Estimating the Probability of an Authorship Attribution

Jacques Savoy
Computer Science Department, University of Neuchatel, Rue Emile Argand 11, Neuchâtel 2000, Switzerland. E-mail: jacques.savoy@unine.ch

In authorship attribution, various distance-based metrics have been proposed to determine the most probable author of a disputed text. In this paradigm, a distance is computed between each author profile and the query text. These values are then employed only to rank the possible authors. In this article, we analyze their distribution and show that we can model it as a mixture of 2 Beta distributions. Based on this finding, we demonstrate how we can derive a more accurate probability that the closest author is, in fact, the real author. To evaluate this approach, we have chosen 4 authorship attribution methods (Burrows' Delta, Kullback-Leibler divergence, Labbé's intertextual distance, and the naïve Bayes). As the first test collection, we have downloaded 224 *State of the Union* addresses (from 1790 to 2014) delivered by 41 U.S. presidents. The second test collection is formed by the *Federalist Papers*. The evaluations indicate that the accuracy rate of some authorship decisions can be improved. The suggested method can signal that the proposed assignment should be interpreted as possible, without strong certainty. Being able to quantify the certainty associated with an authorship decision can be a useful component when important decisions must be taken.

Introduction

In text categorization, various classifiers have been proposed to determine the most appropriate category (or categories) for a query text. These predefined categories could be thematic labels (topical text categorization) (Sebastiani, 2002), the most probable author (authorship attribution) (Stamatatos, 2009), or other discriminative factors (e.g., sentiment or opinion analysis [Pang & Lee, 2008]). In such categorization tasks, the answer could be a Boolean value (binary classifier), a single label selected over a set of possible tags, the probability of belonging to the most likely category, or a ranked list of categories (sorted according to the estimated fitness of the various categories to the query text).

When considering the wide range of possible applications, obtaining only the target category could be enough in some cases (hard classifiers). For example, in handwriting recognition, one simply needs the character or word that corresponds best to the input features. Similar expectations could apply for e-mail filtering in which the user just wants the removal of spam e-mails without further information. In authorship attribution, the system answer cannot be limited to a single person, namely, the most probable author of the disputed text. A better approach is to return a ranked list of possible authors (soft classifiers) with their corresponding fitness to the disputed text. This latter value is, however, difficult to interpret by the user who is unable to predict if a given value or the difference between two values should be viewed as either small or large. Obtaining an accurate probability of the identity of the real author would be more useful, but this information is usually not given. Of course, the proposed decision should also be justified by stylistic considerations that cannot be usually generated automatically.

Focusing on authorship attribution, the aim of this article is to propose a method to estimate the probability that a proposed authorship attribution is correct. In fact, even when an authorship attribution scheme returns a fitness value between an author profile and the doubtful text, such values are usually difficult to interpret. In this article, we model these fitness values as belonging to a mixture of two distributions. The first corresponds to all values obtained when considering correct attributions. The second represents the distribution of incorrect assignments. When the underlying authorship attribution is based on a distance, the values corresponding to correct assignments tend to be smaller than values associated with incorrect assignments. However, a clear and disjoint separation between the two distributions never occurs in practice. Thus, instead of being limited to either a single author name with or without a fitness value, the proposed algorithm will return a precise probability estimation that the proposed solution is correct, or indicate that

Received September 23, 2014; revised October 17, 2014; accepted October 17, 2014

© 2015 ASIS&T • Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23455

the given evidence is not sufficient to reach a decision with a high degree of certainty.

The next section presents an overview of related work while the Evaluation Corpora section depicts the main features of the corpora used in our experiments. In Author Attribution Models, we describe four authorship attribution models used in our study. The Mixture Model section describes our mixture model to represent the distribution of distance values. Based on this model, the last section evaluates four distinct authorship attribution schemes using our corpora.

Related Work

Computer-based authorship attribution (Craig & Kinney, 2009; Stamatatos, 2009) aims to determine, as accurately as possible, the author of a disputed text (e.g., a part of a play or an anonymous letter) based on text samples written by known authors. Using this general definition, we can find the closed-class attribution problem where the real author is one of the given candidates. In the open-set problem, the real author could be one of the specified authors or an unknown one. Authorship attribution can also be implemented to mine demographic or psychological information on an author (profiling) (Argamon, Koppel, Pennebaker, & Schler, 2009) or simply determine whether or not a given author did, in fact, write a given text (chat, e-mail, or testimony) (verification) (Koppel, Schler, & Bonchek-Dokow, 2007).

To solve the closed-set question, various authorship attribution approaches try to derive the particular style of the disputed text and those corresponding to the possible candidates. These stylistic representations are usually based on either the frequency analysis of functional words (e.g., determiners, pronouns, prepositions, conjunctions, and certain auxiliary and modal verbs) or on the top *m* most frequent words, with *m* between 50 and 1,000 (Burrows, 2002; Hoover, 2004; Savoy, 2015a). Other studies have considered the letter frequencies (Merriam, 1998) or short sequences of letters (*n*-gram), but always by focusing on the most frequent ones (Kešelj, Peng, Cercone, & Thomas, 2003).

As additional sources of evidence, we can take account of the part-of-speech (POS) distribution or short sequences of POS tags. Moreover, some studies have also considered structural and layout features, such as the number of lines per sentence or per paragraph, paragraph indentation, number of tokens per paragraph, presence of greetings or particular signature formats, as well as features derived from HTML tags. Both syntactical and layout information tends, at best, to slightly improve the overall performance over word-based features (Zheng, Li, Chen, & Huang, 2006).

Based on such representations, different distance-based methods can be applied to determine the distance, or inversely the similarity, between the possible authors and the disputed text. These distance values are then used to rank the authors, and the closest is defined as the most probable author (Burrows, 2002; Zhao & Zobel, 2007; Labbé, 2007).

Machine-learning is another approach (Mitchell, 1997; Barber, 2012). Based on a set of texts with known authorship, the computer can learn the particular style corresponding to each author. Within this paradigm, a feature selection is usually applied first to extract from the entire vocabulary the words (or *n*-gram of characters) that can best discriminate between the different authors. Various experiments have been conducted based on different models such as the naïve Bayes, the *k*-nearest neighbors (*k*-NN), the support vector machine (SVM), or nearest shrunken centroids (Jockers & Witten, 2010; James, Witten, Hastie, & Tibshirani, 2013).

With these methods, the important and useful output is the name of the author having his profile the closest to the disputed text. The value of the distance (or the fitness) is either not given or difficult to interpret. When a probability that the corresponding author is the real one is given, it is not clear how this estimation is computed for many machine-learning approaches. Moreover, returning a probability = 1.0 the majority of the time tends to appear overoptimistic (or might indicate an overfitting during the learning stage). However, some approaches, such as the logistic regression report on the probabilities of the different categories based on a clear, well-understood theoretical framework (James et al., 2013). Van Halteren, Baayen, Tweedie, Haverkort, and Neijt (2005) also suggest estimating the probability that a given writer is the real author of a text. In this case, the proposed estimation is specific to the suggested method based on an adhoc weighted voting scheme combining a large number of feature-value pair sets (with normalization factors to guarantee that the resulting values are between 0 and 1).

