Author Identification on the Large Scale

David Madigan[1,3] Alexander Genkin[1] David D. Lewis[2]
Shlomo Argamon[4]
Dmitriy Fradkin[1,5] and Li Ye[3]
(1) DIMACS, Rutgers University
(2) David D. Lewis Consulting
(3) Department of Statistics, Rutgers University
(4) Department of Computer Science, Illinois Institute of Technology
(5) Department of Computer Science, Rutgers University

1 Introduction

Individuals have distinctive ways of speaking and writing, and there exists a long history of linguistic and stylistic investigation into authorship attribution. In recent years, practical applications for authorship attribution have grown in areas such as intelligence (linking intercepted messages to each other and to known terrorists), criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (tracking authors of computer virus source code). This activity is part of a broader growth within computer science of identification technologies, including biometrics (retinal scanning, speaker recognition, etc.), cryptographic signatures, intrusion detection systems, and others.

Automating authorship attribution promises more accurate results and objective measures of reliability, both of which are critical for legal and security applications. Recent research has used techniques from machine learning [3, 10, 13, 31, 50], multivariate and cluster analysis [24, 25, 8], and natural language processing [5, 46] in authorship attribution. These techniques have also been applied to related problems such as genre analysis [4, 1, 6, 17, 23, 46] and author profiling (such as by gender [2, 12] or personality [38]).

Our focus in this paper is on techniques for identifying authors in large collections of textual artifacts (e-mails, communiques, transcribed speech, etc.). Our approach focuses on very high-dimensional, topic-free document representations and particular attribution problems, such as: (1) Which one of these K authors wrote this particular document? (2) Did any of these K authors write this particular document?

Scientific investigation into measuring style and authorship of texts goes back to the late nineteenth century, with the pioneering studies of Mendenhall [36] and Mascol [34, 35] on distributions of sentence and word lengths in works of literature and the gospels of the New Testament. The underlying notion was that works by different authors are strongly distinguished by quantifiable features of the text. By the mid-twentieth century, this line of research had grown into what became known as "stylometrics", and a variety of textual statistics had been proposed to quantify textual style. The style of early work was characterized by a search for invariant properties of textual statistics, such as Zipf's distribution and Yule's K statistic.

Modern work in authorship attribution (often referred to in the humanities as "nontraditional authorship attribution") was ushered in by Mosteller and Wallace in the 1960s, in their seminal study *The Federalist Papers* [37]. The study examined 146 political essays from the late eighteenth century, of which most are of acknowledged authorship by John Jay, Alexander Hamilton, and James Madison, though twelve are claimed by both Hamilton and Madison. Mosteller and Wallace showed statistically significant discrimination results by applying Bayesian statistical analysis to the frequencies of a small set of 'function words' (such as 'the', 'of', or 'about'), as stylistic features of the text. Function words, and other similar classes of words, remain the most popular stylistic features used for authorship discrimination. As we shall see below, reliance on a particular representation (e.g., function words) can lead to misplaced confidence in subsequent predictions.

Other stylometric features that have been applied include various measures of vocabulary richness and lexical repetition, based on Zipf's studies on word frequency distributions. Most such measures, however, are strongly dependent on the length of the text being studied, and so are difficult to apply reliably. Many other types of features have been applied, including word class frequencies [2, 18], syntactic analysis [5, 46], word collocations [45], grammatical errors [27], and word, sentence, clause, and paragraph lengths [3, 33]. Many studies combine features of different types using multivariate analysis techniques.

One widely-used technique, pioneered for authorship studies by Burrows [8], is to use principal components analysis (PCA) to find combinations of style markers that can discriminate between a particular pair (or small set) of authors. This method has been used in several studies, including [5]. Another related class of techniques that have been applied are machine learning algorithms (such as Winnow [30] or Support Vector Machines [11]) which can construct discrimination models over large numbers of documents and features. Such techniques have been applied widely in topic-based text categorization (see the excellent survey [42]) and other stylistic discrimination tasks (e.g. [2, 26, 46]), as well as for authorship discrimination [3, 13]. Often, studies have relied on intuitive evaluation of results, based on visual inspection of scatter-plots and cluster-analysis trees, though recent work (e.g. [3, 12, 13]) has begun to apply somewhat more rigorous tests of statistical significance and cross-validation accuracy.

2 Representation

Document representation provides the central challenge in author attribution. Features should capture aspects of author style that persist across topics. Traditional stylometric features include function words, high-frequency words, vocabulary richness, hapax legomena, Yules K, syllable distributions, character level statistics, and punctuation. Much of the prior work focuses on relatively low-dimensional representations. However, newer statistical algorithms as well as increases in computing power now enable much richer representations involving tens or hundreds of thousands of features.

Don Foster's successful attribution of "Primary Colors" to Joe Klein illustrates the value of idiosyncratic features such as rare adjectives ending in "inous" (e.g., vertiginous) or words beginning with hyper-, mega-, post-, quasi-, and semi-. Our own work focuses on word-endings and parts-of-speech in addition to the classical function words.

On key challenge concerns the notion of a "topic-free" feature. The stylometry

literature has long considered function words to be topic-free in the sense that the relative frequency with which an author uses, for example, "with," should be the same regardless of whether the author is describing cooking recipes or the latest news about the oil futures market. We know of no prior work that defines the topic-free notion or formally assesses candidate features in this regard.

3 Bayesian multinomial logistic regression

Traditional 1-of-k author identification requires a multiclass classification learning method and implementation that are highly scalable. The most popular methods for multiclass classification in recent machine learning research are variants on support vector machines and boosting, sometimes combined with error-correcting codes approach. Rifkin and Klautau provide a review [40].

In contrast, we turned to polytomous or multinomial logistic regression because of its probabilistic character. Since this model outputs an estimate of the probability that the input belongs to each of the possible classes, we can easily take into account the relative costs of different misidentifications when making a classification decision. If those costs change, classifications can be altered appropriately, without retraining the model.

Further, the Bayesian perspective on training a multinomial logistic regression model allows training data and domain knowledge to be easily combined. While this study looks at relatively simple forms of prior knowledge about features, in other work we have explored incorporating prior knowledge about predictive features, and hierarchical Bayesian structures that allow sharing information across related problems (e.g. identifying an author's work in different genres).

To begin, let $\boldsymbol{x} = [x_1, ..., x_j, ..., x_d]^T$ be a vector of feature values characterizing a document to be identified. We encode the fact that a document belongs to a class (e.g. an author) $k \in \{1, ..., K\}$ by a K-dimensional 0/1 valued vector $\boldsymbol{y} = (y_1, ..., y_K)^T$, where $y_k = 1$ and all other coordinates are 0.

