran the following experiment: For each model obtained in a cross-validation trial, we chose the 128 most important features (where the importance of a feature in a given model is defined as the absolute value of its weight in the model multiplied by its average frequency in the training set) in each direction and ran the cross-validation trial again using only those 256 features. We then repeated the process using only the 128 most important features (64 from each side) in the obtained model. This process was iterated down to just 8 features from each side. The results are shown in Figure 1. Optimal performance seems to lie between 64 and 128 features per side. For the full feature set, non-fiction reaches accuracy of 83.7% at 128 features per side and fiction reaches 79.9% at 64 features per side. Most significantly, though, the drop-off in performance is extremely slow. Even using as few as eight features (from the full feature set) on a side, we obtain accuracy of 79.1% for non-fiction and 76.8% for fiction.

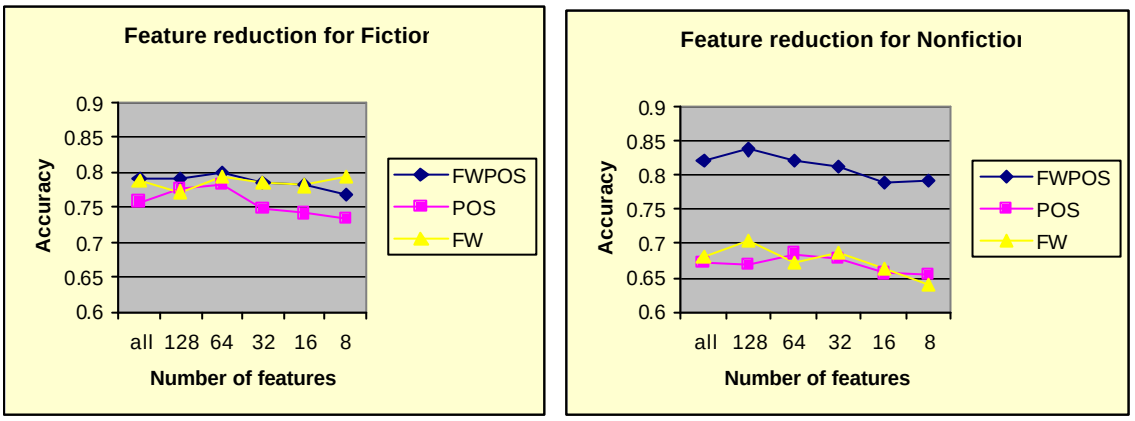

Figure 1: Averaged cross-validation accuracies for training/testing on fiction and on non-fiction, plotted over different numbers of features, ranging from the full feature sets of 1081 (FWPOS), 676 (POS), and 405 (FW) features, down to 8 features in each case. See text for detail on experimental methodology.

In fact, when training and testing on fiction using only function words, optimal performance (79.4%) is achieved when using only eight words from each side. Not surprisingly, certain features tended to survive down to the final iteration in different trials. For example, the function words which consistently appear in the final iteration training on fiction are: male features – *a, the, as*; female features – *she, for, with, not*. When training on non-fiction we find: male features – *that, one*; female features – *for, with, not, and, in*. Similar phenomena appear when using parts-of-speech. Elsewhere (Argamon et al, submitted) we have analyzed differences in usage between males and females for certain classes of words. The picture that emerges is that the male indicators are largely noun specifiers (determiners, numbers, modifiers) while the female indicators are mostly negation, pronouns and certain prepositions. Although a given feature's Winnow weight does not necessarily reflect the feature's mean frequency difference between males and females, a comparison of male and female usage of determiners, pronouns, prepositions, negation, and the conjunction *and* (Table 3) reveals significant differences in usage between males and females both in fiction and in non-fiction. These results bear out and significantly extend results of earlier research on gender differences (Biber et al 1998, Holmes 1993).

Another interesting fact that is explained by analysis of the data is that for non-fiction, using function words and parts-of-speech together yields considerably greater accuracy than either feature type by itself. This is because the combination of features is able to exploit anomalies such as that women use the prepositions *for* and *with* significantly greater frequency than do men, but men use the set of all other prepositions (PRP) with about the same frequency as do women and the preposition *of* (which in the BNC gets its own tag, PRF) with greater frequency than women. Similarly, men use the pronoun *he* with about the same frequency as do women but women use the set of all other pronouns much more than men do. In particular, what distinguishes non-fiction is that – as in fiction – men use significantly more determiners (AT0, DT0) than do women, but – unlike fiction – women use the most frequent determiner, *the*, with about the same frequency as men.

