The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

A unified approach to authorship attribution and verification

Xavier Puig, Martí Font & Josep Ginebra

To cite this article: Xavier Puig, Martí Font & Josep Ginebra (2016): A unified approach to authorship attribution and verification, The American Statistician, DOI: 10.1080/00031305.2016.1148630

To link to this article: http://dx.doi.org/10.1080/00031305.2016.1148630

Accepted author version posted online: 12 Feb 2016.

A unified approach to authorship attribution and verification

Xavier Puig, Martí Font, Josep Ginebra [1]

In authorship attribution one assigns texts from an unknown author to either one of two or more candidate authors by comparing the disputed texts with texts known to have been written by the candidate authors. In authorship verification one decides whether a text or a set of texts could have been written by a given author. These two problems are usually treated separately. By assuming an open-set classification framework for the attribution problem, contemplating the possibility that none of the candidate authors is the unknown author, the verification problem becomes a special case of attribution problem. Here both problems are posed as a formal Bayesian multinomial model selection problem and are given a closed form solution, tailored for categorical data, naturally incorporating text length and dependence in the analysis, and coping well with settings with a small number of training texts. The approach to authorship verification is illustrated by exploring whether a court ruling sentence could have been written by the judge that signs it, and the approach to authorship attribution is illustrated by revisiting the authorship attribution of the Federalist papers and through a small simulation study.

KEY WORDS: Stylometry, Model selection, Bayesian methods, Multinomial distribution.

[1]Departament of Statistics and O.R., Technical University of Catalonia, A Avgda. Diagonal 647, 08028 Barcelona, Spain. e-mail: xavier.puig@upc.edu.

1 Introduction

The statistical analysis of literary style has long been used to characterize the style of texts and authors, and to help settle authorship attribution problems. Early work (see, e.g., Mendelhall, 1887, or Yule, 1938) used word length, sentence length and the frequency of use of words to characterize literary style. Early applications involved the study of literary, religious or legal texts, but recently many new challenging problems have appeared due to widespread availability of electronic texts leading, for example, to new applications in homeland security, computer forensics or spam detection. The range of statistical methods used in this setting is wide, but they most often involve various approaches to classification.

In the analysis of the heterogeneity of the style in a given text or set of texts, one does not always know how many authors might have contributed to the text, and one typically does not have a reference set of candidate authors and training texts. In these settings one needs to resort to cluster analysis, also recognized as unsupervised classification/learning. A Bayesian approach to the analysis of the heterogeneity of style using mixtures of multinomial models is presented in Giron et al (2005).

Instead, in authorship attribution problems one has a set of S candidate authors and a set of texts known to have been written by each one of them. With the help of these training texts, one needs to assign texts by an unknown author to an author in the set, using discriminant analysis, also recognized as supervised classification/learning.

In most of the authorship attribution applications one adapts a closed-set classification framework, assuming that one knows with certainty that the unknown author is among the S candidates. Instead, nothing is lost by adopting a more prudent and flexible open-set classification framework also contemplating the possibility that the unknown author is not in the list. By adopting this

open-set framework, the authorship verification problem that requires to decide whether a text of unknown author has been written by a known author with comparable texts, becomes a special case of authorship attribution with $S = 1$.

A wide variety of statistical tools have been used to tackle authorship attribution and verification problems. Even though Mosteller and Wallace (1964, 1984) already used probability models to drive to the solution of an authorship attribution problem, most of that literature resorts to ad-hoc heuristic classifiers using linear or quadratic discriminant analysis (Stamatatos et al, 2000, Tambouratzis et al, 2004), support vector machines (Joachims, 1998, Diederich et al, 2003, Li et al, 2006), decision trees (Zheng et al, 2006), neural networks (Matthews and Merriam, 1993, Merriam and Matthews, 1994, Tweedie et al, 1996) or other machine learning based feature selection algorithms (Forsyth and Holmes, 1996, Forman, 2003, Binongo, 2003, Koppel et al, 2006). Recent applications of these supervised classification tools in authorship problems can be found, for example, in Stamatatos et al (2001), Holmes et al (2001), Burrows (2002, 2007), Hoover (2001, 2004), Abbasi and Chen (2005), Chaski (2005), Grant (2007), Argamon (2008), or Holmes and Crofts (2010).

Good reviews can be found in Holmes (1985, 1994, 1998, 1999), Stamatatatos (2009) and in Sebastiani (2002), and recent comparisons of some of these classification approaches in authorship attribution can be found in Zhao and Zobel (2005), Juola et al (2006), Yu (2008), Jockers et al (2008), Jockers and Witten (2010)

One shortcoming of most of these algorithmic based approaches is that they implicitly assume data to be continuous, or at least are tuned to work best with continuous data. But the data in authorship attribution problems are mostly categorical, and one should adapt to the specificities of that kind of data. In particular, one needs to adequately take into account the length of texts and to accommodate for the dependence between the counts of different categories of a given stylometric

characteristic, which is not easy to do with most of the classifiers used in authorship attribution.

Other shortcomings of the algorithmic approaches from machine learning is that they are tailored to work with large training samples, and hence do not fare well with a small number of training texts. Moreover, they can not be used in open-set classification frameworks.

In this paper we address the open-set authorship attribution problem using stylometric characteristics that involve counting features that are categorical, have a fixed number of categories, and are frequently observed. That leads to data being a contingency table with as many rows as texts under consideration, and it covers, for instance, counting word lengths, sentence lengths, letters, function words, nouns or adjectives. Our approach excludes the analysis of word frequency counts used in vocabulary richness analysis, because in that case the number of categories grows with text size.

We adopt a formal Bayesian model based approach, in the spirit of Mosteller and Wallace (1984). That approach assesses the uncertainty in the classification by assigning them either to one of the candidate authors or to none of them based on the posterior probabilities that the texts were written by each of the authors. Bayesian models are probability models, and one can check the assumptions on which the analysis is based, which is in stark contrast with algorithmic approaches that do not explicit the stochastic assumptions made.

To illustrate our approach, an authorship verification case study involving a court ruling sentence is presented, and the authorship attribution of the Federalist papers is revisited. A small simulation experiment is also carried out to help assess the performance of our Bayesian model driven approach under repeated use, and to compare it to three of the main alternative approaches available for authorship attribution.

2 Bayesian model building

2.1 Description of the model

In authorship attribution problems one starts with n^0 disputed texts that are assumed to have been written by the same unknown author, and with S potential authors for these texts. One also has n^s texts that are comparable to the disputed ones and are known to belong to the s-th candidate author, for $s = 1, \ldots, S$. In order for texts to be comparable, ideally they all should have been written at around the same time, belong to the same genre and deal with a similar topic, even though in practice that might be difficult to attain.

