A comparative study of machine learning methods for authorship attribution

Matthew L. Jockers
Department of English, Stanford University, Stanford, CA 94305, USA

Daniela M. Witten
Department of Statistics, Stanford University, Stanford, CA 94305, USA

Correspondence:
Matthew L. Jockers,
Department of English,
Stanford University,
Stanford, CA 94305, USA.
E-mail:
mjockers@stanford.edu

Abstract

We compare and benchmark the performance of five classification methods, four of which are taken from the machine learning literature, in a classic authorship attribution problem involving the Federalist Papers. Cross-validation results are reported for each method, and each method is further employed in classifying the disputed papers and the few papers that are generally understood to be coauthored. These tests are performed using two separate feature sets: a ''raw'' feature set containing all words and word bigrams that are common to all of the authors, and a second ''pre-processed'' feature set derived by reducing the raw feature set to include only words meeting a minimum relative frequency threshold. Each of the methods tested performed well, but nearest shrunken centroids and regularized discriminant analysis had the best overall performances with 0/70 cross-validation errors.

1 Introduction

In statistical or quantitative authorship attribution, a work of unknown or disputed authorship is classified to a known author based on a training set of works of known authorship. Unlike typical document classification, however, in authorship attribution one does not desire to classify documents based on document content. Instead, one wishes to perform classification based upon author signal, or ''style.'' While there are several case specific issues (such as register, genre, subject, time period, sample length, etc.) that must be considered when testing for authorship, the two most important factors from a machine learning perspective involve the choice of features and the selection of an appropriate and effective classification technique. With regard to the choice of features, there is a growing consensus that analysis of high frequency words (mostly function, or closed class, words) and/or n-grams provides the most consistently reliable results in authorship attribution problems (Burrows 2002; Diederich *et al.*, 2000; Grieve 2007; Hoover 2003a,b; Koppel *et al.*, 2007; Martindale and McKenzie 1995; Uzuner and Katz 2005; Zhao and Zobel 2005; Yu 2008). There is, however, no similar consensus in terms of a best classifier. This lack of consensus may be attributed to the fact that, with a very few notable exceptions, the available methodologies have not been properly or fully compared.

Literary and Linguistic Computing, Vol. 25, No. 2, 2010.
doi:10.1093/llc/fqq001 Advance Access published on 12 April 2010

Much of the existing attribution research has focused upon testing the efficacy of single methods by conducting experiments within a corpus of certain, known provenance (Argamon 2008; Burrows 2002, 2007; Hoover 2001, 2004a,b; Tweedie *et al.*, 1996). Few studies in authorship attribution have been designed to test the relative merits of one method against another. Some initial benchmarking research includes Yu (2008), Jockers *et al.* (2008), and a more extensive analysis by Zhao and Zobel (2005). In Yu (2008) the goal was to compare methods in a more general text classification problem and not specifically in an authorship attribution setting. Yu compares naïve Bayes and support vector machines in classifying ''kinds of emotion'' such as eroticism in Emily Dickinson and sentimentalism in early American novels. In Jockers *et al.* (2008), two methods (Delta and nearest shrunken centroids) are applied to an authorship attribution problem and the cross-validation results of the two methods are discussed; however, the study was not expressly designed for comparative benchmarking. In Zhao and Zobel (2005), an experiment is constructed for the purpose of benchmarking several machine learning methods. The Zhao and Zobel study presents a limited, though carefully controlled, experiment designed for comparing five machine-learning methods. Using a feature set of ''365 function words,'' the authors evaluate classification techniques that have been previously employed in authorship attribution problems. Of the five methods tested, the researchers conclude that Bayesian networks are generally the most effective and that decision trees perform poorly by comparison. The authors also note that when limited positive training data are available, nearest neighbor methods perform well.