Multinomial logistic regression is a conditional probability model of the form

$$p(y_k = 1|\boldsymbol{x}, \mathbf{B}) = \frac{\exp(\boldsymbol{\beta}_k^T \boldsymbol{x})}{\sum_{k'} \exp(\boldsymbol{\beta}_{k'}^T \boldsymbol{x})}, \tag{1}$$

parameterized by the matrix $\mathbf{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K]$. Each column of $\mathbf{B}$ is a parameter vector corresponding to one of the classes: $\boldsymbol{\beta}_k = [\beta_{k1}, ..., \beta_{kd}]^T$. This is a direct generalization of binary logistic regression to the multiclass case.

Classification of a new observation is based on the vector of conditional probability estimates produced by the model. In this paper we simply assign the class with the highest conditional probability estimate:

$$\hat{y}(\mathbf{x}) = \arg\max_k p(y_k = 1|\mathbf{x}).$$

In general, however, arbitrary cost functions can be used and the classification chosen to minimize expected risk under the assumption that the estimated probabilities are correct [14].

Consider a set of training examples $D = \{(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_i}, \mathbf{y_i}), \ldots, (\mathbf{x_n}, \mathbf{y_n})\}$. Maximum likelihood estimation of the parameters $\mathbf{B}$ is equivalent to minimizing the negated log-likelihood:

$$l(\mathbf{B}|D) = -\sum_i \left[\sum_k y_{ik} \boldsymbol{\beta}_k^T \mathbf{x_i} - \ln \sum_k \exp(\boldsymbol{\beta}_k^T \mathbf{x_i}) \right], \tag{2}$$

Since the probabilities must sum to one: $\sum_k p(y_k = 1|\boldsymbol{x}, \mathbf{B}) = 1$, one of the vectors $\boldsymbol{\beta}_k$ can be set to $\boldsymbol{\beta}_k = \mathbf{0}$ without affecting the generality of the model. This is in fact necessary for maximum likelihood estimation for $\mathbf{B}$ to be identifiable in a formal sense (whether or not in practice identifiable for a given data set). This restriction is not necessary for identifiable in the Bayesian approach, and in some cases there are advantages in not imposing this restriction, as we will discuss.

As with any statistical model, we must avoid overfitting the training data for a multinomial logistic regression model to make accurate predictions on unseen data. One Bayesian approach for this is to use a prior distribution for $\mathbf{B}$ that assigns a high probability that most entries of $\mathbf{B}$ will have values at or near 0. We now describe two such priors.

3.1 Types of priors

Perhaps the most widely used Bayesian approach to the logistic regression model is to impose a univariate Gaussian prior with mean 0 and variance σ_{kj}^2 on each parameter β_{kj}:

$$p(\beta_{kj}|\sigma_{kj}) = N(0, \sigma_{kj}) = \frac{1}{\sqrt{2\pi}\sigma_{kj}} \exp(\frac{-\beta_{kj}^2}{2\sigma_{kj}^2}). \tag{3}$$

By specifying a mean of 0 for each Gaussian, we encode our prior belief that β_{kj} will be near 0. The variances of the Gaussians, σ_{kj}, are positive constants we must specify. A small values of σ_{kj} represents a prior belief that β_{kj} is close to zero, while larger value represents less confidence in this. In the simplest case we let σ_{kj} equal the same σ for all j, k. We assume *a priori* that the components of $\mathbf{B}$ are independent and hence the overall prior for $\mathbf{B}$ is the product of the priors for its components. Finding the maximum *a posteriori* (MAP) estimate of $\mathbf{B}$ with this prior is equivalent to ridge regression (Hoerl and Kennard, 1970) for the multinomial logistic model. The MAP estimate of $\mathbf{B}$ is found by minimizing:

$$l_{ridge}(\mathbf{B}|D) = l(\mathbf{B}|D) + \frac{1}{\sigma_{kj}^2} \sum_j \sum_k \beta_{kj}^2. \tag{4}$$

Ridge logistic regression has been widely used in text categorization, see for example [52, 29, 51]. The Gaussian prior, while favoring values of β_{kj} near 0, does not favor them being exactly equal to 0. Absent unusual patterns in the data, the MAP estimates of all or almost all β_{kj}'s will be nonzero. Since multinomial logistic regression models for author identification can easily have millions of parameters, such dense parameter estimates could lead to inefficient classifiers.

However, sparse parameter estimates can be achieved in the Bayesian framework remarkably easily. Suppose we use double exponential (Laplace) prior distribution on the β_{kj}:

$$p(\beta_{kj}|\lambda_{kj}) = \frac{\lambda_{kj}}{2} \exp(-\lambda_{kj}|\beta_{kj}|). \tag{5}$$

As before, the prior for $\mathbf{B}$ is the product of the priors for its components. For typical data sets and choices of λ's, most parameters in the MAP estimate for $\mathbf{B}$ will be zero. Figure 1 compares the density functions for the Gaussian and Laplace distributions, showing the cusp that leads to zeroes in the MAP parameter estimates.

Finding the MAP estimate is done by minimizing:

$$l_{lasso}(\mathbf{B}|D) = l(\mathbf{B}|D) + \lambda_{kj} \sum_j \sum_k |\beta_{kj}|. \tag{6}$$

Tibshirani [48] was the first to suggest Laplace priors in the regression context. He pointed out that the MAP estimates using the Laplace prior are the same as the estimates produced by applying *lasso* algorithm [48]. Subsequently, the use of constraints or penalties based on the absolute values of coefficients has been used to achieve sparseness in a variety of data fitting tasks (see, for example, [15, 16, 20, 49, 44]), including multinomial logistic regression [28].

In large-scale experiments with binary logistic regression on content-based text categorization we found lasso logistic regression produced models that were not only sparse, but systematically outperformed ridge logistic regression models [19].

The lasso approach is even more appealing with multinomial logistic regression. A feature which is a strong predictor of a single class will tend to get a large β_{kj} for that class, and a β_{kj} of 0 for most other classes, aiding both compactness and interpretability. This contrasts with the ridge, where the β_{kj} for all classes will usually be nonzero. This also suggests we may not want to automatically set $\boldsymbol{\beta}_{\mathbf{k}}$ to $\mathbf{0}$ for a "base" class as is usual in maximum likelihood fitting. If all classes are meaningful (i.e. not "other" class) then the model will be more understandable if all classes are allowed to have their distinctive features.

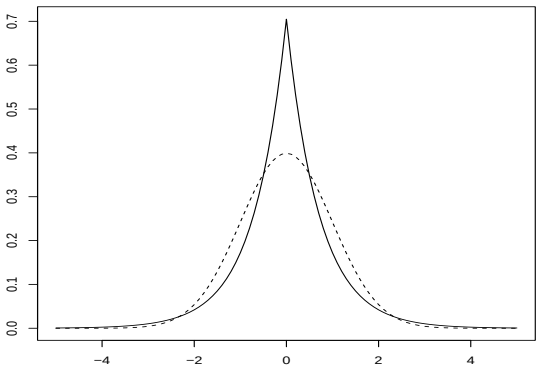

Figure 1: The density of the Laplace and Gaussian (dashed line) distributions with the same mean and variance.