Table 3. Frequency means (per 10,000 words) and standard errors for a variety of features in male/female fiction/non-fiction documents. Parts of speech are indicated using the BNC tag set (PNP=pronouns; AT0=determiners *a*, *an*, *the*, *no*; DT0=other determiners; XX0=*not*, **n't*; PRF=preposition *of*; PRP=other prepositions)

	Fiction		**Non-fiction**	
Feature	**Male μ ± stderr**	**Female μ ± stderr**	**Male μ ± stderr**	**Female μ ± stderr**
PNP	732 ± 14	809 ± 15	291 ± 12	331 ± 17
he	145 ± 4.7	135 ± 4.7	47.5 ± 3.5	48.1 ± 4.3
she	67 ± 4.3	139 ± 6.9	8.73 ± 1.7	21.5 ± 2.3
AT0	735 ± 9.5	626 ± 8.7	884 ± 9.1	822 ± 12
DT0	160 ± 2.9	153 ± 2.0	220 ± 4.0	204 ± 4.6
the	520 ± 8.6	418 ± 7.5	611 ± 8.4	614 ± 12
XX0	84 ± 2.4	98 ± 2.2	54 ± 1.5	55 ± 2.3
PRP	623 ± 6.0	615 ± 5.7	767 ± 5.9	763 ± 7.0
PRF	170 ± 4.2	158 ± 3.7	355 ± 7.2	324 ± 7.9
for	55.7 ± 1.1	61.3 ± 1.0	77.9 ± 1.6	90.7 ± 1.4
with	58.6 ± 1.1	66.5 ± 1.0	56.9 ± 1.1	67.8 ± 1.4
and	234 ± 4.9	249 ± 5.5	242 ± 3.9	287 ± 5.2

The extent to which frequencies of a small number of features can be parlayed into effective categorization is illustrated by the following fact: of the 58 documents in which *the* appears with frequency < 408 and *herself* appears with frequency > 5, all but two are by females.

7. Categorization by Genre

An interesting phenomenon that is evident in Table 3 is that the differences between male and female usages of various features parallel more extreme differences between fiction and non-fiction: determiners, which are used more by men, are used more by all authors in non-fiction; pronouns and negation, which are used more by women, are used more by all authors in fiction. The extreme differences between fiction and non-fiction suggest that distinguishing between the two genres ought to be an easier task than distinguishing between male and female authors. And indeed it is. Using the same corpus and same learning methodology as above on the fiction/non-fiction problem, ten runs of 56-fold cross-validation yields accuracy of 98%. Table 4 shows results for each of the three feature sets.

Feature Set	**Accuracy**
FWPOS	98.2 ± 0.003
POS	97.5 ± 0.003
FW	97.9 ± 0.003

Table 4: Accuracies and standard errors for 10 run of 56-fold cross validation for the fiction/non-fiction problem for all three feature sets.

The misclassified documents are the following:

Fiction:

- *Possession*, A. S. Byatt,
- *The Remains of the Day*, Kazuo Ishiguro
- *Now We Are Thirty-Somethings*, Charles Jennings
- *Now Then Davos,* Martin Wiley, David Harmer, and Ian McMillan
- *The Siege of Krishnapur,* J. G. Farrell
- *A Landing on the Sun,* Michael Frayn

Non-fiction:

- *Thank you for having me,* Maureen Lipman
- *A Crowd is not Company*, Robert Kee
- *T. S. Eliot: A Friendship,* Frederick Tomlin

- *Walking on Water*, Andy Martin
- *Unpublished letters and manuscripts*, Unlisted female author
- *Falling for Love: Teenage Mothers Talk*, Sue Sharpe

Examination of the six misclassified non-fiction documents reveals that all are biographical or diary-like works. Significantly, of the six misclassified fiction documents, five are by male authors (the sole exception is "Possession" by A. S. Byatt).

8. Conclusions

This paper has presented convincing evidence of a difference in male and female writing styles in modern English books and articles. Such a difference is sufficiently pronounced that it can be exploited for automated text classification with accuracy of approximately 80% (and higher in some cases). We have shown some of the features selected as being useful for classification, and have seen for some of them how their frequency distributions in the BNC differ for male and female authors.

The overall approach is essentially that of recent research in text categorization. The main difference between this work and text categorization by topic is in the choice of features: we use the kinds of content-independent features used by researchers in authorship attribution. Best performance is achieved when both function words and parts-of-speech n-grams are used in tandem. We have seen that while reasonable categorization is possible even using a relatively small number of such features, these features should be judiciously selected from a large initial set using principled feature reduction techniques. We have also found that the Winnow-like algorithm we used is superior for this type of problem to less subtle techniques such as decision trees and naïve Bayes.

The methods we have used in this work should work well for other style-based categorization problems. We have already seen that these methods distinguish fiction from non-fiction with about 98% accuracy. Other problems which might be tackled using this approach could include text categorization according to various demographic groupings of authors or intended audience, sub-genres, chronology, publication forum and so forth.

References

Argamon, S., M. Koppel, J. Fine, A. R. Shimony (submitted). Gender, genre, and writing style in formal written texts, submitted for publication.

Argamon-Engelson, S., M. Koppel, G. Avneri (1998). Style-based text categorization: What newspaper am I reading?, in Proc. of AAAI Workshop on Learning for Text Categorization, 1998, pp. 1-4

Baayen, H., H. van Halteren, F. Tweedie (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing, 11, 1996.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36:105-139, 1999

Berryman-Fink, C. L., T. R. Wilcox (1983). A multivariate investigation of perceptual attributions concerning gender appropriateness in language, Sex Roles 9, 1983.

