Comparative evaluation of term selection functions for authorship attribution

Jacques Savoy
Computer Science Department, University of Neuchatel, Neuchâtel, Switzerland

Abstract

Different computational models have been proposed to automatically determine the most probable author of a disputed text (authorship attribution). These models can be viewed as special approaches in the text categorization domain. In this perspective, in a first step we need to determine the most effective features (words, punctuation symbols, part-of-speech, bigram of words, etc.) to discriminate between different authors. To achieve this, we can consider different independent feature-scoring selection functions (information gain, gain ratio, pointwise mutual information, odds ratio, chi-square, bi-normal separation, GSS, Darmstadt Indexing Approach (DIA), and correlation coefficient). Other term selection strategies have also been suggested in specific authorship attribution studies. To compare these two families of selection procedures, we have extracted articles from two newspapers and belonging to two categories (sports and politics). To enlarge the basis of our evaluations, we have chosen one newspaper written in the English language ('Glasgow Herald') and a second one in Italian ('La Stampa'). The resulting collections contain from 987 to 2,036 articles written by four to ten columnists. Using the Kullback–Leibler divergence, the chi-square measure and the Delta rule as attribution schemes, this study found that some simple selection strategies (based on occurrence frequency or document frequency) may produce similar, and sometimes better, results compared with more complex ones.

Correspondence:
Jacques Savoy, Computer Science Department, University of Neuchatel, Neuchâtel, Switzerland.
Email:
Jacques.Savoy@unine.ch

1 Introduction

In automatic authorship attribution, computer systems can be designed and implemented to determine the most probable author behind a disputed document or a text excerpt (Mosteller and Wallace, 1964; Juola, 2006; Stamatatos, 2009). To achieve this, a set of texts written by each of the possible writers must be made available to the classifier. In this study, we focus on the closed-class attribution problem in which the real author is one of the several possible candidates. Other pertinent concerns related to this issue include the mining of demographic or psychological information on an author (profiling) (Pennebaker, 2011), or simply determining whether or not a given author did write a given Internet message or document (verification) (Koppel *et al.*, 2007). Instead of being limited to text, we can also consider other media (e.g. music, song, picture, drawing).

To solve this categorization task, we need to extract and select features that are useful in identifying

doi:10.1093/llc/fqt047 Advance Access published on 6 August 2013

the differences between the authors' writing styles. In this study, we will consider words and punctuation symbols as possible features or terms. In a second step, we determine the discrimination power of each term and then apply a selection procedure to derive a reduced set of terms that can effectively discriminate between the different possible authors. Finally, through applying classification rules or schemes, the system can determine the most probable author of a text excerpt.

The rest of this article is divided as follows. Section 2 presents related work, and Section 3 outlines the main characteristics of the corpora used in our experiments. Section 4 briefly describes the term selection functions applied in our experiments. Section 5 presents the selected attribution methods, and Section 6 evaluates them according to various term selection strategies. Section 7 exposes some practical considerations, and Section 8 draws the main conclusions of this study.

2 Related Work

As far as automatic authorship attribution approaches are concerned, the early solutions were based on a unitary stylometric value that must be constant for a given author but should vary from one writer to another (Holmes, 1998). As measures, previous studies have suggested using vocabulary richness measures, average word length, mean sentence length, and Yule's K measure or other statistics related to type-token ratios (Baayen, 2008). None of these measures has been proven satisfactory in all cases (Hoover, 2003), due in part to word distributions (including word bigrams or trigrams) ruled by a large number of low probability elements (Large Number of Rare Events) (Baayen, 2001). Moreover, these measures are based on a single measurement, and therefore no term selection procedure was needed.

To account for the vocabulary used, Mosteller and Wallace (1964) propose a semiautomatic selection procedure to determine the most useful terms, and particularly the most frequent ones composed mainly of various function words (determiners, prepositions, conjunctions, pronouns, and some adverbs and verbal forms). In this case, we state that the occurrence frequency of some word types is not fully controlled by the author and varies from one person to the other. For example, Mosteller and Wallace (1964) notice that the term 'while' was used 36 times by Hamilton but never by Madison, the second possible writer. In their last study, these researchers have worked with a reduced list to 35 word types.

Following this perspective, Burrows (2002) suggests automatically selecting words that can discriminate between authors. The selection criterion is simply the occurrence frequency, and Burrows (2002) proposes to consider the first 40–150 most frequent word types. In such a sample, we usually find a large proportion of function words. This threshold was first increased to 800, and then to 4,000 (Hoover, 2007). As a variant, Jockers and Witten (2010) derive 2,907 terms (single words and bigrams of words) appearing at least once in texts written by all three possible authors of the 85 'Federalist Papers'. From this list, the researchers extract a reduced set composed of 298 terms, after imposing the condition that for each item the relative frequency must be >0.05%. Various studies have followed this vein suggesting using frequent word types containing many functional words (Damerau, 1975; Holmes and Forsyth, 1995; Baayen and Halteren, 2002; Miranda García and Calle Martín, 2007). In a similar way, Grieve (2007) considers selecting all word types in a k-limit profile, where k indicates that each selected term must occur, at least in k articles written by every author (e.g. a value $k = 5$ imposes the presence of the target word in at least five articles written by every possible author). Thus, this scheme imposes that all selected terms must be used by all authors and not only a fraction of them.