3.2 Algorithm

3.2.1 Algorithmic approaches to multinomial logistic regression

A wide variety of algorithms have been used for fitting the multinomial logistic regression model, and we discuss only a few results here.

Several of the largest scale studies have occurred in computational linguistics, where the maximum entropy approach to language processing leads to multinomial logistic regression models. Malouf [32] studied parsing, text chunking, and sentence extraction problems with very large numbers of classes (up to 8.6 million) and sparse inputs (with up to 260,000 features). He found that for the largest problem a limited memory Quasi-Newton method was 8 times faster than the second best method, a Polak-Ribere-Positive version of conjugate gradient. Sha and Pereira [43] studied a very large noun phrase chunking problem (3 classes, and 820,000 to 3.8 million features) and found limited-memory BFGS (with 3-10 pairs of previous gradients

and updates saved) and preconditioned conjugate gradient performed similarly, and much better than iterative scaling or plain conjugate gradient. They used a Gaussian penalty on the loglikelihood. Goodman [21] studied large language modeling, grammar checking, and collaborative filtering problems using an exponential prior (a Laplace prior truncated at 0). He claimed not find a consistent advantage for conjugate gradient over iterative scaling, though experimental details are not given.

Another interesting study is that of Krishnapuram, Hartemink, Carin, and Figueiredo [28]. They experimented on small, dense classification problems from the Irvine archive using multinomial logistic regression with an L_1 penalty (equivalent to a Laplace prior). They claimed a cyclic coordinate descent method beat conjugate gradient by orders of magnitude but provided no quantitative data.

We base our work here on a cyclic coordinate descent algorithm for binary ridge logistic regression by Zhang and Oles [52]. In previous work we modified this algorithm for binary lasso logistic regression and found it fast and easy to implement [19]. A similar algorithm has been developed by Shevade and Keerthi [44].

3.2.2 Coordinate decent algorithm

Here we further modify the binary logistic algorithm we have used [19] to apply to ridge and lasso multinomial logistic regression. Note that both objectives (4) and (6) are convex, and (4) is also smooth, but (6) does not have a derivative at 0; we'll need to take special care with it.

The idea in the smooth case is to construct an upper bound on the second derivative of the objective on an interval around the current value; since the objective is convex, this will give rise to the quadratic upper bound on the objective itself on that interval. Minimizing this bound on the interval gives one step of the algorithm with the guaranteed decrease in the objective.

Let $Q(\beta_{kj}^{(0)}, \Delta_{kj})$ be an upper bound on the second partial derivative of the negated loglikelihood (2) with respect to β_{kj} in a neighborhood of β_{kj}'s current value $\beta_{kj}^{(0)}$, so that:

$$Q(\beta_{kj}^{(0)}, \Delta_{kj}) \geq \frac{\partial^2 l(\mathbf{B}|D)}{\partial \beta_{kj}^2} \quad \text{for all } \beta_{kj} \in [\beta_{kj}^{(0)} - \Delta_{kj}, \beta_{kj}^{(0)} + \Delta_{kj}].$$

Using Q we can upper bound the ridge objective (4) by a quadratic function of β_{kj}. The minimum of this function will be located at $\beta_{kj}^{(0)} + \Delta v_{kj}$ where

$$\Delta v_{kj} = \frac{-\frac{\partial l(\mathbf{B}|D)}{\partial \beta_{kj}} - 2\beta_{kj}^{(0)}/\sigma_{kj}^2}{Q(\beta_{kj}^{(0)}, \Delta_{kj}) + 2/\sigma_{kj}^2}. \tag{7}$$

Replacing $\beta_{kj}^{(0)}$ with $\beta_{kj}^{(0)} + \Delta v_{kj}$ is guaranteed to reduce the objective only if Δv_{kj} falls inside the trust region $[\beta_{kj}^{(0)} - \Delta_{kj}, \beta_{kj}^{(0)} + \Delta_{kj}]$. If not, then taking a step of size Δ_{kj} in the same direction will instead reduce the objective. The formula for computing the upper bound $Q(\beta_{kj}, \Delta_{kj})$ needed in this computation is described in the Appendix.

The algorithm in its general form is presented in Figure 2. The solution to the ridge regression formulation is found by using (7) to compute the tentative step at Step 2 of the algorithm. The size of the approximating interval Δ_{kj} is critical for the speed of convergence: using small intervals will limit the size of the step, while

(1) initialize $\beta_{kj} \leftarrow 0, \Delta_{kj} \leftarrow 1$ for $j = 1, ..., d,\ k = 1, ..., K$
for $t = 1, 2, ...$ **until** convergence
for $j = 1, ..., d$
for $k = 1, ..., K$
(2) compute tentative step Δv_{kj}
(3) $\Delta\beta_{kj} \leftarrow \min(\max(\Delta v_{kj}, -\Delta_{kj}), \Delta_{kj})$ (reduce the step to the interval)
(4) $\beta_{kj} \leftarrow \beta_{kj} + \Delta\beta_{kj}$ (make the step)
(5) $\Delta_{kj} \leftarrow \max(2|\Delta\beta_{kj}|, \Delta_{kj}/2)$ (update the interval)
end
end
end

Figure 2: Generic coordinate decent algorithm for fitting Bayesian multinomial logistic regression.

having large intervals will result in loose bounds. We therefore update the width, Δ_{kj}, of the trust region in Step 5 of the algorithm, as suggested by [52].

The lasso case is slightly more complicated because the objective (6) is not differentiable at 0. However, as long as $\beta_{kj}^{(0)} \neq 0$, we can compute:

$$\Delta v_{kj} = \frac{-\frac{\partial l(\mathbf{B}|D)}{\partial \beta_{kj}} - \lambda_{kj} s}{Q(\beta_{kj}^{(0)}, \Delta_{kj})}, \tag{8}$$

where $s = \text{sign}(\beta_{kj}^{(0)})$. We use Δv_{kj} as our tentative step size, but in this case must reduce the step size so that the new β_{kj} is neither outside the trust region, nor of different sign than $\beta_{kj}^{(0)}$). If the sign would otherwise change, we instead set β_{kj} to 0. The case where the starting value $\beta_{kj}^{(0)}$) is already 0 must also be handled specially. We must compute positive and negative steps separately using right-hand and left-hand derivatives, and see if either gives a decrease in the objective. Due to convexity, a decrease will occur in at most one direction. If there is no decrease in either direction β_{kj} stays at 0. Figure 3 presents the algorithm for computing Δv_{kj} in the Step 2 of the algorithm in Figure 2 for the lasso regression case.