Given a stylometric characteristic that involves counting features that are categorical with a fixed number of categories, k, the i-th text of the unknown author will become a vector valued categorical observation, $y_i^0 = (y_{i1}^0, \ldots, y_{ik}^0)$, for $i = 1, \ldots, n^0$, where y_{ij}^0 is the number of counts of the j-th category in the i-th disputed text. Analogously, the i-th text known to be by the s-th author will yield the vector of counts $y_i^s = (y_{i1}^s, \ldots, y_{ik}^s)$, for $i = 1, \ldots, n^s$.

The frequency of frequent function words is one of the most reliable stylometric features of the kind considered here (see, e.g., Hoover, 2003, Zhao and Zobel, 2005, Uzuner and Katz, 2005, Grieve, 2007). Even though word length has rarely proven useful in the authorship attribution of English texts, it is useful in other languages (see, e.g., Giron et al, 2005). Table 1 presents two examples of this kind of data, with each row of the table corresponding to either a training or a disputed text, and playing the role of a y_i^s or a y_i^0 observation.

The set of all the n^0 vector valued observations corresponding to the n^0 disputed texts, denoted $y^0 = (y_1^0, \ldots, y_{n^0}^0)$, are assumed to be conditionally independent and multinomially distributed, $\prod_{i=1}^{n^0} \text{Mult}(y_i^0; N_i^0, \theta^0)$, where $N_i^0 = \sum_{j=1}^{k} y_{ij}^0$ is the total count for the i-th disputed text, and where

$\theta^0 = (\theta^0_1, \ldots, \theta^0_k)$ with θ^0_j being the probability of the j-th category for all the disputed texts, and hence with $\sum_{j=1}^{k} \theta^0_j = 1$. Analogously, the set of observations of the s-th author, $y^s = (y^s_1, \ldots, y^s_{n^s})$, are assumed to be $\prod_{i=1}^{n^s} \text{Mult}(y^s_i; N^s_i, \theta^s)$ distributed, with $N^s_i = \sum_{j=1}^{k} y^s_{ij}$ and $\theta^s = (\theta^s_1, \ldots, \theta^s_k)$, where $\sum_{j=1}^{k} \theta^s_j = 1$.

Under the assumption that all the n^0 disputed texts share the same multinomial parameter θ^0, it is possible to combine all the n^0 texts into a single text and work with the vector of aggregated counts; in that case, $y_0 = (\sum_{i=1}^{n^0} y^0_{i1}, \ldots, \sum_{i=1}^{n^0} y^0_{ik})$, is $\text{Mult}(y_0; N^0, \theta^0)$ distributed, where $N^0 = \sum_{i=1}^{n^0} N^0_i$ is the total count in texts by the disputed author. Analogously, if all the observations of the s-th author are indeed conditionally independent and multinomially distributed, and share the same θ^s, then $y_s = (\sum_{i=1}^{n^s} y^s_{i1}, \ldots, \sum_{i=1}^{n^s} y^s_{ik})$ follows a $\text{Mult}(y_s; N^s, \theta^s)$ distribution, with $N^s = \sum_{i=1}^{n^s} N^s_i$.

If the author of the disputed texts was the s-th candidate, one expects that the aggregated counts in the disputed texts, y_0, will be $\text{Mult}(y_0; N^0, \theta^0 = \theta^s)$ distributed. Furthermore, if the sample counts of all texts are conditionally independent, then the probability density function of the whole set of data, $y = (y_0, y_1, \ldots, y_S)$, will be:

$$p_s(y|\theta^1, \ldots, \theta^S) = \text{Mult}(y_0; N^0, \theta^s)\text{Mult}(y_s; N^s, \theta^s) \prod_{r=1, r\neq s}^{S} \text{Mult}(y_r; N^r, \theta^r), \tag{2.1}$$

which will be recognized from now on as the M_s model.

In most authorship attribution studies one adopts a closed-set classification framework, where one acts as if one had the certainty that the unknown author was one of the S candidates. In that case, one would only consider the $M_1, \ldots, M_S$ models. Instead, we adopt an open-set classification framework, contemplating the possibility that disputed texts might not be written by any author in

the list. That is done by considering an extra $(S+1)$-th sub-model, M_0, with pdf:

$$p_0(y|\theta^0,\theta^1,\ldots,\theta^S) = \text{Mult}(y_0;N^0,\theta^0)\prod_{s=1}^{S}\text{Mult}(y_s;N^s,\theta^s). \tag{2.2}$$

The $S=1$ case corresponds to the authorship verification problem.

As prior distribution for the multinomial probabilities, θ^r, for $r=0,1,\ldots,S$, it will be assumed that they are independent and Dirichlet$(a_1^r,\ldots,a_k^r)$ distributed, where $a^r=(a_1^r,\ldots,a_k^r)$ is such that $a_j^r>0$. Depending on the values chosen for a^r, the prior will capture a different type and amount of information. In particular, the expected value of θ^r will be $(a_1^r,\ldots,a_k^r)/(\sum_{j=1}^k a_j^r)$, and one can choose the a_j^r to reflect the fact that some categories might be known to appear with larger probabilities than others. Also, the larger $\sum_{j=1}^k a_j^r$ the smaller the variances of θ_j^r and the more informative the prior chosen for θ^r.

The Dirichlet prior is convenient, because it leads to closed form expressions for the posterior probabilities of each one of the $S+1$ sub-models. In the examples that follow all the $a^r=(a_1^r,\ldots,a_k^r)$ are set to be equal to $(1,\ldots,1)$, which corresponds to assuming a uniform distribution on the simplex for θ^r. The amount of information in this prior is equivalent to the one in a sample text with a count total of $N=k$. Given that the total number of words in texts will be much larger than k, the influence of the uniform prior on the posterior distribution will be a lot weaker than the influence of the data through the likelihood function. As a consequence, varying the parameters of the prior distribution around the chosen $(1,\ldots,1)$ does not alter the conclusions of the analysis.

It is also assumed that all $S+1$ sub-models are equally likely a priori, and hence that their prior probabilities are $P(M_r)=1/(S+1)$, but that can be trivially set to be otherwise.

2.2 Author selection through model selection

A difficulty of the heuristic algorithms for classification is that they often lack a statistically well grounded method for selecting an author for the disputed texts. Here that selection is tackled first through a formal model selection method, based on the posterior probability that each one of the models considered could be the one generating the data. Model checks will also help support the choice of model, and hence of author.