The test corpus utilized by Zhao and Zobel consists of 200,000+ Associated Press newswire articles written by 2,380 authors. These articles are relatively short. The corpus contains many articles by single individual authors, with as many as 800 documents by a single writer in one extreme case. Zhao and Zobel also adjust the size of the document pool in order to assess the effect of sample size on classification performance. The corpus employed here, while useful, is not typical of many authorship problems. Despite a carefully constructed and compelling experiment, one wonders how applicable the approaches recommended would be in more classic authorship attribution problems involving longer, often more ''literary'' texts such as Shakespeare's plays, The Federalist Papers of Madison, Hamilton and Jay, or even the anonymous political exposé *Primary Colors*. Moreover, the techniques Zhao and Zobel test do not include some of the most recent methodologies from the machine learning literature. So while Zhao and Zobels's work marks a significant watershed, it must be likewise seen as a beginning point rather than a definitive or exhaustive analysis.

Though not a benchmarking study per se, it is worth noting another experiment from 2004 in which Juola *et al.* (2006)[1] orchestrated a comparative testing of authorship attribution methods in the form of an authorship attribution contest. The challenge involved twelve participants (or participant teams) working with a controlled corpus that Juola had carefully compiled. The multilingual corpus is divided into thirteen different ''problem sets'' with varying degrees of complexity. Participants were invited to apply methods of their choice to the thirteen diverse problem sets. After the deadline for submissions, Juola compiled the attribution results and ranked the participants in terms of their accuracy, thus providing an empirical evaluation of approaches. The highest scoring researchers (see Koppel and Schler 2004) ''scored an average success rate of 71% . . . using Support Vector Machine with a linear kernel function'' (Juola *et al.*, 2006).[2]

One difficulty with the Juola experiment is that several of the thirteen problem sets are quite small (allowing researchers limited ability to train a classifier on multiple samples of known authorship)[3] and, thus, the corpus does not facilitate thorough cross-validation testing within the known texts. While a corpus with a very small sample size may be more realistic than the massive corpus utilized by Zhao and Zobel, the small corpus size hampered the ability of competitors to properly tune their algorithms. Most classification methods involve one or more tuning parameters (such as the number of features used in the analysis) that must be chosen based on the specific problem at hand.

Given the small sample size, competitors were most likely forced to select tuning parameter values without proper cross-validation. For this reason, the results for at least some of the problems within the Juola experiment are likely to contain an element of randomness that could have been avoided if a larger corpus had been available. Nonetheless, the empirical testing performed in this experiment is certainly worthwhile, and the results of the contest are valuable and should inspire additional analysis.

In this research, we offer a natural extension of the work begun by Zhao and Zobel. Our objective is three-fold: to expose the authorship attribution community to classification methods that have not been previously applied to authorship problems, to compare the relative performance of these methods, and to apply these methods to a classic authorship attribution problem. Additionally, we compare classification results obtained using two possible feature sets. The first feature set is not pre-processed: we let the classification algorithms determine which features to use based on cross-validation of internal tuning parameters. The second feature set is pre-processed in order to filter out context-sensitive features and limit the available feature set to features of a certain frequency that are common to all texts in the corpus. Our classification methods and feature selection methodologies are described in further detail below.

For our test corpus, we decided against using Zhao and Zobel's corpus of newswire articles on the grounds that it is both unfamiliar to authorship researchers and not typical of authorship problems.[4] We decided against employing the Juola problem set corpus due the small sample size and relative obscurity of at least some of the problems. The Federalist Papers corpus was selected for this research on the grounds that it met the two primary criteria of being both familiar to authorship researchers and of adequate size to afford thorough testing. As pointed out by a reviewer, the Federalist corpus is not the only suitable problem set for a benchmarking analysis. However, in addition to being one of the most widely used corpora for authorship attribution testing and investigation, the Federalist Papers is a "real" authorship corpus, with an ample set of works of known authorship and a smaller subset of disputed texts. It has the advantage of being well understood. As early as 1997, Richard S. Forsyth had noted that the Federalist Paper problem "is possibly the best candidate for an accepted benchmark in stylometry" (Forsyth 1997).[5] The Federalist collection has the advantage of being homogeneous and the "closed set" of potential authors tightly constrained. Moreover, significant prior research on the Federalist Papers provides opportunity for comparison with other approaches.[6]