Software implementing this algorithm has been made publicly available [1]. It scales up to 100's of classes, 100,000's of features and/or observations.

3.2.3 Strategies for choosing the upper bound

A very similar coordinate descent algorithm for fitting lasso multinomial logistic regression models has been presented by Krishnapuram, Hartemink, Carin, and Figueiredo [28]. However, they do not take into account the current value of $\mathbf{B}$ when computing a quadratic upper bound on the negated loglikelihood. Instead, they use the following bound on the Hessian of the negated (unregularized) loglikelihood [7]:

$$\mathbf{H} \leq \sum_i \frac{1}{2} \left[\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathbf{T}}/K\right] \otimes \mathbf{x_i}\mathbf{x_i}^T, \tag{9}$$

[1] *http://www.stat.rutgers.edu/∼madigan/BMR/*

if $\beta_{kj} \geq 0$
  compute Δv_{kj} by formula (8) with $s = 1$
  if $\beta_{kj} + \Delta v_{kj} < 0$ (trying to cross over 0)
    $\Delta v_{kj} \leftarrow -\beta_{kj}$
  endif
endif
if $\beta_{kj} \leq 0$
  compute Δv_{kj} by formula (8) with $s = -1$
  if $\beta_{kj} + \Delta v_{kj} > 0$ (trying to cross over 0)
    $\Delta v_{kj} \leftarrow -\beta_{kj}$
  endif
endif

Figure 3: Algorithm for computing tentative step of lasso multinomial logistic regression: replacement for Step 2 in algorithm Fig. 2.

where $\mathbf{H}$ is the $dK \times dK$ Hessian matrix; $\mathbf{I}$ is the $K \times K$ identity matrix; $\mathbf{1}$ is a vector of 1's of dimension K; $\otimes$ is the Kronecker matrix product; and matrix inequality $\mathbf{A} \leq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is negative semi-definite.

For a coordinate descent algorithm we only care about the diagonal elements of the Hessian. The bound (9) implies the following bound on those diagonal elements:

$$\frac{\partial^2 l(\mathbf{B}|D)}{\partial \beta_{kj}^2} \leq \frac{K-1}{2K} \sum_i x_{ij}^2. \tag{10}$$

As before, the exact second partial derivatives of the regularization penalties can be added to 10 to get bounds on the second partial derivatives of the penalized likelihoods. We then can use the result to put a quadratic upper bound on the negated regularized loglikelihood, and derive updates that minimize that quadratic function. For the ridge case the update is

$$\Delta v_{kj} = \frac{-\frac{\partial l(\mathbf{B}|D)}{\partial \beta_{kj}} - 2\beta_{kj}^{(0)}/\sigma_{kj}^2}{\frac{K-1}{2K}\sum_i x_{ij}^2 + 2/\sigma_{kj}^2}, \tag{11}$$

and for the lasso case the tentative update is:

$$\Delta v_{kj} = \frac{-\frac{\partial l(\mathbf{B}|D)}{\partial \beta_{kj}} - \lambda_{kj} s}{\frac{K-1}{2K}\sum_i x_{ij}^2}. \tag{12}$$

As before, a lasso update that would cause a β_{kj} to change sign must be reduced so that β_{kj} instead becomes 0.

The bound in 10 depends only on the number of classes K and the values taken on by each feature j, and holds at all values of $\mathbf{B}$. Therefore, in contrast to our bound $Q(\beta_{jk}, \Delta_{jk})$, it does not need to be recomputed when $\mathbf{B}$ changes, and no trust region is needed. On the downside, it is a much looser bound than $Q(\beta_{jk}, \Delta_{jk})$. In addition, since $Q(\beta_{jk}, \Delta_{jk})$ only uses information that is needed anyway for computation of first derivatives, the constancy of the bound in 10 provides only a

Group name	Contents	Postings	Authors
ARCHCOMP	Computational Archaeology	1007	298
ASTR	Theatre History	1808	224
BALT	Baltic Republics - politics	9842	23
DOTNET-CF	.NET Compact Framework	801	115
ICOM	International Council of Museums	1055	227

Table 1: Some Listserv group statistics.

minor savings. On the other hand, it seemed conceivable that eliminating the trust region might give a larger advantage, so we did an empirical comparison.

We compared training a lasso multinomial logistic regression model using each of the bounds on the data set Abalone from the UCI Machine Learning Repository [41]. This data set contains 27 classes, 11 variables, and 3133 observations. All aspects of the software (including the convergence tolerance) were identical except computation of the bounds, and omission of the trust interval test when using the bound in 10.

Training the classifier using the bound in 10 took 405 passes through the coordinates and 79 sec on a Pentium 4 PC, while with our bound it took only 128 iterations and 31 sec. While we have not conducted a detailed comparison, it appears that the looseness of the bound means that updates, while always valid, are not very large. More aggressive updates that must occasionally be truncated by the trust region boundary, and in turn adapt the size of the trust region, appears to be more efficient.

4 Experiments in one-of-k author identification

4.1 Data sets

Our first data set was based on RCV1-v2[2], a text categorization test collection based on data released by Reuters, Ltd.[3]. We selected all authors who had 200 or more stories each in the whole collection. The collection contained 114 such authors, who wrote 27,342 stories in total. We split this data randomly into training (75% - 20,498 documents) and test (25% - 6,844 documents) sets.

The other data sets for this research were produced from the archives of several listserv discussion groups on diverse topics. Table 1 gives statistics on some of the listserv groups used in the experiments. Each group was split randomly: 75% documents of all postings for training, 25% for test.

The same representations were used with all data sets, and are listed in Figure 4. The representations were produced by first running the perl module *Lingua:EN:Tag*[4] on the text. This broke the text into tokens and (imperfectly) assigned each token a syntactic part-of-speech tag based on a statistical model of English text. The sequence of tokens was then postprocessed in a variety of ways. After postprocessing, each of the unique types of token remaining became a predictor feature. Feature set sizes ranged from 10 to 133,717 features.

The forms of postprocessing are indicated in the name of each representation:

[2] *http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm*
[3] *http://about.reuters.com/researchandstandards/corpus/*
[4] *http://search.cpan.org/dist/Lingua-EN-Tagger/Tagger.pm*

- *noname*: tokens appearing on a list of common first and last names were discarded before any other processing.
- *Dpref*: only the first D characters of each word were used.
- *Dsuff*: only the last D characters of each word were used.
- *~POS*: some portion of each word, concatenated with its part-of-speech tag was used.
- *DgramPOS*: all consecutive sequences of D part-of-speech tags are used.
- *BOW*: all and only the word portion was used (BOW = "bag of words"). There are also two special subsets defined of *BOW*. *ArgamonFW* is a set of function words used in a previous author identification study [26]. The set *brians* is a set of words automatically extracted from a web page about common errors in English usage[5].