Instead of selecting the features based on the available corpus, Zhao and Zobel (2007) propose to define a priori the most useful word types. Their suggested list contains 363 English word types, composed mainly of function words but with some lexical terms (independent of the thematic of the underlying texts).

Finally, other studies suggest applying approaches used in automatic text categorization (or text classification) defined as the task to automatically assign one (or more) predefined label(s) to each input text (Manning and Schütze, 1999; Sebastiani, 2002; Stamatatos, 2009). To build such a system, we need to apply a feature selection procedure to reduce the number of terms needed to discriminate between the different categories. Having fewer terms, the underlying computation can be simplified, and the lexical space needed to be explored is also reduced (Liu and Motoda, 2008). In a comparative study, Yang and Pedersen (1997) evaluated six selection measures for topical text classification, using two corpora and two classifiers (*k*-nearest neighbors and linear least squares fit). Their experiments indicate that the information gain (also called expected mutual information) or the chi-square statistic tends to achieve the best results. For Sebastiani (2002), the odds ratio (OR) and the chi-square are usually the selection functions displaying the best performance.

Topical text categorization and authorship attribution do not, however, strictly follow the same implementation. In the former, we usually remove the most frequent words (stopword list) having no precise and useful meaning. In authorship attribution, these terms are viewed as important style markers because they are used in a less conscious way than other words (Pennebaker, 2011). Thus, their use and occurrence frequency may differ from one author to the other. Moreover, topical text categorization must deal with sparse data because many terms appear only in a few documents. Therefore, numerous synonyms must be taken into account to achieve a high effectiveness. In authorship attribution, we tend to ground the classification decision on frequent terms, thus reducing the problems related to the synonymy and data sparseness.

The main objective of this article is to know if, from the set of all possible terms and punctuation symbols, we can automatically select a smaller pertinent set of terms that can discriminate among authors. This objective can be achieved by automatically ignoring noisy terms. Those terms are irrelevant in discriminating between the possible authors because their occurrence distributions are similar among them. Such noisy terms must be ignored, and their removal might improve the classification accuracy. Moreover, working with a reduced set of terms will speed up the underlying computation and decrease the risk of over fitting the classifier to the available data (Hastie *et al.*, 2009).

3 Evaluation Corpora

To obtain a replicable test collection with authors sharing a common culture and having similar language registers, we opt for a stable and publicly available corpus by pulling out a subset of the CLEF-2003 test suite. The first two collections are written in the English language and correspond to articles appearing in 1995 in the ‘Glasgow Herald’ newspaper. From this news source, we have chosen articles covering two distinct topics, namely, ‘Sports’ and ‘Politics’. For each set, we have selected five journalists having written 1,948 articles on Sports and 987 on Politics. The distribution over authors is depicted in Table 1.

To complement these first two collections, we extracted two additional corpora based on articles appearing in 1994 in the ‘La Stampa’ newspaper (written in the Italian language). As with the first newspaper, we also selected articles covering the categories ‘Sports’ and ‘Politics’. For the Sports subset, we chose four journalists having written 1,321 articles, whereas for the Politics corpus, we had ten columnists having written 2,036 articles. Table 2 shows the distribution over the authors.

These corpora are pertinent for authorship attribution because each collection is formed by texts having the same general topic and genre. Moreover, they originate from the same period. We know that the style may differ from one person to another, but the period (Juola, 2003; Hughes *et al.*, 2012), the topic, the genre (Labbé, 2007), and the text intent also have an obvious impact on the style. Finally, their spelling was controlled and normalized (e.g. to denote the capital of the People’s Republic of China, we can name either Beijing or Peking).

To speed up the computation and derive an effective feature set, we ignored all terms appearing less than ten times in a given corpus and terms appearing only in a single article. Moreover, we also removed terms used only by a single author. Such

Table 1 Distribution of 1,948 articles about Sports and 987 articles about Politics by author name in the 'Glasgow Herald'

	Name	Topics	Number
1	Douglas Derek	Sports	411
2	Gallacher Ken	Sports	409
3	Gillon Doug	Sports	369
4	Paul Ian	Sports	419
5	Traynor James	Sports	340
Total			1,948
1	Johnstone Anne	Politics	73
2	Shields Tom	Politics	174
3	Smith Graeme	Politics	330
4	Trotter Stuart	Politics	337
5	Wishart Ruth	Politics	73
Total			987

Table 2 Distribution of 1,321 articles about Sports and 2,036 articles about Politics by author in 'La Stampa'