2 Classification Methods

We compare five classification methods in this analysis: Delta (Argamon 2008; Burrows 2002; Hoover 2004a,b), k-nearest neighbors (KNN), the support vector machine (SVM), nearest shrunken centroids (NSC; Tibshirani *et al.*, 2003), and regularized discriminant analysis (RDA; Guo *et al.*, 2007). Of these methods, only one (Delta) is specifically designed for authorship attribution. An overview of the other methods can be found in Hastie *et al.* (2009). They are general-purpose classification methods from the machine learning literature. Most of these methods involve one or more tuning parameters, which are chosen via cross-validation on the training data. Delta involves a tuning parameter that determines the number of features used in the classification. KNN's tuning parameter is the number of nearest neighbors to be used in classification of a test observation. Many versions of SVMs exist in the literature. We used a SVM with a linear kernel and one-against-one classification, as implemented in the "e1071" library of the statistical software language R. The SVM has a single tuning parameter, which determines the cost of violating the constraints. NSC has one tuning parameter, which controls the number of features used, and RDA has two tuning parameters, one of which controls the number of features used.[7] It is worth noting that KNN and SVM result in classifiers that use all of the features present in the data, whereas

NSC, RDA and Delta perform built-in feature selection.

3 Text Acquisition, Preparation, and Tokenization

The text of the Federalist Papers was acquired through Project Gutenberg and compared against the versions available online at the Avalon Project of Yale's Law School (*The Federalist Papers* 2009).[8] In some cases formatting corrections to the Gutenberg texts were made and in all cases the boilerplate Gutenberg text was removed before analysis. To allow for word and bigram tokenization using the scripts developed by the authors, the corrected text was first marked up into XML. "Div" elements separated each paper and its author-related metadata: its title and the status of the paper's authorship (e.g. Madison, Hamilton, Jay, Coauthored, Disputed). Using scripts developed for this project, the XML text was lowercased and tokenized in order to produce raw counts and relative frequencies for each word and word bigram within each text sample. The decision to use word features alone was based on the growing consensus that analysis of high frequency words (mostly function, or closed-class words) and/or n-grams provide the most consistently reliable results in authorship attribution problems (see, for example, Burrows 2002; Diederich *et al.*, 2000; Grieve 2007; Hoover 2003a,b; Koppel *et al.*, 2007; Martindale and McKenzie 1995; Uzuner and Katz 2005; Zhao and Zobel 2005; Yu 2008). We formatted the resulting data as a matrix of dimension 85×69,969 (number of texts by number of features). Further analysis and handling of the data was conducted in the open-source R statistical software package (http://cran.r-project.org).

4 Data Pre-processing

We performed all analyses on two different versions of the data, which we will refer to as "raw features" and "pre-processed features".

Despite its name, we did impose one restriction on the features contained in the raw data set. We required that every feature occur at least once in the texts of Jay, Madison, and Hamilton. That is, a feature that occurs in Hamilton and Madison texts but never in Jay was excluded from the analysis. This was done both for computational convenience and in order to avoid context-specific features that might skew the attribution by subject over style. The resulting matrix was of dimension 85×2,907.

The pre-processed feature set was composed of the subset of the raw features that appear across the corpus with a mean relative frequency of at least 0.05%. That is, the pre-processed feature set consists of features that are used by each author and also occur with sufficiently high overall relative frequency. This pre-processing results in the exclusion of features that are likely to be context sensitive (the effect of which is to avoid classifying texts based on a shared subject rather than upon a shared style). The pre-processed data matrix was of dimension 85×298.