Finally, *CorneyAll* is based on a large set of stylometric characteristics of text from the authorship attribution literature gathered and used by Corney [9]. It includes features derived from word and character distributions, and frequencies of function words, as listed in *ArgamonFW*.

4.2 Results

We used Bayesian multinomial logistic regression with Laplace prior to build classifiers on several data sets with different representations. The performance of these classifiers on the test sets is presented in Figure 4.

One can see that error rates vary widely between data sets and representations; however the lines that correspond to representations do not have very many crossings between them. If we were to order all representations by the error rate produced by the model for each data set, the order will be fairy stable across different data sets. This is even more evident from Figure 5, which shows ranks instead of actual error rates. For instance, representation with all words ("bag-of-words", denoted BOW in the chart) almost always results in the lowest error rate, while pairs of consecutive part of speech tags (2gramPOS in the chart) always produces one of the highest error rates. There are some more crossings between representation lines near the right-most column that reflects RCV1, hinting that this data set is essentially different from all listserv groups. Indeed, RCV1 stories are produced by professional writers in the corporate environment, while the postings in the discussion groups are written by people in an uncontrolled environment on topic of their interest.

5 Topic independence in author identification

Topics correlate with authors in many available text corpora for very natural reasons. These days text categorization by topics is a well developed technology, so we have to look into the role that topics play in author identification and see if we confuse one for the other knowingly or unknowingly. Some researchers consciously use topics to help identify authors, which makes perfect sense when dealing

[5] http://www.wsu.edu/~brians/errors/errors.html

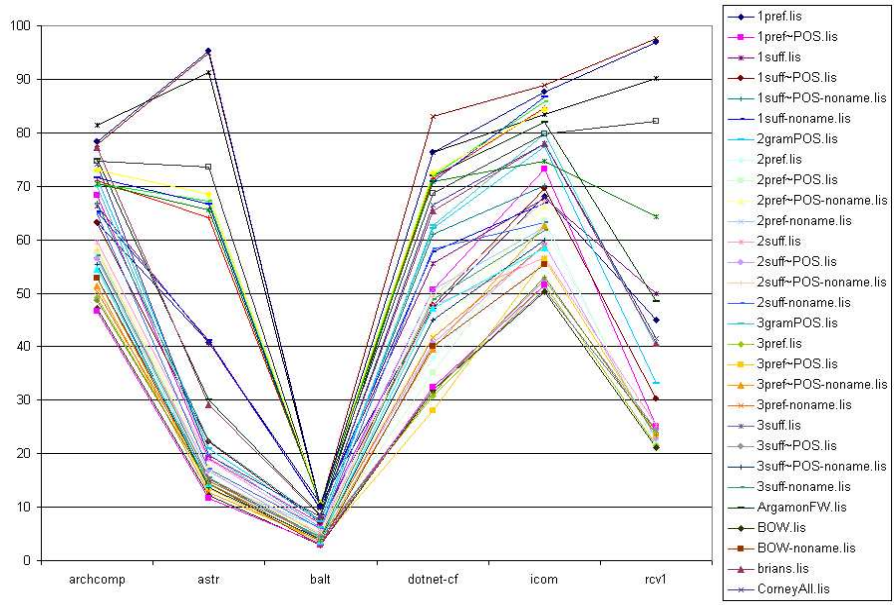

Figure 4: Test set error rates on different data sets with different representations.

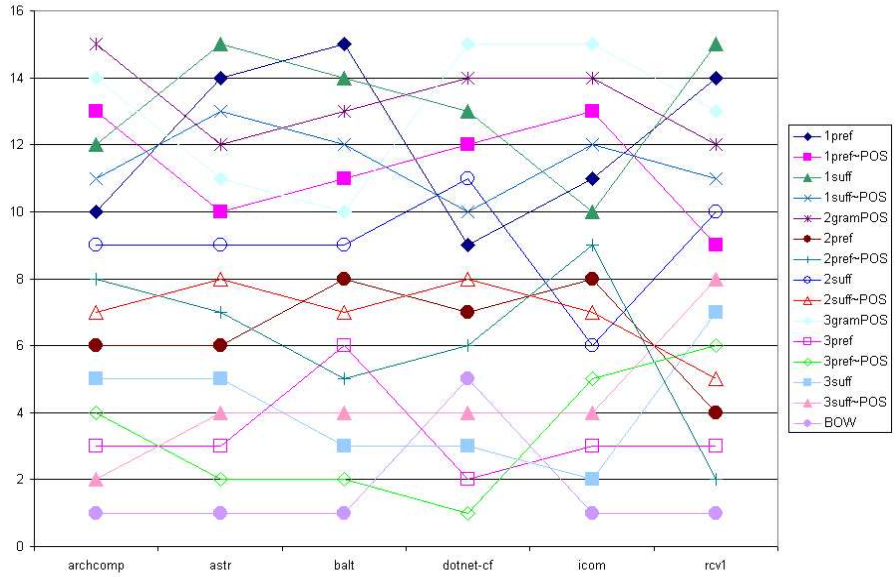

Figure 5: Ranks of test set error rates on different data sets with different representations.

with research articles, see for example [47], [22]. However, forensic, intelligence or homeland security applications seek to identify authors regardless of topic.

Traditionally, researchers used representations like function words that they assumed to be topic-independent. Whether function words really are topic-independent is questionable. A number of other representations may be subject to same concern.

Experimental evidence is needed to determine whether a particular representation is indeed topic-independent. Cross-topic experiments,i.e. experiments on a corpus of documents written on diverse topics by the same authors, are one promising approach to addressing this issue.

It is hard to collect a cross-topic corpus, so we performed a small-scale experiment, which, however, we believe to be illustrative. In the Listserv collection there are few authors that have posted on essentially different topics. We selected two of them who have made considerable number of postings, see Table 2. The postings from the group GUNDOG-L where both authors participated were used for training; the postings from two other groups with radically different topics were used for testing. Results are presented in Figure 6 through side by side comparison with earlier results on different representations. Obviously, there are much more line crossings approaching the right-most column, which reflects our cross-topic experiment. That of course means that the order of representations by produced error rate is radically different. In particular, the "bag of words" representation (BOW on the chart), which is known to be good for content-based categorization, performs poorly in this experiment. In contrast, a representation based on pairs of consecutive part of speech tags (2gramPOS in the chart) becomes one of the best.

Author	GUNDOG-L	BGRASS-L bluegrass music	IN-BIRD-L birds of Indiana
drxxx@aol.com	10	24	
bbxxx@inetdirect.net	6		19

Table 2: Two authors from Listserv for cross-topic experiment: number of postings per group.

6 "Odd-man-out" experiments

Given a list of authors, the "odd-man-out" task is to determine whether a particular document was written by one of these authors, or by someone else. Let us assume that there is a training set of documents available, where each document was written by one of the target authors, and that there is at least one document written by each of those authors.