The posterior probability that the M_r model is the one generating the data is:

$$P(M_r|y) = \frac{P(M_r)P(y|M_r)}{\sum_{r=0}^{S} P(M_r)P(y|M_r)}, \text{ for } r = 0, 1, \ldots, S, \tag{2.3}$$

where $P(M_r)$ is the prior probability of model r and where $P(y|M_r)$ is the density function of the prior predictive distribution under model M_r evaluated at the observed data, also recognized as the marginal likelihood of M_r. Hence, the posterior probability of M_r is proportional to $P(M_r)$ and $P(y|M_r)$. One will select the model with the largest posterior probability, and when each model is considered equally likely a priori, that means picking the M_r with the largest marginal likelihood, $P(y|M_r)$.

Often, computing $P(y|M_r)$ exactly is too complicated to be attempted in practice, and one approximates its logarithm through the BIC, or through the MCMC simulations used to update the model. But in our case, by choosing a Dirichlet prior one has a closed form expressions for $P(y|M_r)$, that can be easily evaluated. In particular, when $y = (y_0, y_1, \ldots, y_S)$ one has that:

$$p(y|M_0) = \text{Dir-Mult}(y_0; N^0, a^0) \prod_{s=1}^{S} \text{Dir-Mult}(y_s; N^s, a^s), \tag{2.4}$$

where Dir-Mult$(x; N, a)$ denotes the pdf of a Dirichlet-Multinomial distribution with parameters N

and $a = (a_1, \ldots, a_k)$ evaluated at $x = (x_1, \ldots, x_k)$,

$$\text{Dir-Mult}(x; N, a) = \frac{N!\Gamma(\sum_{j=1}^{k} a_j)}{\Gamma(N + \sum_{j=1}^{k} a_j)} \prod_{j=1}^{k} \frac{\Gamma(x_j + a_j)}{x_j!\Gamma(a_j)}. \tag{2.5}$$

The marginal likelihood under M_r for $r \in \{1, \ldots, S\}$ becomes:

$$p(y|M_r) = \frac{N^0!N^r!}{(N^0 + N^r)!} \frac{\prod_{j=1}^{k}(\sum_{i=1}^{n^0} y_{ij}^0 + \sum_{i=1}^{n^r} y_{ij}^r)!}{\prod_{j=1}^{k}(\sum_{i=1}^{n^0} y_{ij}^0)! \prod_{j=1}^{k}(\sum_{i=1}^{n^r} y_{ij}^r)!} \times \tag{2.6}$$

$$\text{Dir-Mult}(y_0 + y_r; N^0 + N^r, a^r) \prod_{s=1, s\neq r}^{S} \text{Dir-Mult}(y_s; N^s, a^s). \tag{2.7}$$

In this way, one can compute $P(y|M_r)$, and hence $P(M_r|y)$, exactly.

Note that here one is computing the exact posterior probabilities, $P(M_r|y)$, conditional on both the training as well as the disputed texts, $y = (y_0, y_1, \ldots, y_S)$. That is different from taking an approximate two-stage approach, first "estimating" the posterior distribution of the multinomial probabilities θ^r of the r-th author based only on the counts in the training texts by that author, y_r, and using (2.3) with $y = y_0$ after replacing $P(y = y_0|M_r)$ by $P(y = y_0|\hat{\theta}^r)$, where $\hat{\theta}^r$ is an estimate of θ^r based on its posterior distribution. This two-stage approach is used in Gale et al (1993), McCallum and Nigan (1998), Lewis (1998), Schneider (2003), or Peng et al (2004), but it can not be used in the open-set classification framework adopted here.

2.3 Model checking

Our solution to the authorship attribution and verification problems relies on the model comparison just described, which in turn relies on the assumption that the model considered is correct. Before standing by the conclusions reached, one should check whether that model does indeed capture all the relevant features in the data or not.

The main model assumption is that all the vectors with the counts of the texts by the same author, s, are conditionally independent and distributed as a Mult(N_i, θ^s), where θ^s is identical for all the texts by that author. Even though inference is made after aggregating all texts by the same author in a single text, to check that assumption one needs to resort back to the sample of n^s vectors of counts, $y_1^s, \ldots, y_{n^s}^s$, before aggregation. The two most likely deviations from that assumption, and the way to check them, are:

1. The style of one or several of the texts attributed to the s-th author might not be comparable to the style of the other texts by him, or might not even be by that author. In such a situation, some of the observation(s) assumed to be from the s-th author, y_i^s, for $i = 1, \ldots, n^s$, might be independent and multinomially distributed but with different and unrelated multinomial parameter values.

   To check whether all the n^s texts assumed to be comparable and by the same author are indeed so, we verify whether each one of them is by that author by treating the other $n^s - 1$ texts as a training set. That is, we would go author by author, and resort to the $S = 1$ special case of the model in Section 2.1.

2. The vectors of counts y_i^s, for $i = 1, \ldots, n^s$, corresponding to the training texts from the s-th author, might be multinomially distributed with similar but not identical values of θ_i^s. That leads to the count data from the s-th author being more dispersed than anticipated by (2.1) or (2.2). If these θ_i^s can be assumed to be exchangeable and follow a given distribution, one can switch from the purely multinomial models considered here to multinomial mixtures instead.

   Building a Bayesian model is like building a data simulation model. To check whether the vector of counts for the texts of a given author are identically distributed as a multinomial or not, we assess whether it is plausible that one could simulate data like the data observed through the predictive distributions under the updated model (see, e.g., Gelman et al, 2004). We do not report on the predictive checks carried out in the examples that follow, but we

found that the purely multinomial based models in Section 2.1 match closely the variability of the counts observed.

3 Authorship verification case study

Here, we compare the style of a Spanish patent court ruling sentence, denoted by *D*, with the style of four other patent court ruling sentences written at around the same time and dealing with similar issues, denoted by S_1, S_2, S_3 and S_4. All five sentences were signed by the same judge, but law experts conjecture that the disputed sentence was actually written by someone else. The goal is to examine whether the style of the disputed sentence is similar enough to the style of the other sentences to back the single authorship hypothesis.

The comparison is based both on word length distribution, as well as on the frequency with which the twenty most frequent function words are used in these sentences. Before counting the number of *l*-lettered words and the number of times function words appear in the sentences, we have excluded from the text all citations, acronyms, capital lettered words, numbers, dates and names of persons and of cities. On top of that, we have only considered the factual, the legal basis and the final verdict, excluding from the analysis the formal paragraphs that are always repeated at the end of all sentences.