	Name	Topics	Number
1	Ansaldo Marco	Sports	288
2	Beccantini Roberto	Sports	365
3	Del Buono Oreste	Sports	435
4	Ormezzano Gian Paolo	Sports	232
Total			1,321
1	Battista Pierluigi	Politics	232
2	Benedetto Enrico	Politics	253
3	Galvano Fabio	Politics	348
4	Gramellini Massimo	Politics	119
5	Meli Mari Teresa	Politics	216
6	Nirenstein Fiama	Politics	53
7	Novazio Emanuele	Politics	250
8	Pantarelli Franco	Politics	203
9	Passarini Paolo	Politics	304
10	Spinelli Barbara	Politics	58
Total			2,036

words may be good indicators of the real author, but they are also easy to use by another person aiming to play a masquerade. For example, when analyzing the 'Federalist Papers', this filter will remove the term 'while' used frequently by Hamilton but never by Madison, as well as the term 'whilst' used only by Madison.

With the 'Glasgow Herald', after applying these constraints, the remaining vocabulary contains 6,616 word types for the Sports corpus and 5,128 terms for the political domain. In the 'La Stampa' corpora, the Sports subset comprises 6,780 word types, and the Politics subset is composed of 10,644 terms. The question that then arises is the following: can we extract from these sets of terms subsets having better discrimination power to enhance the classification performance?

4 Selection Functions

In the machine learning domain, we can find different independent feature-scoring functions to rank the features according to their discriminative power. To measure this capability for a term t_k according to a given category (or author) c_j, with $j = 1, 2, \ldots, |C|$, we usually use a contingency table for each pair (t_k, c_j) as depicted in Table 3. In this table, the value a indicates the number of texts belonging to the category c_j in which the term t_k occurs. When considering all other classes (denoted by $-c_j$), the term t_k appears in b other texts. Thus, in the whole corpus, this term occurs in $a + b$ texts, while we can count $a + c$ texts labeled with the category c_j.

To measure the association between a term t_k and a category (or author) c_j, we can compute the 'pointwise mutual information' (PMI) given in Equation (1) (Church and Hanks, 1989).

$$\begin{aligned} \mathrm{PMI}(t_k,c_j) &= \log_2\left[\frac{\mathrm{Prob}[t_k,c_j]}{\mathrm{Prob}[t_k]\cdot\mathrm{Prob}[c_j]}\right] \\ &= \log_2\left[\frac{{}^{a}/_{n}}{(a+b)/_{n}\cdot(a+c)/_{n}}\right] \end{aligned} \qquad (1)$$

This function compares two models to estimate the probability of selecting the term t_k within the category c_j. The first model is based on a direct estimation of the joint probability (and denoted $\mathrm{Prob}[t_k, c_j] = a\ /\ n$). This estimation is the numerator of Equation (1). The second model (denominator of Equation (1)) estimates this probability by considering independently the probability of the occurrence of the term t_k ($\mathrm{Prob}[t_k] = (a + b)\ /\ n$), and the probability of selecting a text belonging to the category c_j ($\mathrm{Prob}[c_j] = (a + c)\ /\ n$). This second model assumes that there is no relationship between the occurrence of the term t_k and the category c_j. When this assumption is true (no real relationship

Table 3 Example of a contingency table for a term t_k and a category c_j

	Category c_j	Category $-c_j$	
Term t_k	a	b	$a+b$
Other $-t_k$	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Table 4 Contingency table for the term Ω and the author A_j

	A_j	Other authors $-A_j$	
Term Ω	10	11	21
Other -Ω	190	1,789	1,979
	200	1,800	2,000

between the category c_j and the term t_k), Prob[t_k, c_j] can be estimated by Prob[t_k] · Prob[c_j]. In such cases, the two probability estimates will be close, and the ratio in Equation (1) will return a value close to 1. Computing the logarithm of such a value, we will find a value close to 0, indicating independence between the term occurrence and the corresponding category.

On the other hand, when a strong association does exist between the term t_k and the category c_j, the value of a will be large. The direct estimation for Prob[t_k, c_j] will be larger than the product Prob[t_k] · Prob[c_j]. The ratio will then be larger than 1 and the logarithm function will return a positive value. With a negative association between the term t_k and the category c_j, the numerator will be smaller than the denominator, returning a value smaller than 1. Taking the logarithm, a negative value is returned, indicating that the term t_k is less frequently used in category c_j than in the rest of the corpus.

To illustrate this idea, we have taken a hypothetical numerical example given in Table 4. In this case, the term Ω appears in twenty-one texts in the corpus, in which we can find ten texts written by author A_j. The other authors have used the word Ω in eleven other texts. The last row in Table 4 indicates that the author A_j has written 200 texts, while we can count 1,800 texts written by all other authors. In other words, A_j has written 10% of all the texts belonging to the corpus.