It also seems natural to assume there are other documents available that do not belong to any of the target authors. We are going to use the authors of these "other" documents as "decoys" for training our classifier. Of course, it's better if these documents have much in common with available documents from the target authors: same genre, close creation date, etc. For the purpose of experimental work, all the documents will be taken from the same corpus.

The idea is to construct binary classifiers to discriminate between the target authors' documents and, ideally, documents from any other author. In our experiments, we are going to pool together some documents from the target authors as positive training examples, documents from the "decoy" authors as negative training examples; other documents from target authors and documents from other authors (not target and not decoy) will form the test sample.

We used the subset of RCV1 data set with 114 authors and train/test split as described earlier. The documents were represented using function words fre-

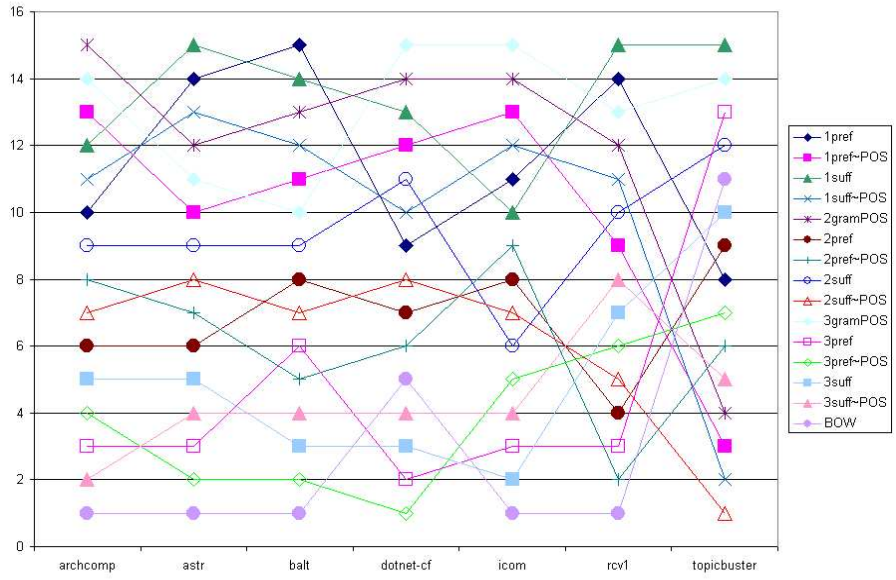

Figure 6: Ranks of test set error rates with different representations for the cross-topic experiment (right-most column) compared to ranks from Figure 5.

quencies. Let K denote the number of target authors; L - number of "decoy" authors and M - number of the rest, test authors. Table 3 shows the results of experiments for different combinations of K,L and M. For each combination, 10 random splits of 114 authors into those three categories were performed and the results averaged. We used our Bayesian logistic regression software (BBR, http://www.stat.rutgers.edu/~madigan/BBR/), essentially a binary specialization of the Bayesian multinomial logistic regression we described above.

In these experiments the multiclass nature of data was completely ignored; all documents for target authors, as well as for "decoy" and test authors, were pooled together. It's interesting to find out if it is possible to improve the results using information about individual authors. The approach we are using here is inspired by the works of Pereversev-Orlov [39]. Consider the set of documents for training as before, i.e. from target and "decoy" authors. We are going to train a multinomial logistic model with $K+L$ classes, regardless of those authors being target or "decoy". Having built this model, for any document $\boldsymbol{x}$ we can compute $K+L$ values of linear scores from that model: $\boldsymbol{\beta}_k^T \boldsymbol{x}, k = 1, ..., K+L$. Higher score value would mean that the document is closer to a particular class (i.e. author) in the view of the model at hand. The intuition behind is that multinomial model would produce feature combinations generally useful for discriminating between authors and capture this in scoring functions.

We now proceed with binary classification as before; the only difference is that, instead of function words or whatever other representation, we are going to use the vector of $K+L$ scores from the multinomial model as document representation. Figure 7 compares error rates produced by both approaches for the same set of K, L, M combinations as above. Obviously, the approach with multinomial model scores produces lower error rates in most cases.

K	L	M	error rate %%
10	30	74	39.02
10	40	64	45.68
10	50	54	24.56
10	60	44	37.14
20	10	84	55.64
20	20	74	41.31
20	30	64	49.60
20	40	54	49.07
20	50	44	34.33
30	10	74	51.72
30	20	64	54.52
30	30	54	48.37
30	40	44	49.99
30	50	34	50.41
40	10	64	53.75
40	20	54	52.41
40	30	44	50.89
50	10	54	50.59
50	10	44	45.09

Table 3: "Odd-man-out" experiments with binary classification: error rates for different combinations of K, L, M values, averaged over 10 random splits of 114 authors into these three categories.

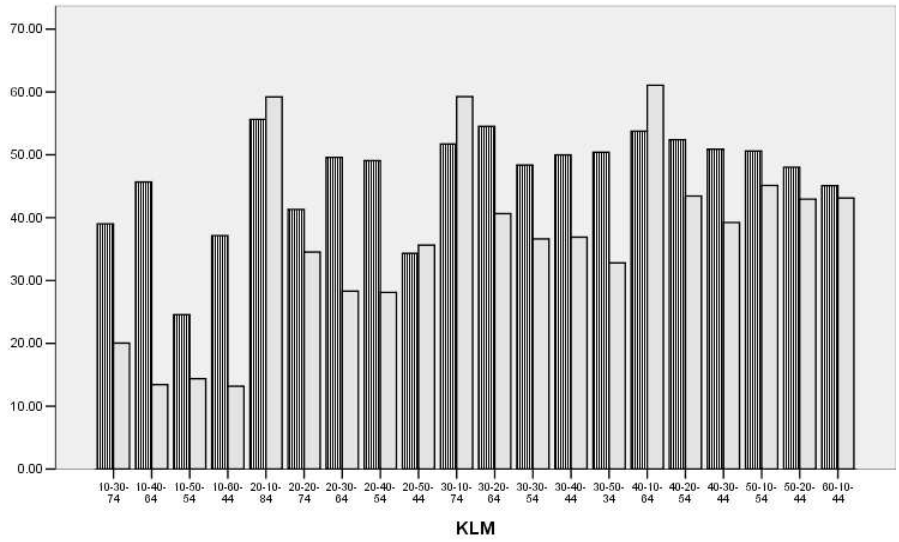

Figure 7: "Odd-man-out" experiments: comparing error rates from Table 3 (dark bars) with those produced by multinomial model scores approach (light bars).

7 Revisiting the Federalist Papers

During 1787-1788 seventy-seven articles were published anonymously in four of New York's five newspapers by Alexander Hamilton, John Jay, and James Madison to

persuade the citizens of the State of New York to ratify the Constitution. These papers together with an additional eight essays that had not previously been published were called the Federalist papers. These articles appeared under the pseudonym Publius and, as it happens, were unsuccessful: 56% of the citizens of New York state voted against ratifying the constitution.