The resulting data, used in the analysis, are partially presented in Table 1. The first row of the first sub-table for example indicates that in the disputed sentence, *D*, there are 598 one-lettered words, 4069 two-lettered words and so on, and a total of 13051 words. The remaining rows of that sub-table have the counts for the four training sentences.

Figure 1 compares the proportion of *l*-lettered words observed in the disputed sentence *D* with the proportion observed in S_1 to S_4. It indicates that the proportion of words of 3, 4, 7, 8 and more

than nine letters in D is the largest, and the proportion of words of 1, 5, 6 and 9 letters in D is the smallest of all five sentences considered. Figure 2 compares the frequency of appearance of the twenty most frequent words in D with the one in S_1 to S_4. Note that the frequency of appearance of *que, en, a, los, las* and *no* in D is the highest, and the frequency of *y, con, o* and *su* is the lowest among all five sentences considered.

To check whether the four sentences used as a training sample do indeed have a similar style, we compare each one of them with the other three training sentences, excluding D. The first four rows of Table 2 present the probability that the counts for S_i share multinomial probabilities with the counts obtained by adding up the ones of the three remaining training texts. These probabilities are all very close to one, which is consistent with the hypotheses that these four sentences were all written by the judge that signed them.

The word length and word count distributions of D are compared with the corresponding distributions of the training sentences by computing $P(M_1|y)$, the probability that the counts for D share the same multinomial probabilities as the sum of the counts for S_1, S_2, S_3 and S_4. According to the last row in Table 2, that probability is zero under both features, which indicates that the style of the disputed sentence is very different from the style of the training sentences. That is consistent with Figures 1 and 2, and it indicates that it is likely that the disputed sentence was actually written by someone other than the one signing it.

4 Authorship attribution case study

The federalist papers were published anonymously between 1787 and 1788 by Alexander Hamilton, John Jay, and James Madison to persuade New Yorkers to adopt a new constitution of the US. Of the 77 essays, having between 900 and 3500 words each, it is generally agreed that Jay wrote

five, Hamilton wrote 43, Madison wrote 14, and three papers are known to be the joint work of Madison and Hamilton. That leaves twelve papers, numbered 49 to 58, 62 and 63, that can not be clearly attributed to Hamilton or Madison.

Mosteller and Wallace (1964, 1984) carried extensive comparisons of the frequencies of a carefully chosen set of common words in writings known to be by Hamilton and by Madison, with the frequencies of these words in the twelve disputed papers. Recent studies re-visiting that problem are, for example, Holmes and Forsyth (1995), Martindale and McKenzie (1995), Tweedie et al. (1996), Bosch and Smith (1998), Khmelev and Tweedie (2001), Collins et al (2004), and Jockers and Witten (2010).

Our approach is Bayesian, as the one taken by Mosteller and Wallace, but it is different from their one in that we model the whole vector of counts jointly, using multinomial distributions, instead of modeling each count separately assuming that they were independent and Poisson or negative binomial distributed. A second difference is that we take the open-set classification approach described in Section 2, instead of a closed-set approach.

Mosteller and Wallace explore the use of word length as a way to help determine authorship, but conclude that this feature does not distinguish Hamilton and Madison styles. Our analysis confirmed that fact, and hence here we focus on word counts.

Different from what happens in authorship verification studies, having more than one candidate author allows one to pick a list of words that best discriminate among them. Mosteller and Wallace base their main analysis on the counts of the 30 frequent words assessed to discriminate best between the styles of Madison and of Hamilton based both on the federalist papers as well as on external texts known to have been written by them.

Besides carrying out our analysis based on the thirty words used by Mosteller and Wallace, we

have also carried out parallel analysis based on two new lists of words. The first list contains the 20 function words that are most frequent in the federalist papers, without taking into consideration their discriminating power. The second list consists of the 30 function words that we found to be most discriminant between the 43 federalist papers by Hamilton and the 14 by Madison, without using any external texts.

To select our list of 30 most discriminant words, we started with the list of 200 most frequent words in the papers by Hamilton and the 200 most frequent words in the papers by Madison. Merging these two lists, leads to a set of 240 words. To assess the discriminating power of these words, we modeled the 240-dimensional vector with the counts of these words in the papers by Hamilton, y^H, and the vector with the counts in the papers by Madison, y^M, as:

$$p(y^H, y^M|\theta^H, \theta^M) = Mult(y^H; N^H, \theta^H)Mult(y^M; N^M, \theta^M) \tag{4.1}$$

where θ^H and θ^M are the multinomial probabilities modeling the relative frequency of these words in the papers by Hamilton and by Madison, and where N^H and N^M are the sum of the counts of these words in these papers. As a prior distribution on θ^H and θ^M, one uses the same one as for θ^r in Section 2. Words are then ranked from having better to having worse discriminating power based on the statistic:

$$T_i = \left| \frac{E(\log \frac{\theta_i^H}{\theta_i^M}|y^H, y^M)}{\sqrt{Var(\log \frac{\theta_i^H}{\theta_i^M}|y^H, y^M)}} \right|, \qquad \text{for} \qquad i = 1, \ldots, 240. \tag{4.2}$$

The thirty words with the largest T_i, after discarding the ones that clearly depended on context, together with their T_i value, were: *on (10,73), would (8,16), upon (7,69), there (7,54), by (7,47), to (6,94), and (6,81), the (5,42), these (4,82), in (4,39), at (4,19), latter (4,16), several (3,96), I (3,8), if (3,69), might (3,62), any (3,51), kind (3,48), had (3,46), between (3,45), those (3,34), an*

(3,2), he (3,19), this (3,19), very (3,17), against (3,12), no (2,95), were (2,9), into (2,89) and *same (2,88)*. Only eight of these words, (*an, by, kind, on, there, this, to* and *upon*), appear also in the list of Mosteller and Wallace.

Figure 3 compares the frequencies of appearance of our list of 30 most discriminating words in the papers by Hamilton and by Madison, with the ones in the twelve disputed papers.

To check whether all the 43 papers used as a training sample of the style of Hamilton do indeed have a similar style, we verify whether the style of each one of these papers is similar to the one of the other 42. And we repeat a similar verification exercise on each one of the 14 papers used as training samples of Madison. In both cases, one classifies all 57 papers as belonging to the presumed author with probability close to one.