Better still, Labbé (2007) suggests taking into account the resulting distance between the disputed text and the possible author. In this case, if the computed distance is large, the proposed attribution should be interpreted as possible, given without certainty. To define this threshold, Labbé (2007) assumes that all intertextual distance values computed with the training set follow a Gaussian distribution. Based on this distribution, all attributions based on a distance smaller than *mean*—2.5 * *standard deviation* (representing approximately 0.5% of the cases) will be interpreted as *good evidence*.

Following this line of thought, we will propose a better model to describe the distribution of the distance values corresponding to correct and incorrect attributions. Moreover, we can derive a probability estimate that the proposed attribution is correct based on clear theoretical basis. To evaluate this proposition, two distinct corpora described in the next section have been used together with four authorship attribution schemes.

Evaluation Corpora

The first corpus used in our experiments contains 224 *State of the Union* addresses (SUAs) delivered by 41 U.S. presidents, from G. Washington (1790) through to B. Obama

(2014). The main objective of these speeches is to inform the Congress and the nation about the state of the country and the world on the one hand and, on the other, to announce legislative projects for the upcoming year. A more detailed analysis of the form and political functions of these addresses can be found in Hoffman and Howard (2006).

Some of these SUA addresses are well known for explaining an important issue or a political position held for decades, such as the Louisiana purchase (1803), the Monroe Doctrine (1823), the Roosevelt corollary to the Monroe Doctrine (1904), the *Four Freedoms* (1941), or the *War on Poverty* (1964). In others, we can find the first occurrence of well-known expressions, such as the *axis of evil* (2002). In a recent study, Savoy (2015b) shows that, when applying a clustering algorithm on this collection, all speeches appearing under the same presidency tend to cluster under the same cluster.

This corpus was generated by downloading all speeches from the website www.presidency.ucsb.edu. There are no speeches for two presidents (W. H. Harrison [1841] and J. A. Garfield [1881]) because their terms were limited to a few months. We have also removed the single speech written by Z. Taylor (1849). In fact, it is not possible to test an authorship attribution scheme with a unique document that cannot be used simultaneously to build an author profile and serve as query text. The Appendix Table A1 presents, with more details, a complete list of all U.S. presidents with the number and date of their SUA addresses.

Each speech has been cleaned by replacing certain UTF-8 punctuation marks with their corresponding ASCII symbol. When needed, the diacritics found in certain words (e.g., *détente*) have been removed. Moreover, the contracted forms are replaced by their equivalent full forms (e.g., *we're* into *we are*).

To represent each text, we can employ the word tokens (e.g., *taken*, *takes*, *took* or *taxes*, *tax*) or the word types (lemmas or entries in the dictionary). Using this last form, word tokens belonging to the same dictionary entry are regrouped under the same word type (e.g., *take* or *tax* in our previous examples). Such an approach has the advantage of ignoring possible variations resulting from syntax.

To define the corresponding word type to each word token, we used the POS tagger proposed by Toutanova, Klein, Manning, and Singer (2003). For each sentence given as input, this system provides the corresponding POS tag to each token. For example, from the sentence "I want to implement the Wall Street reform law.", the POS tagger returns "I/PRP want/VBP to/TO implement/VB the/DT Wall/NNP Street/NNP reform/NN law/NN ./." Tags may be attached to nouns (NN, noun, singular, NNS noun, plural, NNP proper noun, singular), verbs (VB, base form, VBG gerund or present participle, VBP non-third-person singular present, VBZ third-person singular present), adjectives (JJ), personal pronouns (PRP), prepositions (IN), determiners (DT), and adverbs (RB). These morphological tags correspond mainly to those used in the Brown corpus (Francis & Kučera, 1982). This information makes it possible to derive the word type by removing the plural form of nouns (e.g., *laws*/NNS → *law*/NN) or by substituting inflectional suffixes of verbs (e.g., *creates*/VBZ → *create*/VB).

After this preprocessing, this U.S. corpus contains 1,955,699 tokens for 20,589 distinct lemmas (length of the vocabulary). When considering the occurrence frequency, we have 6,242 *hapax legomena* (word types appearing only once and corresponding to 30.3% of the whole vocabulary) and 2,426 *dis legomena* (word types occurring exactly twice, representing 11.8% of the vocabulary). The definite determiner *the* (151,068 occurrences) is the most frequent word type, followed by *of* (97,818), the comma (96,128), *be* (65,455), the full stop (61,563), *to* (60,182), *and* (59,920), *in* (38,335), *an* (33,817), and *we* (31,214).

At the speech level, the mean length is 8,731.2 tokens (standard deviation [*SD*], 5,860). The longest address was written by Taft in 1910 (30,773 tokens) and the shortest by Washington in January 1790 (1,180 tokens). When considering the mean length per president, Adams (1797–1800) wrote the shortest remarks (average of 1,931 tokens per speech) while Taft (1909–1912) is the author, of the longest addresses (24,655 tokens).

As a second evaluation corpus, we have selected the *Federalist Papers* (Rossiter, 2003) composed of 85 articles, a test collection already used in authorship attribution studies (Mosteller & Wallace, 1964; Jockers & Witten, 2010; Savoy, 2013). This set of articles was written to persuade the citizens of New York to ratify the U.S. Constitution (adopted September 17, 1787 in Philadelphia and ratified on July 26, 1788 by the State of New York). These articles were published between October 1787 and April 1788 in three newspapers under the pseudonym of *Publius*. Although, at the time of publication, the authorship of each article was kept secret, contemporaries guessed the joint work of General Alexander Hamilton (1755–1804), James Madison (1751–1836), and John Jay (1745–1829), without being able to explicitly attribute each article to its legitimate author.

From these 85 articles, 70 are undisputed (5 by Jay, 14 by Madison, and 51 by Hamilton) and can be used for training purposes. For three articles, the authorship might be a collaborative effort between Madison and Hamilton, and therefore they are ignored. For the rest (test set), 12 articles could have been written by either Hamilton or Madison (article nos. 49–58 and 62–63).

To generate this corpus, the *Federalist* articles have been downloaded from the Gutenberg project (www.gutenberg.org). From the plain text version, we have ignored the Gutenberg boilerplate as well as all metadata (e.g., title, author name, and source), the footnote texts, and calls. In all articles, the recurrent first sentence ("To the People of the State of New York") has been removed. All the text was transposed to lowercase and tokenized to determine words (sequence of letters or digits) and punctuation symbols. As for the SUA corpus, each article is represented by its lemmas.