With the exception of Delta, all methods were performed on the 85×2,907 or 85×298 matrix of feature frequencies, where the count for a given feature in a given text was converted to a frequency by dividing it by the total number of frequency counts occurring in that text. Delta was performed as specified in Argamon (2008) and Burrows (2002, 2003).

5 Exploratory Data Analysis

To visually explore the data before performing classification, we used principal components analysis (PCA) on the raw feature set. PCA (see Hastie *et al.*, 2009) is a method for projecting data on to a low-dimensional subspace that is frequently used in authorship attribution research.[9] It is not possible to visualize the eighty-five texts directly, since they lie in 2,907-dimensional space. Instead, we compute principal components (PCs), which are directions in the 2,907-dimensional space along which much of the variance of the data occurs. The first PC is the dimension that explains the greatest possible part of the variation in the data, the second PC

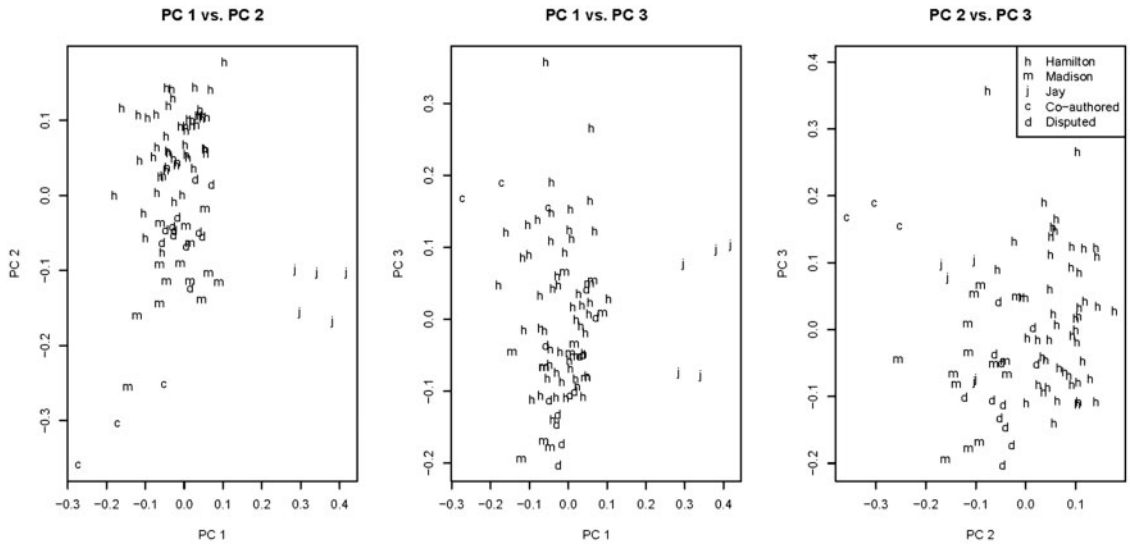

Fig. 1 PCA applied to the eighty-five texts. The first PC direction clearly separates the Jay texts from the others, and the second and third directions separate the coauthored texts. The Madison, Hamilton, and disputed texts lie close together. In the plot of PC 2 against PC 3, there is a suggestion that some of the disputed texts may lie closer to Madison than to Hamilton. The first three PCs explain 22.9, 8.6, and 6.9% of the variance, respectively

explains the next greatest possible part of the variation, and so on.

In Fig. 1, we plot the projections of the eighty-five texts on to the first, second, and third PCs. We see that the first PC shows separation between Jay and the other authors, and the second PC shows separation between the coauthored texts and the others. However, the first PC does not distinguish between Hamilton, Madison, and the disputed samples. The second and third PCs suggest some separation between Hamilton and Madison, but the disputed samples are located between Hamilton and Madison in this projection. There is a suggestion that some of the disputed samples may lie closer to Madison than to Hamilton in the plot of the second and third PCs. Though principal components analysis reveals some degree of separation between the authors, it does not provide a tool for predicting the authorship of a new text.