A quick analysis reveals that the author A_j represents ~50% of the occurrences of the term Ω, but only 10% of all texts. When computing the PMI value (formulation given below), we can see that the joint estimation is larger than the denominator. The resulting ratio between the two models is larger than 1 (4.76 in our example), giving a final value of 2.25. The term Ω is more closely associated to A_j than by pure chance. Because the presence of this term in a text is an indication that this document might have been written by A_j, we can select this term to help discrimination between A_j and all other possible writers.

$$\begin{aligned}\mathrm{PMI}(\Omega,A_j) &= \log_2\left[\frac{{}^{10}\!/_{2000}}{(10+11)/_{2000}\cdot(10+190)/_{2000}}\right]\\ &= \log_2\left[\frac{10\cdot 2000}{21\cdot 200}\right] = \log_2[4.76] = 2.25\end{aligned}$$

On the other hand, if the value for a in Table 4 had been 2 instead of 10, and keeping the same values for the last row and the last column, we would find no relationship between the term Ω and the author A_j. Observing two texts written by A_j with the term Ω corresponds closely to the independent model (pure chance). Because A_j wrote 10% of the whole corpus, it is not surprising to see close to 10% of them with the term Ω. In this case, the PMI function will return the value −0.07, a value close to 0.

Using such a term selection function seems a pertinent choice. Such a procedure may reveal the terms useful in discriminating between the possible authors. Moreover, this approach may rank all terms according to their discriminative power and we can limit the selection of the top m most discriminative terms.

As a second function, we can estimate the probability Prob[c_j | t_k], a measure denoted Darmstadt Indexing Approach (DIA) (Fuhr *et al.*, 1991). Based on Table 3, this probability is estimated by a / ($a+b$) (to simplify the presentation, all formulae are re-grouped in the Appendix). The DIA function is based on different arguments than those justifying the PMI function.

As a third function, we can use the odds ratio (OR) (Manning and Schütze, 1999), which always returns a positive value. A positive association between the term t_k and the category c_j is indicated by a value larger than 1, although a value close to 0 signifies an opposition. A value close to 1 denotes independence between the term and the underlying class.

As a fourth selection function, we can use the information gain (IG) or the expected mutual information. The value returned by this function is large if a positive association exists. A small positive value signifies the absence of a discriminative power for the term t_k and the category c_j. Following the same interpretation, we can compute the chi-square, $\chi^2(t_k, c_j)$, statistics (Manning and Schütze, 1999).

As a sixth function, we can apply the gain ratio (GR) returning a positive value to signal either a positive or negative association between the term t_k and the category c_j. Independence is indicated by a value close to 0.

As a seventh term selection function, derived from the χ^2, we can provide the correlation coefficient, CC(t_k, c_j) (Ng *et al.*, 1997), which indicates a positive association by a positive value (although a negative value signifies an opposition). The independence between the term and the category is denoted by a value close to 0. Following the same interpretation, we can compute the GSS coefficient (Gavalotti *et al.*, 2000).

Another interesting selection function is the bi-normal separation (BNS) suggested by Forman (2003). In the presence of independence, this function returns a small positive value. A larger value indicates either a positive or a negative association between the term t_k and the underlying category c_j.

In addition to these nine selection functions, we can simply consider the document frequency (*df*) indicating the number of texts indexed by the term t_k. The larger this value, the better the corresponding term. Moreover, we can also assume that the style of a given author may be revealed by the frequent use of certain forms. In this perspective, we can follow Burrows (2002) and use the absolute term frequency (*tf*). As for the *df* value, the higher the absolute term frequency (or *tf*), the better the usefulness of this term.

When applying one of the aforementioned selection functions, we can compute a local utility value denoted f(t_k, c_j) for each t_k and each category c_j. When faced with a binary classification problem (two authors), such a function is enough to define the overall selective value for each term. In authorship attribution in general, the number of authors (categories) is larger than two. In such cases, we need to aggregate the local utility values over the |C| categories. To define such a global utility measure for a term t_k (denoted $U_{op}(t_k)$), we can take the maximum over the |C| categories or compute the sum or a weighted mean as shown in Equation (2).

$$\begin{aligned} U_{max}(t_k) &= Max_j f(t_k, c_j), \\ U_{\mathrm{sum}}(t_k) &= \sum_{j=1}^{|C|} f(t_k, c_j), \\ U_{\mathrm{wmean}}(t_k) &= \sum_{j=1}^{|C|} \mathrm{Prob}[c_j] \cdot f(t_k, c_j) \end{aligned} \tag{2}$$

Finally, to select the *m* most adequate terms, we extract the *m* terms having the highest utility values U$_{op}(t_k)$ according to one of the aggregate operators given above (max, sum, or weighted mean).