Based on this representation, the *Federalist* corpus contains 203,190 tokens for 7,191 distinct lemmas (length of the vocabulary). In this collection, we can find 2,616 *hapax legomena* (corresponding to 36.4% of the whole vocabulary) and 1,059 *dis legomena* (representing 14.7% of the vocabulary). The definite determiner *the* (16,849 occurrences) is the most frequent word type, followed by the comma (12,644), *of* (11,192), *be* (8,367), *to* (6,767), the full stop (4,973), *and* (4,833), *an* (4,768), *in* (4,276), and *it* (3,212).

The average length of an article is 2,478 tokens (*SD*, 840). When considering the mean length per author, Hamilton shows an average of 2,423 tokens per article, Madison 3,107, and Jay 1,875. For the 12 disputed articles, this mean is 2,230 tokens per text.

Authorship Attribution Models

As for authorship attribution schemes, we have selected three distance-based approaches, namely, the Delta rule described in the next subsection, the Kullback-Leibler divergence, and Labbé's distance. Derived from the machine-learning paradigm, we have chosen to evaluate the naïve Bayes model presented in the last subsection.

Delta Rule

To determine the most probable author of a disputed text, Burrows (2002) suggests taking into account the occurrence of very frequent terms (e.g., the top $m = 50$ to 200 most frequent word types). In this set, we can find many function words (e.g., determiners, prepositions, conjunctions, pronouns, and certain auxiliary verbal forms). In this attribution scheme, the underlying idea is to consider only those very frequent words used unconsciously by an author and able to reveal his or her own style markers.

Burrows proposes to not directly consider the absolute occurrence frequencies, but rather their standardized scores. This Z-score value is computed for each selected term t_i in a corpus by calculating its relative term frequency rtf_{ij} in a particular document d_j, as well as the mean ($mean_i$), and standard deviation (S_i) of term t_i according to all texts belonging to the underlying corpus, as depicted in Equation (1) (Hoover, 2004).

$$Z\ score(t_{ij}) = \frac{rtf_{ij} - mean_i}{S_i} \qquad (1)$$

Once these dimensionless quantities are obtained for each selected term, we can then compute the distance to those obtained from author profiles. Given a query text Q, an author profile A_j, and a set of terms t_i, for $i = 1, 2, \dots, m$, Burrows (2002) suggests to compute the Delta value (or the distance) by applying Equation (2).

$$Delta(Q, A_j) = {}^{1}\!/_{m} \cdot \sum_{i=1}^{m} \left| Z\ score(t_{iq}) - Z\ score(t_{ij}) \right| \qquad (2)$$

When computing this distance, we attribute the same importance to each term t_i, regardless of its absolute occurrence frequencies. Large differences may occur when, for a given term, both Z scores are large and have opposite signs. In such cases, one author tends to use the underlying term more frequently than the mean while the other employs it very infrequently. Finally, to determine the most probable author, we select the author A_j depicting the smallest distance Delta.

Kullback-Leibler Divergence

Zhao and Zobel (2007) suggest considering a limited number of predefined terms to discriminate between different possible authors of a disputed text. Their proposed list consists of 363 English terms, mainly function words (e.g., *the*, *in*, *but*, *not*, *am*, *of*, *can*), as well as certain frequently occurring forms (e.g., *became*, *nothing*). Other entries are not very frequent (e.g., *howbeit*, *whereafter, whereupon*), whereas some reveal the underlying tokenizer's expected behavior (e.g., *doesn, weren*) or seem to correspond to certain arbitrary decisions (e.g., *indicate*, *missing*, *specifying*, *seemed*).

After defining the feature set, the probability of occurrence of each word associated with a given author or a disputed text has to be estimated. Based on these estimations, we can define the degree of disagreement between the two probabilistic distributions. To do so, Zhao and Zobel (2007) suggest using the Kullback-Leibler divergence (KLD) formula, also known as relative entropy (Manning & Schütze, 1999). The KLD value expressed in Equation (3) indicates how far the term distribution derived from the query text Q diverges from the *j*th author profile distribution A_j (concatenation of all texts written by the same writer):

$$KLD(Q \parallel A_j) = \sum_{i=1}^{m} \mathrm{Prob}_q[t_i] \cdot \log_2 \left[\frac{\mathrm{Prob}_q[t_i]}{\mathrm{Prob}_j[t_i]} \right] \qquad (3)$$

where $\mathrm{Prob}_q[t_i]$ and $\mathrm{Prob}_j[t_i]$ indicate the occurrence probability of the term t_i in the query text Q or in the author profile A_j, respectively. With this definition, and when the two distributions are identical, the resulting value is zero, whereas in all other cases the returned value is positive. As a decision rule, we assign the query text to the author whose profile shows the smallest KLD value.

To estimate the underlying probabilities, we may consider the term occurrence frequency (denoted tf_i) and the length in tokens of the corresponding text (n) (e.g., $\mathrm{Prob}[t_i] = tf_i/n$, estimation based on the maximum likelihood principle). This solution tends, however, to overestimate the occurrence probability of terms appearing in the sample to the detriment of missing words. To resolve this anomaly, we suggest smoothing the probability estimates using the Lidstone technique (Manning & Schütze, 1999) based on the following estimation: $\mathrm{Prob}[t_i] = (tf_i+\lambda) / (n + \lambda|V|)$, with $|V|$ indicating the vocabulary size. Based on past experiments (Savoy, 2012), this λvalue was fixed to 0.1, which produces a slightly better performance over other choices. Finally, as for the Delta rule, the smallest KLD value indicates the most probable author of the disputed text.

Labbé's Distance

To determine the most probable author of a disputed text, Labbé (2007) suggests computing a distance between the disputed text Q and each author profile A_j. The proposed distance measure is a function of the overlap between the two texts, with a minimum value 0 and the maximum equal to 1. Between these two limits, the distance depends on both the number of terms in common and their frequencies. To be precise, the distance between the author profile A_j and the disputed text Q is given by Equation (4), assuming that the author profile A_j is longer than the text Q:

$$\mathrm{Dist}(A_j, Q) = \frac{\sum_{i=1}^{m} \left| \hat{tf}_{ij} - tf_{iq} \right|}{2 \cdot n_q}$$
$$\text{with } \hat{tf}_{ij} = tf_{ij} \cdot \frac{n_{Aj}}{n_q}, \quad n_q = \sum_{i}^{m} tf_{iq} \quad \text{and} \quad n_q = n_{\hat{A}j} = \sum_{i}^{m} \hat{tf}_{ij} \tag{4}$$

where tf_{iq} indicates the term frequency (number of occurrences) of term t_i in Q, n_q and n_{Aj} the length of document Q, respectively, the author profile A_j.

The important aspect in this computation is to compare two texts with the same length. To achieve this, Equation (4) reduces the term frequencies of the longest document (assumed to be the author profile A_j in Equation [4]) by multiplying them with the ratio of the two text sizes. After this normalization, $n_{\hat{A}j} = n_q$. Then for each term t_i that appears either in A_j or Q, we compute the absolute value of the difference in term frequencies.