6 Results

We will refer to the Hamilton, Madison, and Jay samples as the training data, and the disputed and coauthored texts will be called the test data. For each method, we performed ten-fold cross-validation on the training data in order to estimate its accuracy and select tuning parameter values. Then the methods were fit on the full training data set (using the tuning parameter values that resulted in smallest cross-validation errors) and tested on the test data set. The resulting output constitutes our predictions for the disputed and coauthored texts. Each method was performed on two feature sets: the

Table 1 Fewest cross-validation errors obtained using each method

Method	Number of CV errors	Number of features used
Delta pp	3/70	75
Delta Raw	3/70	400
KNN pp	3/70	298
KNN Raw	2/70	2,907
NSC pp	0/70	199
NSC Raw	0/70	718
RDA pp	1/70	243
RDA Raw	0/70	312
SVM pp	4/70	298
SVM Raw	10/70	2,907

pp, pre-processed.

Table 2 Attributions by method for disputed and coauthored papers

Paper	NSC Raw	SVM Raw	KNN Raw	Delta Raw	RDA Raw	NSC pp	SVM pp	KNN pp	Delta pp	RDA pp
No. 18	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Hamilton	Madison	Madison
No. 19	Madison	Madison	Hamilton	Madison	Madison	Madison	Madison	Hamilton	Madison	Madison
No. 20	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Hamilton	Madison	Madison
No. 49	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Hamilton	Madison	Madison
No. 50	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Hamilton	Madison
No. 51	Madison	Madison	Hamilton	Madison	Madison	Madison	Madison	Madison	Madison	Madison
No. 52	Madison	Madison	Hamilton	Madison	Madison	Madison	Madison	Madison	Madison	Madison
No. 53	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison
No. 54	Madison	Hamilton	Madison	Madison	Madison	Madison	Madison	Hamilton	Madison	Madison
No. 55	Madison	Hamilton	Madison	Madison	Hamilton	Madison	Madison	Madison	Madison	Hamilton
No. 56	Madison	Madison	Hamilton	Madison	Madison	Madison	Madison	Madison	Madison	Madison
No. 57	Madison	Hamilton	Hamilton	Madison	Madison	Madison	Madison	Hamilton	Madison	Madison
No. 58	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison
No. 62	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison
No. 63	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison	Madison

Papers eighteen to twenty are thought to be coauthored by Hamilton and Madison, and the remaining are of unknown authorship.

raw feature set, and the pre-processed feature set, as described earlier.

The number of cross-validation errors resulting from each method is given in Table 1. These results suggest that in this experiment all of the methods performed quite well. NSC and RDA show very few cross-validation errors, and SVM performs the least effectively, with ten cross-validation errors reported with the raw feature set and four with the pre-processed data. The number of features used by each method is shown in Table 1. The NSC, RDA, and Delta classifiers derive added interpretability from the fact that they use only a subset of the words and bigrams; this subset can be examined in order to determine what types of words and bigrams differ between the authors. The smallest cross-validation error for KNN on the raw data was achieved using $K = 6$ neighbors.

The attributions for the disputed and coauthored papers are found in Table 2. The attribution data displayed in Table 2 is in general agreement with prior attribution work suggesting that most of the disputed papers are likely written by Madison. There is a hint that a few may be from Hamilton.

Since NSC was the best overall classifier in these tests, we provide a few additional observations regarding the NSC results. Table 3 shows the fifty most important features used by NSC. A positive score indicates that the given author used the word or word bigram more than average, and a negative score indicates underuse.

The NSC output includes probabilities indicating the classifier's certainty of each sample text belonging to a given candidate author, within the closed set of candidate authors. The probabilities are listed in Table 4.