Historians did a lot of research on the identity of Publius at the time. It was believed that General Alexander Hamilton had written most of the articles. Jay wrote five and these were identified. Hamilton died in a duel with Aaron Burr in 1804, and in 1807 a Philadelphia periodical received a list, said to have been made by Hamilton just before his fatal duel, assigning specific papers to specific authors. But in 1818, Madison claimed to have written numbers 49-58 as well as 62 and 63 which had been ascribed to Hamilton himself in his list. Thus twelve of the eight-five papers were claimed by both Hamilton and Madison. These papers were called disputed papers. An additional three No 18,19,20 are usually referred to as "Hamilton and Madison" since Hamilton said they were joint papers.

Many previous statistical studies have attempted to attribute the disputed Federalist papers and most assign all the disputed papers to Madison. Mosteller and Wallace (1962) used a function word representation and a naive Bayes classifier. They concluded: "Madison is the principal author. These data make it possible to say far more than ever before that the odds are enormously high that Madison wrote the 12 disputed papers."

Traditionally, most of the statistical analyses are based on a small numbers of features. Table 4 lists the features sets we used in this analysis.

Features	Name in Short
The length of each word	charcount
Part of speeches	POS
Two-letter-suffix	Suffix2
Three-letter-suffix	Suffix3
Words, numbers, signs, punctuations	Words
The length of each word plus part of speech tags	Charcount+POS
Two-letter-suffix plus part of speech tags	Suffix2+POS
Three-letter-suffix plus part of speech tags	Suffix3+POS
Words, numbers, signs, punctuations plus part of speech tags	Words+POS
484 function words from Koppel et al's paper	484 features
Mosteller and Wallace function words	Wallace features
Words appear at least twice	Words(¿=2)
Every word in the Federalist papers	Each word

Table 4: Feature sets for the Federalist analysis.

Word lengths vary from 1 to 20. The suffix2 features are features like ly, ed, ng, and there are 276 of them. The suffix3 features are features like ble, ing, ure, and there are 1051 of them. The word features include each word and numbers and signs like # $ % and punctuations like ;,". The 484 features are given by Koppel et al. There are three feature sets in Mosteller and Wallace paper, we choose the third one which has 165 features. The part of speech feature set includes 44 features.

One way to assess the usefulness of a representation is to examine preditive performance. Table 5 below shows error rate estimates for the different representations as assessed by ten-fold cross-validation on the 65 undisputed (i.e., labeled) papers and using the BBR software.

Features	Error Rate
charcount	0.216
POS	0.189
Suffix2	0.117
Suffix3	0.086
Words	0.099
Charcount+POS	0.120
Suffix2+POS	0.078
Suffix3+POS	0.041
Words+POS	0.083
484 features	0.047
Wallace features	0.047
Words($\geq$2)	0.047
Each word	0.051

Table 5: The results of the error rates on the training data set for each feature set.

We can see that the feature set suffix3 plus POS has the lowest error rate but several other representations provide similar performance.

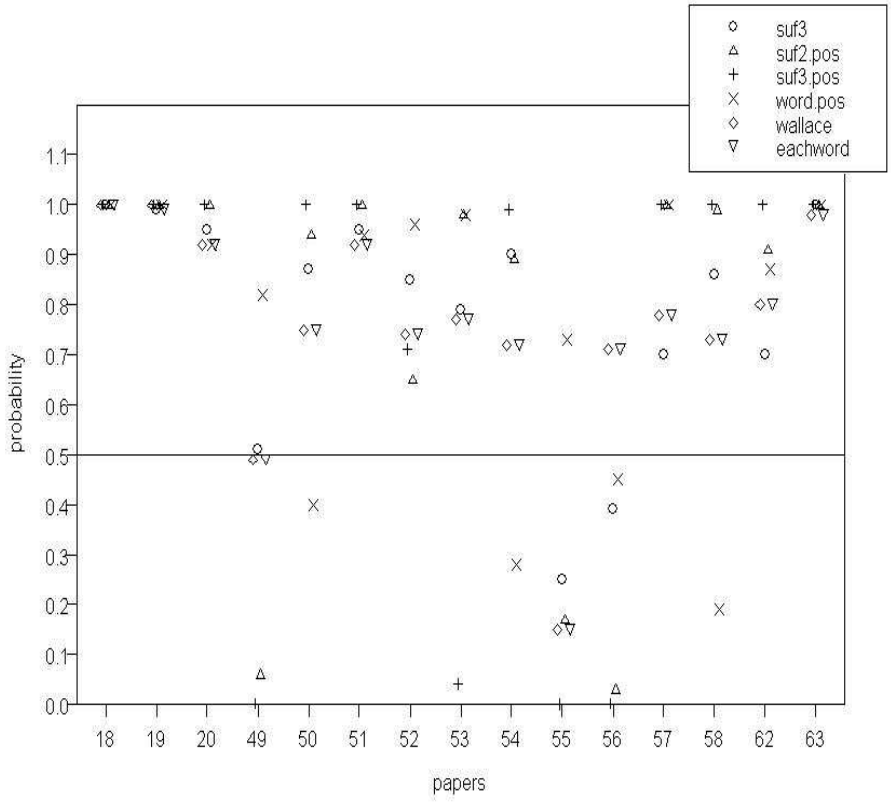

Figure 8: predicted probability of Madison for each of the disputed papers for six of the representations.

Figure 8 shows the predicted probability of Madison for each of the disputed papers for six of the representations. For four of the papers (18, 19, 20, and 63) the probability of Madison is close to one for all representations. For all the other papers, however, the predicted probability depends on the representation. For three of the papers (49, 55, and 56), Suffix3+POS, the representation that provided the best predictive performance on the training examples, actually assigns zero probability to Madison! The confidence Mosteller and Wallace placed in their findings seems inappropriate. We speculate that many of published attribution studies may suffer from similar over-confidence.

We note that Collins et al. using 18 "representational effects" as the features claimed No. 49, 55, 57, 58 were written by Hamilton. The Madison scores for No. 53 and No. 56 are also very low in their paper.

8 Conclusion

Our initial experiments suggest that sparse Bayesian logistic regression coupled with high-dimensional document representations shows considerable promise as a tool for authorship attribution. However, significant challenges concerning representation remain; different document representations can lead to different attributions and no clear method exists for accounting for this uncertainty.