Based on this formulation, and when all term frequencies are equal in both texts (we have twice the same texts), the summation will be zero. On the other hand, when the two texts have nothing in common, the sum returns the value $n_{\hat{A}j} + n_q$ or $2 \,.\, n_q$. Dividing this value by $2 \,.\, n_q$, the final distance will be 1.

Unlike the two previous authorship attribution methods, this scheme could be based on the whole vocabulary (as suggested by Labbé [2007]) or on a selection of the *m* most important terms. In order to achieve effective attributions, Labbé (2007) specifies some limits for this approach. First, each text should be longer than 5,000 tokens. This is not a hard constraint, but it is useful to have an idea about a minimal length. Second, the ratio between the largest and smallest text must be maximally around 1:8.

Naïve Bayes

The three previously described authorship attribution methods belong to the classical distance-based paradigm. In the machine-learning domain, other approaches have been suggested for text categorization (Sebastiani, 2002) or authorship attribution (Jockers & Witten, 2010). As a learning scheme, we select the naïve Bayes model (Mitchell, 1997; Manning, Raghavan, & Schütze, 2008) to determine the probable writer between the set of possible authors (or hypotheses), denoted by A_j. To define the most probable author, the naïve Bayes model selects the one maximizing Equation (5), in which t_i represents the *i*th term included in the query text Q, and *m* indicates the number of selected words.

$$\begin{aligned} \mathrm{ArgMax}_{A_j} \mathrm{Prob}[A_j \mid Q] &\propto \mathrm{Prob}[A_j] \cdot \prod_{i=1}^{m} \mathrm{Prob}[t_i \mid A_j] \\ &\propto \ln(\mathrm{Prob}[A_j]) + \sum_{i=1}^{m} \ln(\mathrm{Prob}[t_i \mid A_j]) \end{aligned} \tag{5}$$

To estimate the previous probabilities of each author ($\mathrm{Prob}[A_j]$), we simply take into account the proportion of texts written by each author or assume a uniform distribution over all writers. This second possibility was applied in our experiments.

To determine the term probabilities $\mathrm{Prob}[t_i \mid A_j]$, all texts belonging to the same author are regrouped to define the author profile. For each term t_i, we compute the ratio between its occurrence frequency in the author profile A_j (tf_{ij}) and the length of this sample (n_{Aj}). As with the previous methods, the Lidstone smoothing approach was applied to estimate each probability $\mathrm{Prob}[t_i \mid A_j] = (tf_{ij}+\lambda) / (n_{Aj}+\lambda.|V|)$, with λ as a parameter (set to 0.1) and $|V|$ indicating the vocabulary size.

Unlike the three previous distance-based attribution schemes, the largest value will indicate the most probable author of the disputed text. This number is not strictly speaking a probability, but a value proportional to the corresponding probability than the author A_j is the real author of text Q. Moreover, instead of multiplying *m* probabilities (and some of them could be very small), this equation is transformed by taking the natural logarithm. The sequence of multiplications is then transformed into a series of additions, as shown in the left part of Equation (5).

Mixture Model

When applying any of the previously described attribution schemes, the system returns a list of possible authors ranked according to their fitness values to the disputed text. This value could be proportional to the estimated probability of being the right author (naïve Bayes), an intertextual distance between 0 and 1 (Labbé), a positive distance (Delta rule), or a measure of the divergence between two distributions (KLD). How can we interpret this value? What is a small versus a large value? When does this value correspond to a correct attribution with a high degree of certainty or given without strong support?

To answer these questions, the distance value *d* (or fitness) can be modeled as a mixture model of two distributions. The first one, denoted D_1, corresponds to correct assignments, while the second, indicated by D_2, matches the incorrect attributions. The general probability density corresponding to a mixture of two distributions is given in

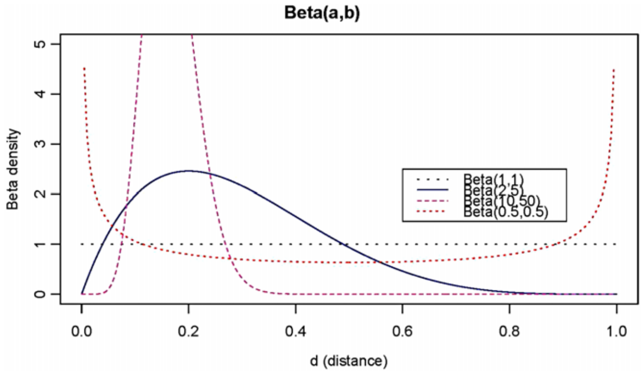

FIG. 1. Four different density functions (Beta distribution) produced by different parameter settings. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Equation (6), where α denotes the relative importance of each distribution, and u_1, u_2 indicate the underlying parameters of the two distributions.

$$f(d) = \alpha \cdot D_1(d \mid u_1) + (1-\alpha) \cdot D_2(d \mid u_2) \quad \text{with} \quad 0 \le \alpha \le 1 \qquad (6)$$

Even if D_1 and D_2 can be modeled by two different probability distributions, we usually select the same one, and the Gaussian is a classical choice (Bishop, 2006).

In authorship attribution, empirical distributions are usually not symmetrical, a reason not to choose a Gaussian distribution. Moreover, the possible values for the distance d tend to be limited and certainly do not cover a large spectrum (as is possible with a Gaussian one). Finally, when the variable d represents a distance, the distribution D_1 (correct assignments) will show smaller values than the distribution based on incorrect attributions. When d indicates a similarity measure (e.g., or argmax with the naïve Bayes approach), the reasoning is reversed.

Considering these facts, we prefer representing the distribution of the distance d using the Beta probability density function defined by Equation (7). In this representation, the variable d can take any value in the interval [0,1]. The precise curvature of the density function is defined by the two parameters a and b. For example, when $a = b = 1$, the distribution is uniform (dotted horizontal line in Figure 1); all possible values of d have the same chance of occurring. When both a and b are greater than 1, the distribution is unimodal. This is the case in the authorship attribution studies. Moreover, when $b > a > 1$, the mode is smaller than 0.5 (two examples are given in Figure 1). In our context, this situation is very common for distributions corresponding to the correct attributions. On the other hand, when $a > b > 1$, the mode is greater than 0.5. Figure 1 illustrates the Beta density function for different parameter settings.

$$Beta(d \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} \cdot d^{a-1} \cdot (1-d)^{b-1} \quad \text{with} \qquad (7)$$
$$a > 0, b > 0, \Gamma() \text{ is the Gamma function}$$