The results in Table 4 suggest a surprisingly high degree of certainty that Madison authored the texts in question. In particular, NSC is confident that Madison authored even the papers that are known to have been coauthored. In interpreting these probabilities, is it important to keep in mind that NSC (like any classifier) is quite sensitive to the parameters used to fit the model, such as the feature set and specific choice of training samples used. The extremely high probabilities in Table 4 may be due in part to the use of context-specific words by the classifier. That is, if the Madison training texts and the test texts address a particular topic that is not addressed by the Hamilton or Jay training texts, then the NSC classifier might use these words as very strong evidence that the test texts were written by Madison. Therefore, one should interpret the probabilities in Table 4 as the probability of each test text being written by a given author *under the NSC model*. The sensitivity of an individual classifier argues for the application of multiple classifiers for a single problem, as we have done in this study.[10]

Table 3 Fifty most important features for NSC classifier based on raw data

Feature	Hamilton Score	Madison Score	Jay Score
a	0.2412	−0.105	−1.7906
america	−0.1449	0	1.4893
an	0.2422	−0.2594	−1.3678
and	−0.3347	0.297	2.2061
arms and	−0.0289	0	0.6999
be more	−0.0283	0	0.8047
been	0	0.0773	−0.7353
by	−0.2041	0.805	0
by the	−0.1438	0.8094	−0.2656
confederacies	−0.009	0	1.1143
four	0	−0.0512	0.6787
given	−0.0156	0	0.8156
has been	0.023	0	−0.7837
importance	−0.0301	0	0.7586
in the	0.0859	0	−0.6783
in	0.2102	−0.2745	−0.9989
independent	0	0	0.7267
is	0	0.0859	−0.9019
national government	0	−0.1011	0.8511
nations	−0.024	−0.0017	1.287
of	0.3769	−0.1734	−2.9831
of a	0.2567	−0.3661	−1.2168
of america	−0.1082	0	1.102
of the	0.1998	0	−2.0792
on	−0.3511	1.31	0
on the	−0.2469	1.0483	0
one	−0.063	0	0.6553
only to	−0.0737	0	0.7166
others	−0.0558	0	0.69
powers	−0.1778	0.8355	0
soon	−0.0077	0	0.7338
that they	−0.0644	0	1.0944
the	0.1427	0.5186	−3.9449
them	−0.0568	0	0.736
there	0.3111	−0.791	−0.5824
they	−0.0204	0	0.8312
they should	−0.0283	0	0.7041
this	0.2419	−0.2932	−1.2704
three	−0.0182	0	0.8145
to	0.4428	−1.0046	−1.3278
to be	0.0591	0	−0.7669
to the	0.1271	0	−1.2119
treaties	−0.0486	0	1.0861
treaties and	−0.0634	0	0.9153
upon	0.5181	−1.3328	−1.1763
useful	−0.0103	0	0.6865
well	−0.0465	0	0.982
well as	−0.0576	0	0.7236
which	0.0504	0	−1.1303
wise	0	−0.0307	1.0514

Table 4 NSC probabilities

Paper	Hamilton	Madison	Jay
No. 18	0	1	0
No. 19	0	1	0
No. 20	0	1	0
No. 49	0	1	0
No. 50	0	1	0
No. 51	0	1	0
No. 52	0	1	0
No. 53	0	1	0
No. 54	0	1	0
No. 55	0.0038	0.9962	0
No. 56	1e−04	0.9999	0
No. 57	0	1	0
No. 58	0	1	0
No. 62	0	1	0
No. 63	0	1	0

A reviewer asked how our results would change if we did not initially restrict the ''raw'' data set to consist of words that occur in all three of the authors' training texts. To answer this question, all analyses were repeated on the data set of dimension $85 \times 69,969$ consisting of all features that occur in any of the texts of known authorship. Restricting this data to the features that occur with a mean relative frequency of at least 0.05% (the ''pre-processed'' data) resulted in a data set of dimension 85×158. (The pre-processed data set is smaller when the raw data set contains more features, since this reduces the relative frequency of each feature.) The results on these new raw and pre-processed data sets were substantively similar to those reported earlier. In particular, NSC assigned all texts to Madison using both the raw (3/70 CV errors) and the pre-processed (1/70) data sets.