5 Authorship Attribution Methods

To evaluate the different term selection functions, we have selected three authorship attribution approaches to ground our finding on a relatively broad basis. As a first method, we can apply the approach suggested by Zhao and Zobel (2007), who propose to compute the distance between the author profile A_j (concatenation of all his/her writings) and the query text *Q* by using the Kullback–Leibler divergence (KLD) (also called relative entropy (Manning and Schütze, 1999)). This measure is given in Equation (3), where $\mathrm{Prob}_q[t_i]$ (or $\mathrm{Prob}_{Aj}[t_i]$) indicates the occurrence probability of a term t_i in the query *Q* (or in the author profile A_j), for $i = 1, 2, \ldots, m$.

$$KLD(Q||A_j) = \sum_{i=1}^{m} \mathrm{Prob}_q[t_i] \cdot \log_2 \left[\frac{\mathrm{Prob}_q[t_i]}{\mathrm{Prob}_{Aj}[t_i]} \right] \tag{3}$$

When two distributions are identical, the KLD measure is 0. Otherwise, the formula returns a positive value. This value grows as the distance (disagreement) increases between the two underlying distributions.

To estimate the needed probabilities, we can apply the maximum likelihood principle and access Prob$[t_i] = tf_i/n$, with tf_i indicating the occurrence frequency of a term t_i, and n the size of the document. However, it is usually better to smooth such estimates to avoid null probabilities (Manning and Schütze, 1999). In our evaluations, we have applied the Lidstone's rule where the probabilities are then estimated as $(tf_i + \lambda)$ / $(n + \lambda \cdot |V|)$, with $|V|$ indicating the vocabulary size and λ a parameter fixed to 0.01 (showing the best performance).

As a second authorship attribution method, we can compare the representation of a given text Q with an author profile A_j using the chi-square statistic (Grieve, 2007) defined by Equation (4) (the same general method can be used as term selection and attribution scheme). In this formulation, $rtf_q(t_i)$ represents the relative frequency of the ith term in the query text, and $rtf_{Aj}(t_i)$ the same information in the jth author profile. When comparing a query text Q with different author profiles A_j, we simply select the lowest chi-square value to determine the most probable author of a disputed text.

$$\chi^2(Q, A_j) = \sum_{i=1}^{m} \left(rtf_q(t_i) - rtf_{A_j}(t_i)\right)^2 \Big/ rtf_{A_j}(t_i) \quad (4)$$

As a third authorship attribution method, we used the Delta model (Burrows, 2002) measuring the distance between two texts according to the standardized frequency (Z score) of their terms. This value is obtained from the relative occurrence frequency (denoted rtf_{ij} for term t_i in the document d_j) by subtracting the mean ($mean_i$) and dividing by the standard deviation (sd_i), the mean and standard deviation estimated by considering the underlying corpus (see Equation (5)).

$$Z\ score(t_{ij}) = \frac{rtf_{ij} - mean_i}{sd_i} \quad (5)$$

Once these dimensionless quantities are obtained for each selected word, 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 $i = 1, 2, \ldots, m$, we compute the Delta value (or the intertextual distance) by applying Equation (6).

$$\Delta(Q, A_j) = 1/\mathrm{m} \cdot \sum_{i=1}^{m} \left| Z\ score(t_{iq}) - Z\ score(t_{ij}) \right| \quad (6)$$

In this formulation, proposed by Burrows (2002), we assign the same weight to each term t_i. A large difference between Q and A_j will appear, when, for a given term, both Z scores are large but with opposite signs. On the other hand, when a term appears with similar relative occurrence frequencies in both texts, the difference in Z scores will be small. Finally, when for all m terms the differences in Z scores are small, the resulting Δ distance will be slight, indicating that the same person probably wrote both texts.

6 Evaluation

To achieve unbiased performance estimations, we cannot use the same instances for both training the classifier and testing it. The set of available examples must, therefore, be divided into a training set and a distinct test set. In the current study, we opted for the leave-one-out approach (Hastie *et al.*, 2009). When applying this methodology with the Sports corpus extracted from the 'Glasgow Herald', each of the 1,948 articles, in turn, will form the query text, whereas the remaining 1,947 texts will generate the training set used to determine the most useful terms. The accuracy rate reported in this study corresponds to the micro-average value, the mean over all documents.

Some authors suggest not using the same training set to let the classifier learn (i.e. the author profiles) and to select the features. Thus, it is recommended to use a disjoint set of instances for feature selection and for learning. If from a theoretical point of view a bias exists, from a practical viewpoint the impact of this bias is rather limited (Singhi and Liu, 2006).

As a first authorship attribution model, we have evaluated the KLD model (Zhao and Zobel, 2007). With the 'Glasgow Herald' corpus, the feature selection is based on 363 predefined word types. This list mainly contains function words (the, in, but, not, am, of, can . . .) and frequent items (became, nothing . . .). Some entries are less frequent (howbeit, whereafter, whereupon), whereas others indicate the expected behavior of the tokenizer (doesn, weren) or correspond to an arbitrary decision (indicate, missing, seemed). With the Italian corpora, we first need to define a list of frequent terms usually appearing in all documents. To achieve this, we have chosen a stopword list used by search technology with this language (Savoy, 2001). This list includes 399 terms containing mainly function words (il, la, del, in, con, nostro, essi, fare . . .) and frequent items (anno, casa . . .).