To settle the authorship attribution of the twelve disputed texts, we carried out the analysis described in Section 2 on each one of these papers separately, considering as tentative hypothesis that they had been authored by Hamilton, by Madison, or by an unknown someone else. The results, based on our set of thirty most discriminating words, appear in Table 3. They indicate that all disputed papers except 55 should be attributed to Madison. Figure 3 indicates what is it that makes the style of paper 55 different from the style of the rest of disputed papers, and closer to the one of Hamilton.

When we do the analysis based on the 30 most discriminating words of Mosteller and Wallace, the only difference is that the posterior probability that paper 55 follows Hamilton style is .06. When we base the analysis on the 20 most frequent function words instead, without filtering out words that do not discriminate, we find that all the disputed papers except 49 and 55 are again attributed to Madison with probability close to one. All these findings are in agreement with the ones in the other studies looking into this problem.

5 Simulation study

To assess the performance of the Bayesian multinomial model driven method, denoted here as BM, and to compare it to alternative supervised classification techniques, two simulation scenarios are designed. In the first one, word length data from five training texts by Author 1 and from five training texts by Author 2 are simulated, to be used to help settle the authorship of three disputed texts, D1, D2 and DU. In the second scenario, word length data from fifty texts by Author 1 and from fifty texts by Author 2 are simulated, to be used to settle the authorship of D1, D2 and DU. All texts are set to have $N = 500$ words.

The multinomial probabilities used to simulate the word length data by Author 1 are $\theta^1 = (.04, .17, .22,$

$.20, .14, .09, .06, .04, .02, .02), while the ones used for Author 2 are \theta^2 = (.035, .16, .23, .19, .15, .095, .065,$

$.045, .015, .015). The disputed text D1 is simulated to be by Author 1, with \theta^0 = \theta^1$, D2 is simulated to be by Author 2, with $\theta^0 = \theta^2$, and DU is simulated to neither be by Author 1 nor by Author 2, with multinomial probabilities $\theta^0 = (.07, .13, .17, .15, .13, .11, .09, .06, .05, .04, .07)$.

Under each of these two simulation scenarios, we first check how our BM method behaves under repeated use. Second, we compare the performance of BM with the performance of three popular supervised classification methods. In both cases, the assessment will be based on repeating the two simulation experiments 1000 times, each time simulating the word length data of all the training texts as well as the one of the three disputed texts.

To assess how the BM approach fares under repeated use, Figure 4 presents the histograms of the 1000 posterior probabilities of the three authorship hypotheses, (Author is 1, Author is 2, and Author is neither 1 nor 2 and hence unknown), for each one of the three disputed papers under the two simulation scenarios.

In the case of D1, known to be by Author 1, we find that in 733 (824) of the 1000 realizations for the 5 training texts (50 training texts) scenario the posterior probability that it is by Author 1 is the largest one, while in 267 (176) of these realizations the probability that it is by Author 2 is the largest one. In almost all these realizations, these posterior probabilities are far from 0 or 1, due to the styles of Authors 1 and 2 being similar, which makes the classification problem significantly more difficult than the ones faced in the previous case studies. In contrast, Figure 4 also indicates that all 1000 realizations lead to a posterior probability close to 0 that D1 is by an unknown author. Something similar is observed through the histograms of the posterior probabilities for D2.

Instead, the style of DU is purposely set to be very different from the styles of Authors 1 and 2, and therefore in most (but not in all) the 1000 realizations our BM method assigns a posterior probability close to 1 to the author being unknown. The scenario with 50 training texts per author is more conclusive than the one with 5.

Next, our BM method is compared to: i) a decision tree classification method, denoted DT, ii) a support vector machine method, denoted SVM, and iii) a logistic regression method, denoted LR. To do that, the three alternative methods together with the BM method proposed here are used to classify each one of the 1000 realizations of the D1, D2 and DU disputed texts based on each one of the corresponding 1000 realizations of the training texts. And that is done again under both simulation scenarios.

For a description on how these classification methods work, see Chapters 4, 8 and 9 of Gareth et al (2014). To implement the DT method, the tree() function from the tree library in R has been used, to implement the SVM method, the svm() function from the e1071 library has been used, and to implement the LR method, the glm() function has been used. The optimal level of model complexity under each one of these three approaches has been determined through cross validation.

By restricting consideration to texts that have 500 words, one avoids the need to decide how to incorporate text length in these three alternative analysis, which is an issue not adequately settled in authorship attribution practice. Note also that these alternative approaches are tailored to work with large training samples and hence with many training texts. In contrast, the BM approach naturally incorporates text size in the analysis, and it works well in instances with a few, or even a single, training text.

Table 4 presents the proportion of times each one of the three disputed texts is correctly attributed to the author that actually wrote it. These proportions are estimates of the long run (frequentist) probability that the method correctly classifies the disputed text to the actual author. The first row of that table, for example, indicates that the DT approach correctly classifies D1 to be by Author 1 in 639 out of the 1000 realizations, the SVM approach does that 588 times, and the LR approach does that 653 times, all compared to the 733 times that the BM approach correctly classifies D1. Different from the BM method, the three top-of-the-counter alternative supervised classification approaches do not allow for an open-set classification framework, because they can not handle the hypothesis that neither Author 1 nor Author 2 wrote a text. Hence, no proportion of correct classifications can be provided for DU under these alternative approaches.

Table 4 indicates that the BM method implemented with a uniform prior for the multinomial parameters performs better than the LR approach and that, in turn, the LR approach performs better than the DT and the SVM approaches. The performance of the three alternative methods is specially poor in the five training texts per author scenario.

When the text length and/or the number of training samples increase, the problem becomes easier, and we find the performance of the LR and the SVM methods to become closer to the one of the BM method. We have repeated this simulation exercise under many other scenarios and different classification methods, reaching similar conclusions.

6 Final Comments

Different from the algorithmic based classification methods typically used for authorship attribution, the BM approach advocated for here has the advantage of being tailored for categorical data, of naturally incorporating text size and dependence in the analysis, of handling settings with a small number of training texts, and of easily adapting to open-set classification contexts. On top of that, it also comes with the scientific advantage of making explicit the list of distributional assumptions made; by checking whether those assumptions are adequate, one checks the validity of the analysis.

Even though the main goal in authorship attribution is to classify the disputed texts by making inference about M_r, one also benefits from exploring the posterior distributions for $(\theta^0, \theta^1, \ldots, \theta^S)$, to learn about what distinguishes the style of authors.

7 Acknowledgments

This work was funded in part by Grant No. MTM2013-43992-R of the Ministerio de Ciencia e Inovación of Spain. We are very grateful for the comments made by the Editor, the Associate Editor, two referees and Marta Perez-casany; they helped us improve the manuscript a lot.