7 Conclusions

Machine learning methods that are not specific to authorship attribution perform very well on this problem and may perform well on other authorship attribution problems as well.[11] While SVM has been employed rather extensively on authorship problems (e.g. Fung 2003; Hirst and Feiguina 2007; Juola *et al.*, 2006), NSC and RDA have not, and

both performed particularly well in these experiments. Authorship attribution researchers would likely benefit from greater exposure to and further testing of these approaches. NSC has the advantage over RDA of having only one tuning parameter, as well as greater interpretability. With the exception of our prior work with NSC (Jockers *et al.*, 2008), neither NSC nor RDA has been employed in authorship attribution problems. These are two examples of high-dimensional classification methods that result from regularization of classical methods intended for low-dimensional settings. We believe that these methods and other penalized classification approaches show promise for the problem of authorship attribution, which is characterized by high dimensionality and low sample size.

References

Argamon, S. (2008). Interpreting Burrows's delta: geometric and probabilistic foundations. *Literary Linguistic Comput: J Assoc Literary Linguistic Comput*, **23**(2): 131–47.

Bosch, R. A. and Smith, J. A. (1998). Separating hyperplanes and the authorship of the disputed Federalist Papers. *Am Math Monthly*, **105**(7): 601–08.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary Linguistic Comput: J Assoc Literary Linguistic Comput*, **17**(3): 267–87.

Burrows, J. (2003). Questions of authorship: attribution and beyond. *Comput Humanities*, **37**(1): 5–32.

Burrows, J. (2007). All the way through: testing for authorship in different frequency strata. *Literary Linguistic Comput: J Assoc Literary Linguistic Comput*, **22**(1): 27–47.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2000). Authorship attribution with support vector machines. *Appl Intell*, **19**(1–2): 109–23.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM '98)*. Bethesda, Maryland. November 3–7 pp. 148–55.

Forsyth, R. S. (1997). Towards a text benchmark suite. *Association for Computers and the Humaniteis*. Ontario: Kingston.

Fung, G. (2003). The disputed Federalist Papers: Svm feature selection via concave minimization. *Proceedings of the 2003 Conference on Diversity in Computing*. Atlanta, GA, pp. 42–6.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary Linguistic Comput: J Assoc Literary Linguistic Comput*, **22**(3): 251–70.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn*. New York: Springer.

Hirst, G. and Feiguina, O. G. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary Linguistic Comput: J Assoc Literary Linguistic Comput*, **22**(4): 405–17.

Holmes, D. I. and Forsyth, R. S. (1995). The Federalist revisited: new directions in authorship attribution. *Literary Linguist Comput: J Assoc Literary Linguist Comput*, **10**(2): 111–27.

Hoover, D. L. (2001). Statistical stylistics and authorship attribution: an empirical investigation. *Literary Linguist Comput: J Assoc Literary Linguist Comput*, **16**(4): 421–44.

Hoover, D. L. (2003a). Another perspective on vocabulary richness. *Comput Humanities*, **37**(2): 151–78.

Hoover, D. L. (2003b). Multivariate analysis and the study of style variation. *Literary Linguist Comput: J Assoc Literary Linguist Comput*, **18**(4): 341–60.

Hoover, D. L. (2004a). Delta prime? *Literary Linguist Comput: J Assoc Literary Linguist Comput*, **19**(4): 477–95.

Hoover, D. L. (2004b). Testing Burrows's delta. *Literary Linguist Comput: J Assoc Literary Linguist Comput*, **19**(4): 453–75.

Hoover, D. L. (2008). Quantitative analysis and literary studies. *A Companion to Digital Literary Studies*. In Schreibman, S. and Siemens, R. (eds), Oxford: Blackwell.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *European Conference on Machine Learning (ECML)*. Berlin, pp. 137–42.