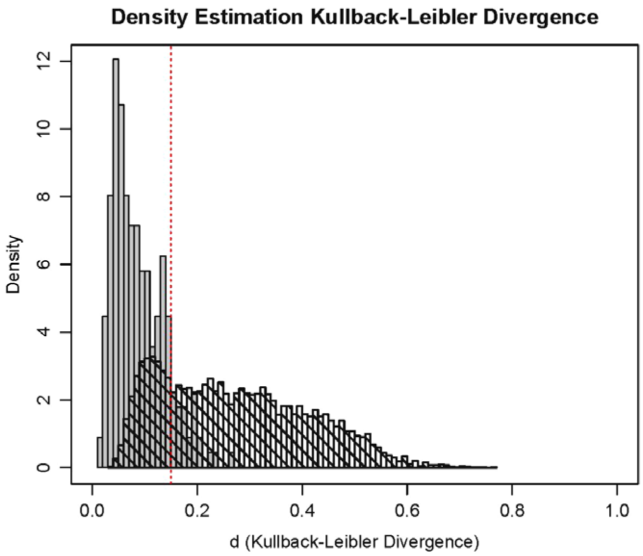

FIG. 2. Density estimations for KLD based on the SUA corpus (correct attributions in dark gray, incorrect assignments with stripes). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Having defined a value for the two parameters a and b, the mean and the variance of a Beta distribution can be computed according to the following equations (Equation [8]).

$$\text{Mean } \mu = \frac{a}{a+b} \quad \text{and Variance } \sigma^2 = \frac{a \cdot b}{(a+b)^2 \cdot (a+b+1)} \qquad (8)$$

The precise values of the mean and the variance are not usually known when the exact values of the parameters a and b are unknown. We can estimate them using the following formulations (Equation [9]).

$$\hat{\mu} = \bar{\mathrm{d}} = \frac{1}{n} \cdot \sum_{i=1}^{n} d_i \quad \text{and} \quad \hat{\sigma}^2 = \mathrm{S}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (d_i - \bar{\mathrm{d}})^2 \qquad (9)$$

Finally, the values for a and b are usually unknown, but can be estimated by the following formula (Equation [10]).

$$\hat{\mathrm{a}} = \bar{\mathrm{d}} \cdot \left(\frac{\bar{\mathrm{d}} \cdot (1-\bar{\mathrm{d}})}{\mathrm{S}^2} - 1 \right) \quad \text{and} \quad \hat{\mathrm{b}} = (1-\bar{\mathrm{d}}) \cdot \left(\frac{\bar{\mathrm{d}} \cdot (1-\bar{\mathrm{d}})}{\mathrm{S}^2} - 1 \right) \qquad (10)$$

Figure 2 depicts two empirical distributions obtained when using the KLD with the SUA corpus. The distribution D_1 (correct assignments, depicted in light gray) shows, on average, smaller values than the distribution D_2 based on incorrect attributions (dark stripes in Figure 2). For example, the mean of distribution D_1 is 0.0889, whereas for D_2 the mean is 0.2711 (median D_1: 0.0778; median D_2: 0.2546). Moreover, the distribution D_1 has a smaller standard deviation than D_2 (in our example, 0.048 vs. 0.14). The values of the distribution D_1 are more concentrated, showing less variability.

DOI: 10.1002/asi

When a small KLD value is obtained between a disputed text and a given author, there is a greater chance that this possible attribution is correct than if a larger one. In Figure 2, the vertical dotted line depicted at position 0.15 splits the KLD values into two; a KLD value smaller than 0.15 tends to indicate a correct assignment, a larger value tends to indicate the opposite.

With this example, the minimal *d* value is 0.0186, found between the SUA of 1905 by T. Roosevelt and T. Roosevelt's profile. The maximum value is 0.76 when computing the KLD between 1827 Monroe's speech and Bush's profile (father). As an interesting case, we can mention the 1916 SUA by Wilson. When computing the KLD with all possible profiles, the minimum value of 0.2617 is achieved with Wilson's profile. Thus, considering only the author with the minimum value, the automatic attribution scheme assigns this text to Wilson. However, when looking at Figure 2, we can see that this minimal value is very close to the median of the distribution D_2 (incorrect assignments). Is such a minimal *d* value a strong support for the attribution to Wilson?

Finally, when analyzing the examples of the Beta distribution displayed in Figure 1 and the empirical distributions depicted in Figure 2, we can infer that for both distributions, the parameters *a* and *b* are greater than 1, and that $a < b$ for the D_1 distribution (correct assignments), leading to a mode smaller than 0.5.

Until now, we have assumed that the maximum value for *d* is 1. This is the case, for example, with the Labbé measure defined between 0 and 1. For the KLD measure, a value larger than one might be obtained, and with the Delta measure, this will certainly be the case. Thus, before applying the mixture model of two Beta distributions, all *d* values have to be divided by the maximum value.

After this normalization, all possible values for *d* are between 0 and 1. Applying this to the data used in Figure 2, Figure 3 illustrates the mixture model after normalizing the KLD data with the approximation for D_1 (light gray and in red) and for D_2 (gray stripes and in blue).

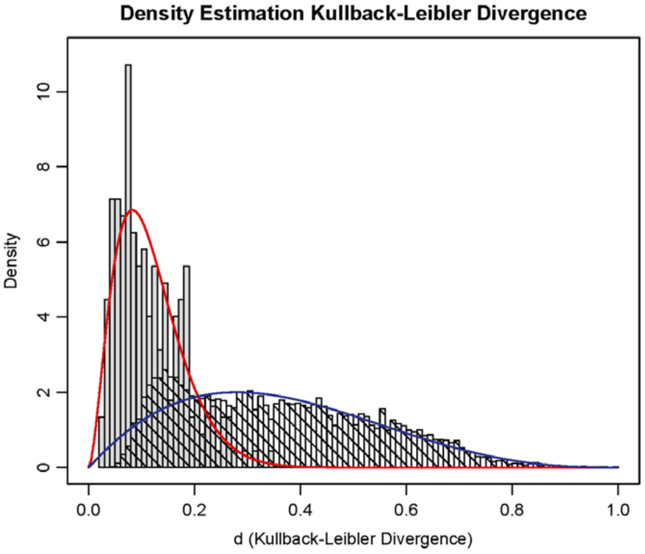

FIG. 3. The mixture of two Beta distributions for the density estimations for KLD based on the SUA corpus (correct attribution in dark gray, incorrect attributions with stripes). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 1. Evaluation (leaving-one-out evaluation methodology) of the four attribution schemes with the 224 SUAs.

Word selection	Delta	KLD	Labbé	Naïve Bayes
300 MFWs	196 **87.5%**	197 **88.0%**	188 **83.9%**	192 **85.7%**
309 words proposed by Hughes et al. (2012)	165 73.7%	182 81.3%	165 73.7%	183 81.6%
344 words proposed by Zhao and Zobel (2007)	165 67.6%	182 81.3%	160 65.6%	178 79.5%

Evaluation

The first experiment is based on the *State of the Union* corpus containing 224 addresses. In this case, it is assumed that all the speeches by one president are written by the same author(s) (or ghost writer[s]). As authorship attribution models, the three distance-based schemes have been applied and, from the machine-learning paradigm, the naïve Bayes approach. To discriminate between the specific style of each writer, the 300 most frequent words (MFWs) are used, a number that tends to produce relatively high performance levels across different authorship schemes (Savoy, 2015a). This list contains many functional words (determiners [*the*, *a*], prepositions [*in*, *for*], conjunctions [*and*, *but*], pronouns [*we*, *where*], some adverbs [*always*, *very*], and some auxiliary and modal verbal forms [*is*, *can*, *will*]). Of course, the main topics of the underlying corpus also produce many entries (such as *citizen*, *federal*, *defense*, *treasury*, or *constitution*).