8 Bibliography

Abbasi, A. and Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20, 67-75.

Argamon, S. (2008). Interpreting Burrow's delta: geometric and probabilistic foundations. Literary and Linguistic Computing, 23, 131-147.

Binongo, J.N.G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16, 9-17.

Bosch, R.A. and Smith, J.A. (1998). Separating hyperplanes and the authorship of the disputed Federalist Papers. *American Mathematical Monthly*, 105, 601-608.

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267287.

Burrows, J.F. (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22, 27-47.

Chaski, C.E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4, 1-13.

Collins, J., Kaufer, D., Vlachos, P., Butler, B. and Ishizaki, S. (2004). Detecting collaborations in text : Comparing the authors rhetorical language choices in the Federalist Papers. *Computers and the Humanities*, 38, 15-36.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, 19, 109-123.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 1289-1305.

Forsyth, R. and Holmes, D. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, 11, 163-174.

Gale, W.A., Church, K.W., and Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26, 415-439.

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Gelman A, Carlin JC, Stern H, Rubin DB (2004). *Bayesian Data Analysis* (2nd ed). New York: Chapman and Hall.

Giron, J., Ginebra, J. and Riba, A. (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, 59, 19-30.

Grant, T.D. (2007). Quantifying evidence for forensic authorship analysis. *International Journal of Speech Language and the Law*, 14, 1-25.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, 22, 251-270.

Holmes, D.I. (1985). The analysis of literary style. A review, *Journal of the Royal Statistical Society, Ser A*, 148, 328–341.

Holmes, D.I. (1994). Authorship attribution. *Computers and the Humanities*, 28, 87-106.

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

Holmes, D.I. (1999). *Encyclopedia of Statistical Sciences; Update Vol.3*. pp. 721-727. New York: Wiley.

Holmes, D.I. and Forsyth, R. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10, 111-127.

Holmes, D.I., Gordon, L. and Wilson, C. (2001). A widow and her soldier: Stylometry and the american civil war. Literary and Linguistic Computing, 16, 403-420.

Holmes, D.I. and Crofts, D.W. (2010). The diary of a public man: a case study in traditional and non-traditional authorship attribution. *Literary and Linguistic Computing*, 25, 179-197,

Hoover, D.L. (2001). Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing*, 10, 111-127.

Hoover, D.L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, 18, 341-360.

Hoover, D.L. (2004). Testing Burrow's Delta. *Literary and Linguistic Computing*, 19, 453-475.

Joachims, T.T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceed.of the 10th European conference on machine learning*, pp. 137-142.

Jockers, M.L. and Witten, D.M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25, 215-223.

Jockers, M.L., Witten, D.M. and Criddle, C.S. (2008). Reassessing authorship in the book of Mormon using nearest Shrunken centroid classification. *Literary and Linguistic Computing*, 23, 465-491.

Juola, P., Sofko, J. and Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21, 169-178.

Khmelev, D.V. and Tweedie, F.J. (2001). Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16, 299-307.

Koppel, M., Akiva, N. and Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57, 1519-1525.

Lewis, D.D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *Proceed. of the 10th European Conference on Machine Learning*, pp. 4-15

Li, J., Zheng, R. and Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM*, 49, 76-82.

Martindale, C. and McKenzie, D. (1995). On the utility of content analysis in author attribution: The Federalist. *Computers and the Humanities*, 29, 259-270.

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry: A application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, 8, 203-209.

McCallum, A. and Nigan K. (1998). A comparison of event models for naive Bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*, Madison, Wisconsin.

Mendelhall, T.C (1887). The characteristic curves of composition, *Science*, IX, 237–249.

Merriam, T. and Matthews, R. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9, 1-6.

Mosteller, F. and Wallace, D.L. (1964, 84). *Applied Bayesian and Classical Inference; the Case of The Federalist Papers*, 1rst and 2nd edn, Berlin: Springer-Verlag.

Oakes, M.P. (1998). *Statistics for Corpus Linguistics*, Edimburg:Edimburgh University Press.

Peng, F., Shuurmans, D. and Wang, S. (2004). Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval Journal*, 7, 317-345.

Schneider, K.M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering. Proceed. of the tenth conference on the European chapter of the Association for Computational Linguistics, Vol. 1, pp. 307-314

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1-47.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60, 538-556.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26, 471-495.

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

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G. and Tambouratzis, D. (2004). Discriminating the registers and styles in the Modern Greek language-Part 2. Exteding the feature vector to optimize author discrimination. Literary and Linguistic Computing. 19, 221-242.

Tweedie, F., Singh, S. and Holmes, D. (1996). Neural network applications in stylometry: The Federalist papers. *Computers and the Humanities*, 30, 1-10.

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

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23, 327-342.

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.

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

Zheng, R., Li, J., Chen, H. and Huang, Z. (2006). A framework for authorship identification of onlie messages: Writing style features and classification techniques. *Journal of the American Society of Information Science and Technology*, 57, 378-393.

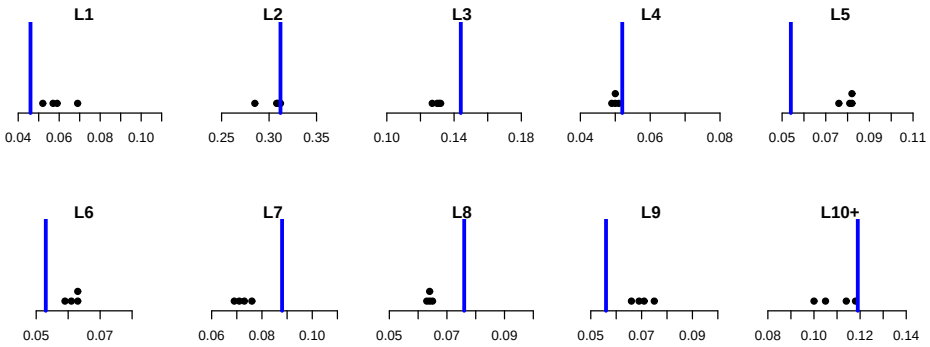

Figure 1: Dots indicate the proportion of *l*-lettered words, *Ll*, in S_1 to S_4. Lines indicate the proportions in *D*.