Using the KLD method with the predefined set of 363 English words, we achieved an accuracy rate of 83.5% for the Sports corpus and 81.0% for Politics (these values are reported in the line 'a priori' in Table 5). To obtain a more complete picture, we also considered all available terms (words and punctuation symbols, no selection). In this case, the KLD scheme produces an accuracy rate of 74.6% (Sports, based on 6,616 terms) and 93.0% (Politics, based on 5,128 terms) (line denoted 'All' in Table 5).

With the Italian language and the predefined set of 399 words, we achieved an accuracy rate of 94.9% for the Sports subset and 88.7% for Politics (these performances appear in the line 'a priori' in Table 6). When considering all terms without any selection, an accuracy rate of 97% was obtained with the Sports subset (based on 6,780 terms) and 88.4% with the Politics subset (based on 10,644 terms) (line denoted 'All' in Table 6).

We then compare these baselines (a priori selection) with eleven other term selection methods and three aggregation operators (max, sum, or weighted mean (denoted wmean)). As the number of selected terms (parameter m in the previous formulae depicted in Section 5), we have tested the following values {150; 300; 500; 800; 1,000; 1,500; 2,000; 3,000; 4,000; and 5,000}. The last value (5,000) corresponds to ∼75.6% of all 6,616 terms available for the Sports subset in the 'Glasgow Herald' (or 97.5% of all 5,128 available terms for the Politics subset). Similar percentages can be obtained when analyzing the two Italian text collections. When considering larger numbers, the term space is not really reduced; therefore, we did not attempt this variation. Instead of reporting all possible combinations of the number of features with the three aggregation functions, we have only reported the best parameter setting for each selection function (number of terms / aggregation operator).

Table 5 Feature selection methods with KLD and 'Glasgow Herald', with the Sports corpus and Politics

Function	Sports		Politics	
	Parameter	Accuracy (%)	Parameter	Accuracy (%)
a priori	363 terms	83.5	363 terms	81.0
All	6,616 terms	74.6†	5,128 terms	93.0†
$tf(t_k, c_j)$	3,000 / max	92.8†	800 / max	95.1†
$df(t_k, c_j)$	3,000 / max	**93.5**†	2,000 / max	95.1†
IG(t_k, c_j)	1,500 / wmean	93.3†	800 / sum	95.0†
GR(t_k, c_j)	3,000 / max	93.1†	500 / max	**96.1**†
GSS(t_k, c_j)	3,000 / max	92.7†	1,000 / max	95.1†
$\chi^2(t_k, c_j)$	1,500 / sum	93.3†	300 / sum	95.0†
CC(t_k, c_j)	2,000 / max	92.8†	300 / max	95.4†
BNS(t_k, c_j)	5,000 / wmean	90.7†	1,500 / sum	94.4†
PMI(t_k, c_j)	5,000 / wmean	90.3†	5,000 / wmean	92.2†
OR(t_k, c_j)	3,000 / wmean	91.5†	500 / wmean	94.1†
DIA(t_k, c_j)	3,000 / max	86.4†	5,000 / wmean	93.1†

Table 6 Feature selection methods with KLD and 'La Stampa', with the Sports collection and Politics

Function	Sports		Politics	
	Parameter	Accuracy (%)	Parameter	Accuracy (%)
a priori	399 terms	94.9	399 terms	88.7
All	6,780 terms	97.0†	10,644 terms	88.4
$tf(t_k, c_j)$	5,000 / max	97.4†	500 / sum	94.4†
$df(t_k, c_j)$	5,000 / max	**97.7**†	500 / max	**95.6**†
IG(t_k, c_j)	5,000 / max	**97.7**†	4,000 / wmean	95.3†
GR(t_k, c_j)	4,000 / sum	97.3†	3,000 / wmean	95.2†
GSS(t_k, c_j)	5,000 / max	97.0†	500 / max	95.1†
$\chi^2(t_k, c_j)$	4,000 / max	97.3†	3,000 / wmean	94.9†
CC(t_k, c_j)	5,000 / max	97.0†	1,500 / max	94.0†
BNS(t_k, c_j)	2,000 / max	97.3†	3,000 / wmean	93.8†
PMI(t_k, c_j)	4,000 / max	96.2	5,000 / max	87.3
OR(t_k, c_j)	5,000 / max	96.4	1,000 / wmean	93.0†
DIA(t_k, c_j)	4,000 / max	95.5	500 / max	90.1

The question that then arises is 'can we obtain a better performance using fewer terms?' If yes, can the set of terms defined by Zhoa and Zobel (2007) produce a better performance than that produced from sets of terms defined by the various feature selection functions?

The accuracy rates depicted in Tables 5 and 6 indicate that different selection functions may produce better performance levels than either the manual selection or when ignoring the selection procedure (lines labeled 'All'). The manual selection (lines labeled 'a priori') produces relatively low accuracy rates for the English language compared with the others (see Table 5). For the Politics subset of the 'Glasgow Herald', considering all possible terms clearly improves the performance (from 81.0 to 93.0%). Within the same category, but with the Italian language (Table 6), the manual selection, or considering all terms, tends to produce similar performance levels (88.7 versus 88.4%). However, those accuracy rates are lower than those achieved by other selection functions.