The first row of Table 1 depicts the evaluation based on this feature set for the four strategies using the 224 speeches with 41 possible authors. During the evaluation, the query text is not used, in any way, to generate the author profile the (leaving-one-out evaluation method). The resulting performance levels are thus not biased.

As an alternative to the MFWs, we have considered 309 words selected by Hughes, Foti, Krakauer, and Rockmore (2012) containing many functional words, but also some frequently used words (e.g., *go*, *show*, *fifty*, *side*, *detail*, or *serious*). This list was applied to analyze the stylistic variations across the last four centuries (Hughes et al., 2012) and therefore may also contain some spelling variations (e.g., *amongst* and *amougst*). The intersection of Hughes's list and the 300 MFWs contains 111 words.

The second row of Table 1 indicates the performance achieved using Hughes's word list. These accuracy rates are lower for all attribution schemes than when considering the 300 MFWs. In the third row of Table 1, we have used the

TABLE 2A. Evaluation (leaving-one-out evaluation methodology) of the attribution schemes with the 65 *Federalist Papers* (training set).

	Delta	KLD	Labbé	Naïve Bayes
300 MFWs	63 96.9%	64 98.5%	65 **100%**	63 96.9%
309 words proposed by Hughes et al. (2012)	64 **98.5%**	65 **100%**	63 96.9%	64 98.5%
344 words proposed by Zhao and Zobel (2007)	63 96.9%	65 **100%**	63 96.9%	65 **100%**

TABLE 2B. Evaluation of the attribution schemes with the 12 disputed *Federalist Papers*.

	Delta	KLD	Labbé	Naïve Bayes
300 MFWs	12 **100%**	11 91.7%	11 **91.7%**	12 **100%**
309 words proposed by Hughes et al. (2012)	11 91.7%	12 **100%**	11 **91.7%**	12 **100%**
344 words proposed by Zhao and Zobel (2007)	10 83.3%	12 **100%**	11 **91.7%**	12 **100%**

TABLE 3. Evaluation of the four attribution schemes with the 224 SUA.

	Delta	KLD	Labbé	Naïve Bayes
300 MFWs	196 / 224 87.5%	197 / 224 88.0%	188 / 224 83.9%	192 / 224 85.7%
Mean	113 / 118 **95.8%**	106 / 113 93.8%	108 / 121 89.3%	97 / 113 85.8%
Undecidable	106	111	103	111
Probability ≥ 0.5	112 / 117 95.7%	110 / 117 **94.0%**	107 / 118 **90.7%**	96 / 112 **85.7%**
Undecidable	107	107	106	112

344 words included in Zhao and Zobel's list and appearing in the corpus as discriminative terms (Zhao & Zobel, 2007). The overall performance level of this feature set is similar to that achieved with Hughes's list and is inferior to that obtained when considering the 300 MFWs.

Using the *Federalist* test collection, a first experiment is performed on the 65 articles with known attribution (14 by Madison and 51 by Hamilton). For each evaluation, one article is removed from the corpus, and the rest used to generate the two author profiles. The different authorship attribution strategies can be tested using the remaining articles. The results of this evaluation approach the (leaving-one-out evaluation method) are given in Table 2A. As shown, the three lists tend to produce similar performance levels.

In Table 2B, the 65 articles with known authorship are used to form the author profiles. The remaining 12 disputed articles are automatically assigned using the four authorship attribution schemes (assuming that the real author is Madison). In this case, both Hughes's and Zhao and Zobel's word lists tend to produce similar performance levels, compared to the 300 MFWs.

The evaluation results in Tables 1, 2A and 2B use the distance value between the disputed text and the different author profiles only to rank the possible authors. Thus, having a very small or a large distance value is not directly taken into account; only the minimal value (or maximal for the naïve Bayes) is used to indicate the most probable author, whatever this value may be. Moreover, it is not easy to interpret a given distance value as small, moderate, or large.

Based on the proposed model to describe the distribution of the correct assignments and the incorrect attributions by a mixture of two Beta distributions, we can use this representation to derive a better understanding of each distance value.

As a first interpretation, we can specify that an automatic decision can be taken only when the minimal distance is smaller than the mean of the correct assignment distribution. When the minimal value is smaller than this threshold, a *good certainty* that the proposed author is the real one can be achieved. On the other hand, when the minimal value is larger than the mean, the proposed attribution cannot be given with a high degree of certainty and should be interpreted more as an *indication* that the suggested author is the real one.

Based on this interpretation, and using the 300 MFWs, the achieved performances are depicted in the second row of Table 3. Using the 224 speeches, the proposed attribution scheme is unable to determine the real author with a high degree of certainty for 103 (the Labbé model) or 111 speeches (KLD or naïve Bayes). The accuracy rate obtained when considering only assignments having a good certainty ranges from 85.8% (naïve Bayes, with 97 correct attributions over 113) up to 95.8% (Delta, 113 correct assignments over 118).

As a second interpretation, we begin again with the minimal distance denoted *d'* obtained with author profile A_j. Estimating the Beta distribution corresponding to the correct attribution (denoted D_1), we can compute the Prob[$d \geq d' \mid D_1$]. This probability is called the *support* in favor of the hypothesis that the real author is A_j. If this probability is larger than 0.5 (or another specified threshold), we can assume with good certainty than the proposed author A_j is the real author of the query text.

As for the first interpretation, we can find a relatively large value for *d'*, rendering the Prob[$d \geq d' \mid D_1$] smaller than 0.5. In this case, the proposed assignment must be interpreted as an indication of a possible authorship. Of course, we can also compute the support of the alternative hypothesis, namely, that the distance *d'* is coming from the Beta distribution corresponding to the incorrect attribution. In this case, this support is given by Prob[$d \leq d' \mid D_2$].