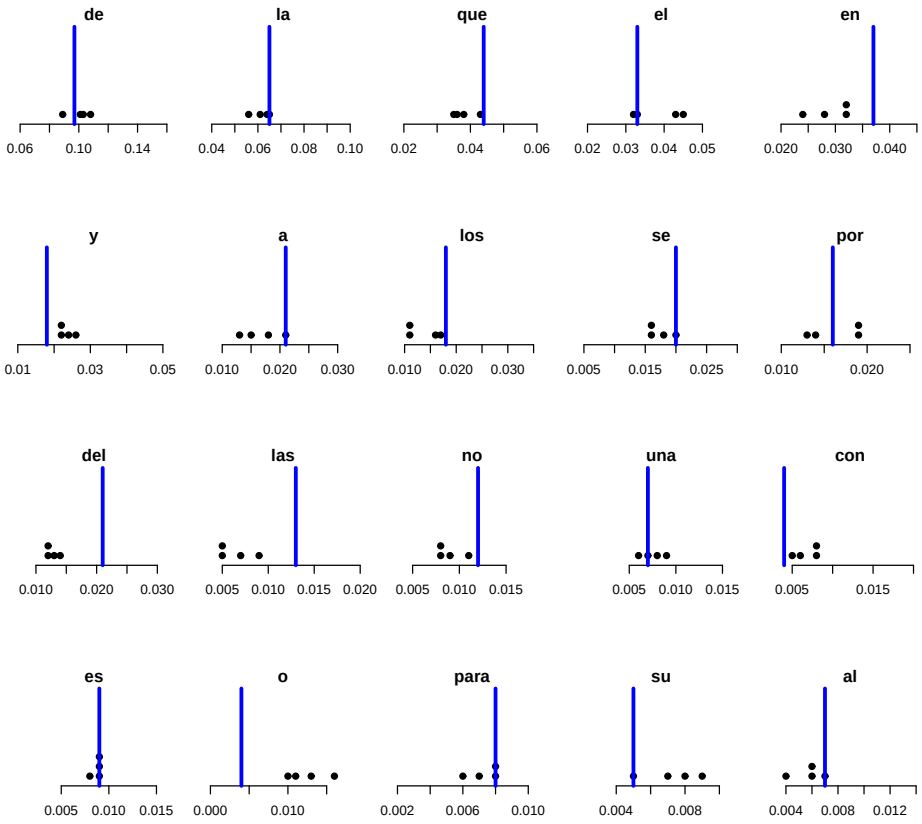

Figure 2: Dots indicate the frequency of the twenty most frequent function words in S_1 to S_4. Lines indicate the corresponding frequency in D.

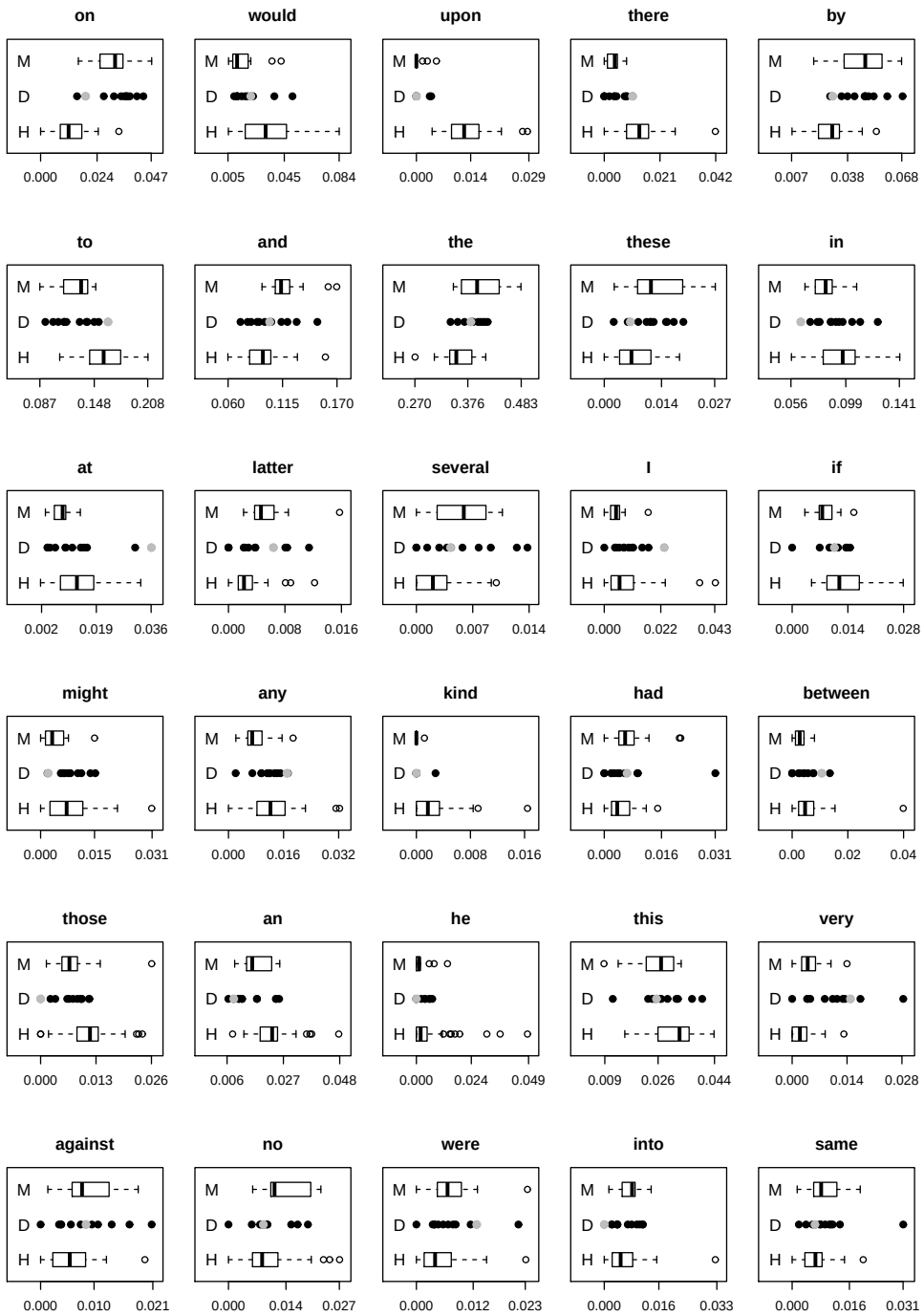

Figure 3: Comparison of the frequencies of appearance of the thirty most discriminating words in the papers known to be by Hamilton and by Madison, and in the twelve disputed papers. The counts for the disputed paper 55, with a style closer to Hamilton than to Madison are shaded lighter.