Overall, Tables 5 and 6 tend to show that we can achieve high-performance results when considering relatively simple selection methods such as *df*, or the absolute *tf*. In these tables, the best performance is shown in bold. The performance differences with IG, GR, GSS, chi-square (χ^2), or CC functions are usually small and not significant. However, the PMI, OR, and DIA functions seem to offer lower performance levels than those produced by other selection schemes. When inspecting the different aggregate operators (max, sum, or weighted mean), we can see that the maximum function tends to occur frequently in Tables 5 and 6, indicating that this aggregation function tends to achieve the best results.

To verify whether a performance difference is statistically significant between two term selection procedures, we opted for the Sign test (Conover, 1980; Yang and Liu, 1999) (bilateral test, significance level $\alpha = 1\%$). In this case, the null hypothesis H_0 assumes that both selection methods perform at a similar level. In Tables 5–10, we use the first line as a baseline, and any statistically significant performance difference is indicated by the symbol '†'. As we can see in Tables 5 and 6, the performance differences are usually significant compared with the first row, the selection strategy based on a predefined set of words.

When applying the chi-square metric (Grieve, 2007), we have tried different *k*-limits and found that for the 'Glasgow Herald' the 5-limit produces the best performance (83.9% as depicted in Table 7 in the row '*k*-limit') by selecting 1,633 terms for the Sports corpus (or with $k = 10$ for the Politics subset, selecting 434 terms and producing an accuracy of 76.0%). When specifying *k*-limit = 10, all selected terms must appear in at least ten articles written by all journalists. Thus, such a selection strategy imposes that every possible author must have used all selected terms.

With the two Italian corpora, the best accuracy rates were achieved when considering the 200-limit (selecting a small set of thirty-one terms, accuracy rate = 83.7%) for the Sports subset or with $k = 10$ for the Politics subset (selecting 252 terms and producing an accuracy of 71.9%).

When ignoring the selection procedure, we take into account all possible terms. In this case, we can achieve an accuracy rate of 72.5% (Sports, 6,616 available terms) and 79.5% (Politics, based on 5,128 terms) with the English corpora (see Table 7, line with the label 'All'). In Table 8, for the Italian collections, the accuracy rate is of 85.5% (Sports, 6,780 available terms) or 72.6% (Politics, based on 10,644 terms) when we ignore the selection procedure.

Instead of strictly following the selection scheme proposed by Grieve (2007), we can apply different feature-scoring selection functions. The best parameter settings and accuracy rates are reported in Table 7 for the 'Glasgow Herald' newspaper and in Table 8 for the Italian corpora.

As for the KLD method, the results depicted in Tables 7 and 8 indicate that an appropriate feature-scoring function (with their parameter values) might produce higher performance levels than when considering all terms or when selecting terms according to the best *k*-limit principle. Overall, when comparing the different selection strategies, the *df*, or *tf* selection schemes tend to produce high-performance levels. With this chi-square-based attribution scheme only, the BNS selection function also offers a high effectiveness. After applying the Sign-test, we can see that the performance differences with the best *k*-limit approaches are usually statistically significant (indicated with the symbol '†'). Finally, and for both

Table 7 Feature selection methods with chi-square method and the 'Glasgow Herald', with the Sports corpus and Politics

Function	Sports		Politics	
	Parameter	Accuracy (%)	Parameter	Accuracy (%)
k-limit	1,633 terms	**83.9**	434 terms	76.0
All	6,616 terms	72.5†	5,128 terms	79.5
$tf(t_k, c_j)$	500 / max	79.6†	300 / sum	**90.4†**
$df(t_k, c_j)$	500 / sum	83.6	800 / sum	90.3†
$\mathrm{IG}(t_k, c_j)$	4,000 / sum	75.3†	150 / wmean	85.1†
$\mathrm{GR}(t_k, c_j)$	4,000 / sum	74.9†	150 / sum	84.0†
$\mathrm{GSS}(t_k, c_j)$	5,000 / max	73.8†	150 / max	87.9†
$\chi^2(t_k, c_j)$	4,000 / sum	74.1†	150 / wmean	82.7†
$\mathrm{CC}(t_k, c_j)$	5,000 / max	73.5†	5,000 / max	79.5
$\mathrm{BNS}(t_k, c_j)$	3,000 / sum	77.6†	2,000 / max	86.5†
$\mathrm{PMI}(t_k, c_j)$	4,000 / max	74.1†	3,000 / sum	24.5†
$\mathrm{OR}(t_k, c_j)$	2,000 / max	74.8†	4,000 / wmean	79.7†
$\mathrm{DIA}(t_k, c_j)$	5,000 / max	73.2†	500 / max	48.0†

Table 8 Feature selection methods with chi-square method and 'La Stampa', with the Sports collection and Politics