Jockers, M. L., Witten, D. M., and Criddle, C. S. (2008). Reassessing authorship in the book of Mormon using nearest Shrunken centroid classification. *Literary Linguist Comput J Assoc Literary Linguist Comput*, **23**: 465–91.

Juola, P., Sofko, J., and Brennan, P. (2006). A prototype for authorship attribution studies. *Literary Linguist Comput: J Assoc Literary Linguist Comput*, **21**(2): 169–78.

Khmelev, D. V. and Tweedie, F. J. (2001). Using Markov chains for identification of writers. *Literary Linguist Comput J Assoc Literary Linguist Comput*, **16**(3): 299–307.

Koppel, M. and Schler, J. (2004). Ad-hoc authorship attribution competition approach outline. *Ad-Hoc Authorship Attribution Contest,* In Juola, P. ACH/ALLC.

Koppel, M., Schler, J., Argamon, S., and Messeri, E. (2006). Authorship attribution with thousands of candidate authors. *Proceedings of the 29th International Conference of the Special Interest Group on Information Retrieval*. Seattle, WA, pp. 659–60.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability: unmasking pseudonymous authors. *J Mach Learn Res*, **8**: 1261–76.

Luyckx, K. and Daelemans, W. Authorship attribution and verification with many authors and limited data. *Proceedings of the 22nd international Conference on Computational Linguistics - Volume 1*, pp. 513–20.

Martindale, C. and Mckenzie, D. (1995). On the utility of content analysis in author attribution: The Federalist. *Comput Humanities*, **29**(4): 259–70.

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

The Federalist Papers. Access (2009). Yale Law School, Lillian Goldman Law Library 2009 [cited 1/20/2009 2009]. Available from http://avalon.law.yale.edu/subject_menus/fed.asp.

Tweedie, F. J., Singh, S., and Holmes, D. I. (1996). Neural network applications in stylometry: The Federalist Papers. *Comput Humanities*, **30**(1): 1–10.

Uzuner, O. and Katz, B. (2005). A comparative study of language models for book and author recognition. *Lecture Notes in Computer Science*. Berlin: Springer.

Yang, Y. and Pedersen, J. (1997). A comparative study on feature selection in text categorization. *Proceedings of the 14th International Conference on Machine Learning (ICML '97),* Nashville, Tennessee, 8–12 July, pp. 412–20.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary Linguist Comput*, **23**: 327–43.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science*. Berlin: Springer.

Notes

1 See also http://www.mathcs.duq.edu/~juola/authorship_contest.html.

2 For more on the use of SVM, see also Koppel *et al.* (2006, 2007).

3 And, the individual samples were frequently too short to allow for text segmentation.

4 Several other studies (Dumais *et al.*, 1998; Joachims 1998; Yang and Pedersen 1997) evaluating text classification algorithms use texts drawn from similar resources, i.e. news articles and web documents.

5 See http://www.ach.org/abstracts/1997/p026.html.

6 Among others, see in particular (Bosch and Smith, 1998; Fung, 2003; Holmes and Forsyth, 1995; Khmelev and Tweedie, 2001; Martindale and McKenzie, 1995; Mosteller and Wallace, 1964; Tweedie *et al.*, 1996).

7 The second RDA parameter controls regularization of the covariance matrix.

8 See http://avalon.law.yale.edu/subject_menus/fed.asp.

9 Hoover (2008) provides a useful overview of PCA in authorship research.

10 Luyckx and Daelemans (2008) argue that many authorship studies overestimate the performance of their systems by limiting the pool of potential candidates to too small a set of potential authors. The performance observed here should be viewed in the context of this particular problem and not necessarily extended to other authorship problems involving hundreds (Luyckx and Daelemans, 2008) or thousands (Koppel *et al.*, 2006) of potential authors.

11 For example, we found (Jockers *et al.*, 2008) that NSC performed substantially better than Delta.