Biber, D., S. Conrad, R. Reppen (1998). Corpus Linguistics Investigating Language Structure and Use, (Cambridge University Press, Cambridge, 1998).

Burrows, J. F. (1992). "Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information". Literary and Linguistic Computing, 7, 1992, 91-109

Dagan, I., Y. Karov, D. Roth (1997), Mistake-driven learning in text categorization in EMNLP-97: 2nd Conf. on Empirical Methods in Natural Language Processing 1997, pp. 55-63.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. Journal of the Society for Information Science, pages 391-407.

Eckert, P. (1997). Gender and sociolinguistic variation, in J. Coates ed., Readings in Language and Gender (Blackwell, Oxford 1997), pp. 64-75.

Forsyth, R. S. and Holmes, D. I. (1996). Feature finding for text classification.
Literary and Linguistic computing, 11(4):163-174.

Herring, S. (1996). Two variants of an electronic message schema, in S. Herring ed., Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives (John Benjamins, Amsterdam, 1996), pp. 81-106.

Holmes, D. I. and Forsyth, R. S. (1995). The federalist revisited: New directions in authorship attribution. Literary and Linguistic computing, 10(2):111-126

Holmes, D. (1998). The evolution of stylometry in humanities scholarship, Literary and Linguistic Computing, 13, 3, 1998, pp. 111-117.

Holmes, J. (1990). Hedges and boosters in women's and men's speech, Language & Communication 10, 3, 1990.

Holmes, J. (1993). Women's talk: The question of sociolinguistic universals, Australian Journal of Communications 20, 3, 1993.

Key, M. R. (1972). Linguistic behavior of male and female, Linguistics 88, 1972.

Kivinen, J., M. Warmuth, (1997). Additive versus exponentiated gradient updates for linear prediction, Information and Computation, 132, 1, 1997, pp 1-64.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change, Language Variation and Change 2, 1990.

Lakoff, R. T. (1975). Language and Women's Place (Harper Colophon Books, New York, 1975).

Lewis, D., R. Schapire, J. Callan, R. Papka (1996). Training algorithms for text classifiers, in Proc. 19th ACM/SIGIR Conf. on R&D in IR, 1996, pp 306-298.

Littlestone, N. (1987). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, Machine Learning, 2, 4, 1987, pp. 285-318.

Matthews, R. A. J. and Merriam, T. V. N. (1997). Distinguishing literary styles using neural networks. In Fiesler, E. and Beale, R., editors, Handbook of Neural Computation, chapter 8. IOP publishing and Oxford University Press.

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry : An application to the works of shakespeare and fletcher. Literary and Linguistic computing, 8(4):203-209.

McEnery, A., M. Oakes (2000). Authorship studies/textual statistics, in R. Dale, H. Moisl, H. Somers eds., Handbook of Natural Language Processing (Marcel Dekker, 2000).

Merriam, T. and Matthews, R. (1994). Neural computation in stylometry : An application to the works of shakespeare and marlowe. Literary and Linguistic computing, 9(1):1-6.

Mosteller, F. and Wallace, D. L. (1964). Inference and Disputed Authorship: The Federalist. Reading, Mass. : Addison Wesley.

Mulac, A., L. B. Studley, S. Blau (1990). The gender-linked language effect in primary and secondary students' impromptu essays, Sex Roles 23, 9/10, 1990.

Mulac, A., T. L. Lundell (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects, Language & Communication 14, 3, 1994.

Palander-Collin, M. (1999). Male and female styles in 17th century correspondence, Language Variation and Change 11, pp. 123-141.

Schutze, H., D. A. Hull, J. O. Pedersen (1995). A Comparison of Classifiers and Document Representations for the Routing Problem, in Proc. of 18th ACM/SIGIR Conf. on R&D in IR, 1995, pp. 229-237.

Sebastiani, F. (2002). Machine learning in automated text categorization, ACM Computing Surveys, forthcoming

Simkins-Bullock, J. A., B. G. Wildman (1991). An investigation into the relationship between gender and language, Sex Roles 24, 1991.

Stamatatos, E., N. Fakotakis & G. Kokkinakis, (2001). Computer-based authorship attribution without lexical measures, Computers and the Humanities 35, pp. 193—214.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich, Language in Society 1, 1972.

Wolters, Maria and Kirsten, Mathias (1999): Exploring the Use of Linguistic Features in Domain and Genre Classification. in: Proceedings of the Meeting of the European Chapter of the Association for Computational Linguistics

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, Vol 1, No. 1/2, pp 67–88, 1999.

Yang, Y. and Pedersen, J.O. (1997). A comparative study on feature selection in text categorization, Proceedings of ICML-97, 14th International Conference on Machine Learning,412--420

Yule, G.U. (1938). "On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship", Biometrika, 30, 363-390, 1938.