To illustrate these ideas, Figure 4 depicts the mixture model of two Beta distributions with a value $d' = 0.2$. Based on this figure, when a distance *d'* tends toward 0.0, the

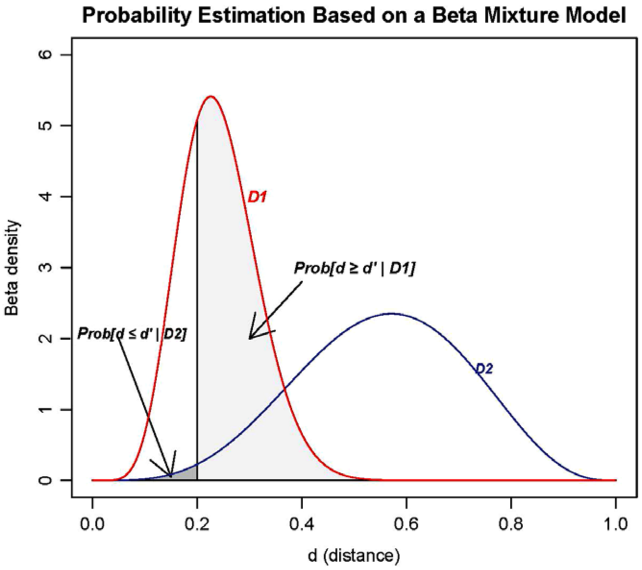

FIG. 4. Support for the two hypotheses with a minimal distance $d' = 0.2$ according to a mixture model of two Beta distributions. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

support for the correct attribution, or Prob[$d \geq d' \mid D_1$], increases and will eventually reach the maximal value of 1.0. On the other hand, the support for an incorrect attribution, or Prob[$d \leq d' \mid D_2$], will decrease and tends toward 0.0. On the other hand, when the value of d' is increasing, the Prob[$d \geq d' \mid D_1$] is decreasing and will tend to 0.0 whereas the support of an incorrect assignment will increase and, finally, will reach the value 1.0 when $d' \rightarrow 1.0$.

The evaluation of this second interpretation is given in the third row of Table 3. As shown, this approach produces similar performance levels to the first interpretation, with between 106 and 112 speeches that cannot be assigned with a high degree of certainty. As for the first interpretation, the accuracy rates are higher than the baseline, given more certainty when proposing an author for a disputed text.

As illustrative examples with the KLD scheme, we can take again the 1905 SUA delivered by T. Roosevelt. This speech obtains the minimal d value (0.0186) over all SUAs. The attribution model suggests T. Roosevelt as the most probable author. When computing the support for this decision (Prob[$d \geq d' \mid D_1$]), the value 0.975 is achieved, indicating a very strong support for T. Roosevelt. On the other hand, the support for an incorrect attribution, or Prob[$d \leq d' \mid D_2$], is 0.004, too small to reject the proposed assignment. Our second example is formed by the 1916 SUA by Wilson. For this speech, the minimal distance is 0.2617 with Wilson's profile. When computing the support for this attribution, we obtain a value of 0.004 whereas the support for an incorrect attribution is 0.51. Of course, the KLD scheme assigns this speech correctly, but with a distance of 0.2617, this decision should be interpreted as possible, without certainty.

TABLE 4A. Evaluation (leaving-one-out evaluation methodology) of the four attribution schemes with the 65 *Federalist Papers*.

	Delta	KLD	Labbé	Naïve Bayes
300 MFWs	63 96.9%	64 98.5%	65 100%	63 96.9%
Mean	37 / 38 97.4%	35 / 35 100%	36 / 36 100%	32 / 32 100%
Undecidable	27	30	29	33
Probability ≥ 0.5	37 / 38 97.4%	37 / 37 100%	37 / 37 100%	31 / 31 100%
Undecidable	27	33	28	34

TABLE 4B. Evaluation of the attribution schemes with the 12 disputed *Federalist Papers*.

	Delta	KLD	Labbé	Naïve Bayes
300 MFWs	12 100%	11 91.7%	11 91.7%	12 100%
Mean	2 / 2 100%	3 / 3 100%	6 / 6 100%	1 / 1 100%
Undecidable	10	9	6	11
Probability ≥ 0.5	2 / 2 100%	3 / 3 100%	6 / 6 100%	1 / 1 100%
Undecidable	10	9	6	11

Using the *Federalist* corpus, the evaluations of this first and second interpretation are reported in Table 4A for the 65 articles with known authorship (leaving-one-out evaluation methodology) and in Table 4B for the 12 disputed articles. In both cases, the performance differences between the two interpretations are small and not significant. With the corpus of 65 articles (Table 4A), the proposed method is able to provide a decision with good certainty for approximately 55% of the cases. For those cases, the attribution accuracy rate is rather high (between 97.4% and 100%).

With the 12 disputed articles (Table 4B), using only the distance to rank the possible authors, the performance level is shown in the first row. This approach produces only two "errors" (article no. 57 with KLD and article no. 55 with the Labbé scheme). When considering only reliable assignments according to the two interpretations, the error rate is zero. The number of assignments with such high certainty is, however, rather low, except for the Labbé attribution scheme that is able to assign 50% of the articles correctly.

In order to analyze the disputed articles of the *Federalist* corpus in more detail, Table 5 reports the degree of support for a correct attribution (or Prob[$d \geq d' \mid D_1$]). Under each column, we find this probability computed according to the smallest distance (or argmax for the naïve Bayes approach). The assignments having a good support according to the second interpretation (or Prob[$d \geq d' \mid D_1$] > 0.5) are presented in bold and corresponding to the numbers depicted in

TABLE 5. Degree of support (Prob[$d \geq d' \mid D_1$]) obtained by each attribution scheme with the 12 disputed papers.

Article no.	Delta	KLD	Labbé	Naïve Bayes
49	0.443	0.484	**0.633**	0.304
50	0.001	0.005	0.001	0.234
51	0.214	0.360	0.243	0.128
52	0.216	0.345	**0.535**	0.333
53	0.337	**0.538**	**0.694**	0.048
54	0.232	0.194	0.453	**0.516**
55	0.115	0.182	*0.309*	0.035
56	0.001	0.003	0.021	0.012
57	0.236	*0.308*	0.367	0.145
58	0.463	0.389	**0.740**	0.299
62	**0.683**	**0.734**	**0.666**	0.172
63	**0.760**	**0.678**	**0.862**	0.419

the third row of Table 4B. For example, under the Labbé approach, six probabilities are shown in bold, corresponding to the six attributions indicated in the third row of Table 4B.

For the three distance-based strategies, the attribution of articles. 62 and 63 to Madison is rather high. For articles 50 and 56, the support for an assignment to Madison is weak, but the intertextual distance is still the smallest with this possible author. Between these two extremes, we can find articles 53, 52, and 58 with good support in favor of Madison.

When inspecting the two attribution errors (mentioned in the first row of Table 4B), we first have article 55 assigned to Hamilton by the Labbé scheme with a probability of 0.309 (value indicated in italics in Table 5). The second divergence appears with article 57 assigned to Hamilton according to the KLD scheme. In this case, the support for this attribution is 0.308.

Conclusion

In authorship attribution, the main research focus has been on proposing more effective classification schemes. The emphasis is usually on the accuracy rate of the suggested method compared to a baseline. Relatively few studies tend explain why and when the proposed assignment is correct or more doubtful.