9 Appendix

Here we are giving the formula for the function $Q(\beta_{kj}, \Delta_{kj})$, defined in Section 3.2, as the least upper bound for the second partial derivative of the negated loglikelihood (2) in the Δ_{kj}-vicinity of β_{kj}, where $\Delta_{kj} > 0$:

$$Q(\beta_{kj}, \Delta_{kj}) = \sum_i x_{ij}^2 / \left(F(\mathbf{B}, \mathbf{x_i}, \Delta_{kj}) + 2\right).$$

To define F we need some auxiliary notation:

$$r_{ik} = \boldsymbol{\beta}_k^T \mathbf{x_i}$$
$$E_{ik} = \left(\sum_{k'} \exp(\boldsymbol{\beta}_{k'}^T \mathbf{x_i})\right) - \exp(r_{ik})$$

Finally:

$$F(\mathbf{B}, \mathbf{x_i}, \delta) = \begin{cases} \exp(r_{ik} - \delta)/E_{ik} + E_{ik}/\exp(r_{ik} - \delta), & \text{if } E_{ik} < \exp(r_{ik} - \delta) \\ 2, & \text{if } \exp(r_{ik} - \delta) \le E_{ik} \le \exp(r_{ik} + \delta) \\ \exp(r_{ik} + \delta)/E_{ik} + E_{ik}/\exp(r_{ik} + \delta), & \text{if } \exp(r_{ik} + \delta) < E_{ik}. \end{cases}$$

The inference is straightforward and omitted here for the lack of space.

References

[1] S. Argamon, M. Koppel, and G. Avneri. Routing documents according to style. In *Proc. Int'l Workshop on Innovative Internet Information Systems*, Pisa, Italy, 1998.

[2] S. Argamon, M. Koppel, J. Fine, and A. R. Shimony. Gender, genre, and writing style in formal written texts. *Text*, 23(3), 2003.

[3] S. Argamon, M. Šarić, and S. S. Stein. Style mining of electronic messages for multiple author discrimination. In *Proc. ACM Conference on Knowledge Discovery and Data Mining*, 2003.

[4] S. Argamon-Engelson, M. Koppel, and G. Avneri. Style-based text categorization: What newspaper am i reading? In *Proc. AAAI Workshop on Learning for Text Categorization*, pages 1–4, 1998.

[5] H. Baayen, H. van Halteren, and F. Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131, 1996.

[6] D. Biber. *Variations Across Speech and Writing*. Cambridge University Press, 1988.

[7] D. Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(9):197200, 1992.

[8] J. Burrows. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Clarendon Press, Oxford, 1987.

[9] M. Corney. Analysing e-mail text authorship for forensic purposes. master of information technology (research) thesis, 2003.

[10] M. Corney, A. Anderson, G. Mohay, and O. de Vel. Identifying the authors of suspect e-mail. *Computers and Security*, 2001.

[11] N. Cristianini and J. Shawe-Taylor. *An Introduction To Support Vector Machines*. Cambridge University Press, 2000.

[12] O. de Vel, M. Corney, A. Anderson, and G.Mohay. Language and gender author cohort analysis of e-mail for computer forensics. In *Proc. Digital Forensic Research Workshop*, Syracuse, NY, August 2002.

[13] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 2000.

[14] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, New York, 1973.

[15] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407499, 2004.

[16] M. A. T. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159, 2003.

[17] A. Finn and N. Kushmerick. Learning to classify documents according to genre. In S. Argamon, editor, *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[18] R. S. Forsyth and D. I. Holmes. Feature finding for text classification. *Literary and Linguistic Computing*, 11(4):163–174, 1996.

[19] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization., 2004.

[20] F. Girosi. An equivlance between sparse approximation and support vector machines. *Neural Computation*, 10:1445–1480, 1998.

[21] J. Goodman. Exponential priors for maximum entropy models. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 305–312, 2004.

[22] S. Hill and F. Provost. The myth of the double-blind review? author identification using only citations. *SIGKDD Explorations*, 5(2):179–184, 2003.

[23] J. Karlgren. *Stylistic Experiments for Information Retrieval*. PhD thesis, SICS, 2000.

[24] D. Khmelev. Disputed authorship resolution using relative entropy for markov chain of letters in a text. In R. Baayen, editor, *4th Conference Int. Quantitative Linguistics Association*, Prague, 2000.

[25] B. Kjell and O. Frieder. Visualization of literary style. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 656–661, Chicago, 1992.

[26] M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 2003.

[27] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico*, 2003.

[28] B. Krishnapuram, A. J. Hartemink, L. Carin, and M. A. T. Figueiredo. Sparse multinomial logistic regression: Fast algorithms and generalized bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):957–968, 2005.

[29] F. Li and Y. Yang. A loss function analysis for classification methods in text categorization. In *The Twentith International Conference on Machine Learning (ICML'03)*, pages 472–479, 2003.

[30] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285318, 1988.

[31] D. Lowe and R. Matthews. Shakespeare vs Fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities*, pages 449–461, 1995.

[32] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002).*, pages 49–55, 2002.

[33] D. Mannion1 and P. Dixon. Authorship attribution: the case of oliver goldsmith. *Journal of the Royal Statistical Society (Series D): The Statistician*, 46(1):1–18, 1997.

[34] C. Mascol. Curves of pauline and pseudo-pauline style i. *Unitarian Review*, 30:452460, 1888.

[35] C. Mascol. Curves of pauline and pseudo-pauline style ii. *Unitarian Review*, 30:539546, 1888.

[36] T. Mendenhall. The characteristic curves of composition. *Science*, 214:237249, 1887.

[37] F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. Series in behavioral science: Quantitative methods edition. Addison-Wesley, Massachusetts, 1964.

[38] J. Pennebaker, M. R. Mehl, and K. Niederhoffer. Psychological aspects of natural language use:our words,our selves. *Annual Review of Psychology*, 54:547–577, 2003.

[39] V. S. Pereversev-Orlov. *Models and methods of automatic reading*. Nauka, Moscow, 1976.

[40] R. M. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[41] C. B. S. Hettich and C. Merz. UCI repository of machine learning databases, 1998.

[42] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 2002.

[43] F. Sha and F. Pereira. Shallow parsing with conditional random fields, 2003.

[44] S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19:2246–2253, 2003.

[45] F. Smadja. Lexical co-occurrence: The missing link. *Journal of the Association for Literary and Linguistic Computing*, 4(3), 1989.

[46] E. Stamatatos, G. Kokkinakis, and N. Fakotakis. Automatic text categorization in terms of genre and author. *Comput. Linguist.*, 26(4):471–495, 2000.

[47] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD Conference*, August 2004.

[48] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[49] M. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, June 2001.

[50] F. Tweedie, S. Singh, and D. Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30(1):1–10, 1996.

[51] J. Zhang and Y. Yang. Robustness of regularized linear classification methods in text categorization. In *Proceedings of SIGIR 2003: The Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 190–197, 2003.

[52] T. Zhang and F. Oles. Text categorization based on regularized linear classifiers. *Information Retrieval*, 4(1):5–31, April 2001.