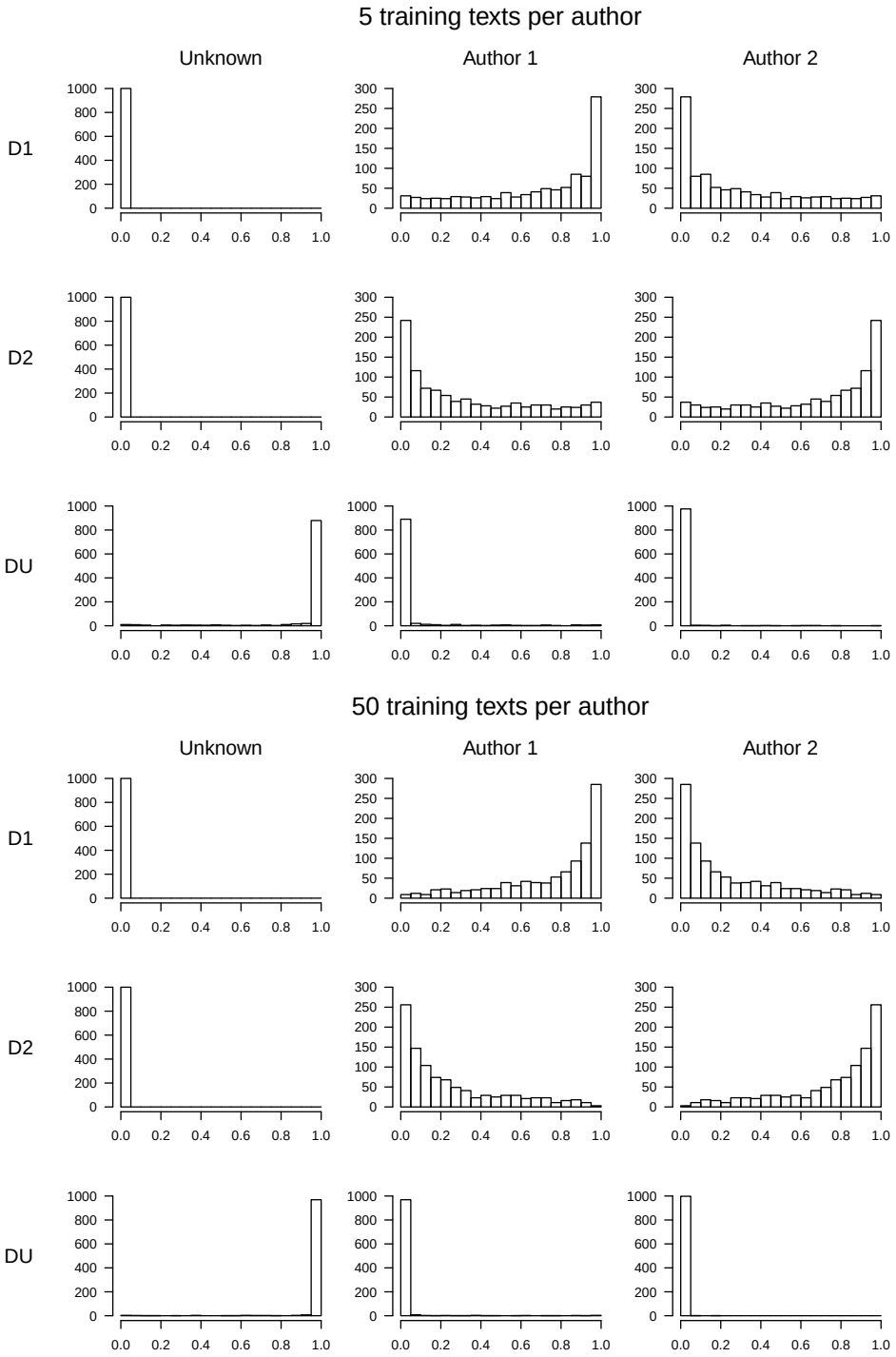

Figure 4: Histogram of the sample of 1000 posterior probabilities of the three authorship hypotheses, with D1 being by Author 1, D2 being by Author 2, and with DU being by an unknown author.

Table 1: Number of l-lettered words for $l = 1, 2, \ldots, 9$ and for $l > 9$, and number of times that the ten most frequent words appear in the sentences. The other words used in the analysis are *del, las, no, una, con, es, o, para, su* and *al*. D is the disputed sentence, and S_1, S_2, S_3 and S_4 is the training set.

word length counts											
court ruling	1	2	3	4	5	6	7	8	9	10+	N_i
D	**598**	**4069**	**1882**	**673**	**707**	**689**	**1145**	**997**	**737**	**1554**	**13051**
S_1	158	942	397	149	249	191	220	196	200	318	3020
S_2	629	2587	1200	450	690	573	631	579	680	1070	9089
S_3	186	978	413	160	257	192	241	198	224	316	3165
S_4	560	3049	1257	499	810	582	705	629	683	1126	9900
Function word counts											
court ruling	de	la	que	el	en	y	a	los	se	por	...
D	**1269**	**851**	**568**	**437**	**480**	**240**	**277**	**229**	**260**	**204**	**...**
S_1	310	184	107	129	85	67	39	34	54	56	...
S_2	806	509	392	297	289	236	192	144	147	116	...
S_3	320	202	115	143	77	77	58	36	62	61	...
S_4	1067	642	376	312	317	214	147	164	157	137	...

Table 2: Posterior probability that the style of a sentence is the same as the style in the other ones, $P(M_1|y)$. D is not used in the first four rows, checking whether S_1 to S_4 share the same style.

Sentence	word length	function words
S_1	1.00	1.00
S_2	0.99	1.00
S_3	1.00	1.00
S_4	1.00	1.00
D	0.00	0.00

Table 3: Posterior probabilities of the three authorship hypotheses, for each one of the disputed papers, using our set of thirty most discriminant words.

	text											
author	49	50	51	52	53	54	55	56	57	58	62	63
Hamilton	0.	0.	0.	0.	0.	0.	.78	0.	0.	0.	0.	0.
Madison	1.	1.	1.	1.	1.	1.	.22	1.	1.	1.	1.	1.
Unknown	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.

Table 4: Estimated probability of correct classification under the Bayesian multinomial method (BM), a decision tree method (DT), a support vector machine method (SVM), and a logistic regression method (LR).

5 training texts per author				
text	BM	DT	SVM	LR
D1	0.733	0.639	0.588	0.653
D2	0.717	0.577	0.584	0.616
DU	0.946	–	–	–
50 training texts per author				
text	BM	DT	SVM	LR
D1	0.824	0.671	0.784	0.793
D2	0.816	0.674	0.704	0.793
DU	0.989	–	–	–