Here we have analyzed the distribution of the intertextual distance values between the different author profiles and the disputed text. Usually, these values are just used to rank the authors, from the closest and most probable author to the less likely one. This distance (or similarity) measure can, however, be useful to discriminate between more or less certain assignments.

To achieve this goal, we propose to model the distance values as a mixture of two Beta distributions. We prefer this probabilistic distribution to the Gaussian because the possible distance values are limited into a subset of the real numbers (e.g., between 0.0 and 1.0). Moreover, the empirical distributions are clearly asymmetric, rendering the choice of the Gaussian less attractive. On the other hand, the Beta function is limited to values between 0 and 1 and can represent different asymmetric distributions. When considering correct and incorrect attribution, we propose to represent the resulting distribution as a mixture of two Beta distributions.

Based on this model, we demonstrate how we can compute an estimation of the probability that the smallest distance between an author profile and the disputed text indicates the real author. If this probability is too small, we suggest viewing the resulting assignment as only an indication of possible authorship. On the other hand, when the probability is larger, we have a higher degree of certainty associated with the proposed attribution.

Using a test collection comprising 224 speeches delivered by 41 U.S. presidents, the suggested evaluation method shows that we can improve the accuracy of the prediction from approximately 86% to 96%, when the underlying support is strong enough. Based also on the most frequent words and using the *Federalist Papers*, the suggested approach shows that two disputed articles (articles 50 and 56) cannot be assigned without some doubt to Madison. On the other hand, the proposed evaluation method suggests with a high degree of certainty that articles 62 and 63 were written by Madison.

Acknowledgments

This research was supported, in part, by the Swiss NSF under Grant No. 200021_149665/1. The author thanks the anonymous reviewers for their helpful suggestions and remarks.

References

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119–123.

Barber, D. (2012). Bayesian reasoning and machine learning. Cambridge, UK: Cambridge University Press.

Bishop, C. (2006). Pattern recognition and machine learning. New York: Springer.

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. Literary & Linguistic Computing, 17(3), 267–287.

Craig, H., & Kinney, A.F. (Eds.). (2009). Shakespeare, computers, and the mystery of authorship. Cambridge, UK: Cambridge University Press.

Francis, W.N., & Kučera, H. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.

Hoffman, D.R., & Howard, A.D. (2006). Addressing the State of the Union. The evolution and impact of the president's big speech. Boulder, CO: Lynne Rienner.

Hoover, D.L. (2004). Testing Burrows's delta. Literary and Linguistic Computing, 19(4), 453–475.

Hughes, J.M., Foti, N.J., Krakauer, D.C., & Rockmore, D.N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the National Academy of Sciences of the United States of America, 109(20), 7682–7686.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. New York: Springer.

Jockers, M.L., & Witten, D.M. (2010). A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing, 25(2), 215–223.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03 (pp. 255–264). Halifax: Dalhoisie Uni. Press.

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research, 8(6), 1261–1276.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. Journal of Quantitative Linguistics, 14(1), 33–80.

Manning, C.D., & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA: MIT Press.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press.

Merriam, T. (1998). Heterogeneous authorship in early Shakespeare and the problem of *Henry V*. Literary and Linguistic Computing, 13(1), 15–28.

Mitchell, T.M. (1997). Machine learning. New York: McGraw-Hill.

Mosteller, F., & Wallace, D.L. (1964). Inference and disputed authorship. The Federalist. Reading, MA: Addison-Wesley.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

Rossiter, C. (Ed.). (2003). The Federalist Papers. New York: Signet Classic.

Savoy, J. (2012). Authorship attribution based on specific vocabulary. ACM—Transactions on Information Systems, 30(2), 170–199.

Savoy, J. (2013). *The Federalist Papers* revisited: A collaborative attribution scheme. Proceedings ASIST 2013, Montreal, November 2013.

Savoy, J. (2015a). Comparative evaluation of term selection functions for authorship attribution. Digital Scholarship in the Humanities, 2015, DOI: http://www.wada-ama.org/.

Savoy, J. (2015b). Text clustering: An application with the *State of the Union* addresses. Journal of the American Society for Information Science & Technology, DOI: 10.1002/asi.23283.

Sebastiani, F. (2002). Machine learning in automatic text categorization. ACM Computing Survey, 34(1), 1–27.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science & Technology, 60(3), 538–556.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. In Proceedings of HLT-NAACL 2003 (pp. 252–259). Edmonton, Canada: ACL.

Van Halteren, H., Baayen, R.H., Tweedie, F., Haverkort, M., & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. Journal of Quantitative Linguistics, 12(1), 65–77.

Zhao, Y., & Zobel, J. (2007). Searching with style: Authorship attribution in classic literature. In Proceedings ACSC 2007 (pp. 59–68). Ballarat, Australia: CRPIT.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science & Technology, 57(3), 378–393.

Appendix

TABLE A1. List of the U.S. presidents with their number of the SUAs and the years of these speeches.

No.	President name	No. of speeches	From	To
1	George Washington	8	1790	1796
2	John Adams	4	1797	1800
3	Thomas Jefferson	8	1801	1808
4	James Madison	8	1809	1816
5	James Monroe	8	1817	1824
6	John Quincy Adams	4	1825	1828
7	Andrew Jackson	8	1829	1836
8	Martin van Buren	4	1837	1840
9	William H. Harrison	0	1841	1841
10	John Tyler	4	1841	1844
11	James Polk	4	1845	1848
12	Zachary Taylor	1	1849	1849
13	Millard Fillmore	3	1850	1852
14	Franklin Pierce	4	1853	1856
15	James Buchanan	4	1857	1860
16	Abraham Lincoln	4	1861	1864
17	Andrew Johnson	4	1865	1868
18	Ulysses S. Grant	8	1869	1876
19	Rutherford B. Hayes	4	1877	1880
20	James A. Garfield	0	1881	1881
21	Chester A. Arthur	4	1881	1884
22	Grover Cleveland	4	1885	1888
23	Benjamin Harrison	4	1889	1892
24	Grover Cleveland	4	1893	1896
25	William McKinley	4	1897	1900
26	Theodore Roosevelt	8	1901	1908
27	William H. Taft	4	1909	1912
28	Woodrow Wilson	8	1913	1920
29	Warren Harding	2	1921	1922
30	Calvin Coolidge	6	1923	1928
31	Herbert Hoover	4	1929	1932
32	Franklin D. Roosevelt	12	1933	1945
33	Harry S. Truman	7	1947	1953
34	Dwight D. Eisenhower	9	1953	1960
35	John F. Kennedy	3	1961	1963
36	Lyndon B. Johnson	6	1964	1969
37	Richard Nixon	5	1970	1974
38	Gerald R. Ford	3	1975	1977
39	Jimmy Carter	3	1978	1980
40	Ronald Reagan	7	1982	1988
41	George H.W. Bush	4	1989	1992
42	William J. Clinton	8	1993	2000
43	George W. Bush	8	2001	2008
44	Barack Obama	6	2009	2014