Function	Sports		Politics	
	Parameter	Accuracy (%)	Parameter	Accuracy (%)
k-limit	31 terms	83.7	252 terms	71.9
All	6,780 terms	85.5	10,644 terms	72.6
$tf(t_k, c_j)$	150 / sum	90.1†	300 / wmean	89.5†
$df(t_k, c_j)$	500 / wmean	91.9†	500 / wmean	**92.8†**
$\mathrm{IG}(t_k, c_j)$	5,000 / max	86.3	150 / wmean	78.9†
$\mathrm{GR}(t_k, c_j)$	5,000 / max	89.3†	150 / wmean	73.4
$\mathrm{GSS}(t_k, c_j)$	5,000 / max	87.9†	300 / max	78.2†
$\chi^2(t_k, c_j)$	5,000 / sum	85.9	5,000 / wmean	73.0
$\mathrm{CC}(t_k, c_j)$	5,000 / max	86.5	5,000 / max	67.9†
$\mathrm{BNS}(t_k, c_j)$	3,000 / max	**93.1†**	2,000 / max	81.8†
$\mathrm{PMI}(t_k, c_j)$	5,000 / max	87.0	5,000 / max	69.7
$\mathrm{OR}(t_k, c_j)$	5,000 / max	87.9†	4,000 / max	62.2†
$\mathrm{DIA}(t_k, c_j)$	5,000 / max	79.7†	5,000 / max	50.3†

Table 9 Feature selection methods with Delta method and 'Glasgow Herald', with the Sports corpus and Politics

Function	Sports		Politics	
	Parameter	Accuracy (%)	Parameter	Accuracy (%)
Most freq.	300 terms	74.2	200 terms	**83.5**
All	6,616 terms	21.1†	5,128 terms	43.0†
$tf(t_k, c_j)$	300 / max	81.1†	150 / sum	83.0
$df(t_k, c_j)$	300 / max	78.7†	150 / wmean	81.8
$\mathrm{IG}(t_k, c_j)$	500 / sum	83.4†	150 / sum	69.1†
$\mathrm{GR}(t_k, c_j)$	500 / wmean	**85.6†**	150 / sum	72.0†
$\mathrm{GSS}(t_k, c_j)$	300 / max	79.8†	2,000 / max	72.1†
$\chi^2(t_k, c_j)$	1,000 / sum	74.2	4,000 / wmean	66.1†
$\mathrm{CC}(t_k, c_j)$	300 / sum	42.2†	2,000 / max	63.8†
$\mathrm{BNS}(t_k, c_j)$	1,500 / max	51.0†	1,500 / sum	62.7†
$\mathrm{PMI}(t_k, c_j)$	5,000 / max	56.4†	1,500 / max	52.2†
$\mathrm{OR}(t_k, c_j)$	150 / max	21.0†	4,000 / wmean	74.4†
$\mathrm{DIA}(t_k, c_j)$	800 / max	54.5†	1,000 / max	64.8†

Table 10 Feature selection methods with Delta method and 'La Stampa', with the Sports collection and Politics

Function	Sports		Politics	
	Parameter	Accuracy (%)	Parameter	Accuracy (%)
Most freq.	150 terms	85.7	500 terms	84.1
All	6,780 terms	29.8†	10,644 terms	15.5†
$tf(t_k, c_j)$	300 / max	**93.3†**	150 / max	85.1
$df(t_k, c_j)$	300 / max	**93.3†**	150 / max	84.9
$\mathrm{IG}(t_k, c_j)$	3,000 / max	80.7†	150 / max	87.9†
$\mathrm{GR}(t_k, c_j)$	4,000 / max	93.0†	150 / sum	**88.7†**
$\mathrm{GSS}(t_k, c_j)$	500 / max	76.4†	150 / max	82.6
$\chi^2(t_k, c_j)$	4,000 / max	81.2†	150 / sum	78.6†
$\mathrm{CC}(t_k, c_j)$	500 / max	70.0†	150 / max	68.5†
$\mathrm{BNS}(t_k, c_j)$	2,000 / wmean	62.5†	300 / wmean	28.0†
$\mathrm{PMI}(t_k, c_j)$	3,000 / max	62.6†	800 / sum	32.9†
$\mathrm{OR}(t_k, c_j)$	5,000 / sum	67.5†	300 / sum	32.8†
$\mathrm{DIA}(t_k, c_j)$	4,000 / max	60.0†	150 / max	28.2†

languages, usually the PMI, the OR, and the DIA functions tend to return less pertinent term sets, achieving lower performance levels.

To obtain a broader view, we have also depicted the best performances achieved with the Delta rule (Burrows, 2002). As depicted in Table 9 for the 'Glasgow Herald', the best performance using the Delta rule is achieved when considering the 300 most frequent terms for the Sports subset (accuracy rate 74.2%, under the label 'Most freq.') or 200 with the Politics corpus (83.5%). With the Italian text collections, the most effective number of terms is 150 for the Sports subset (85.7%) or 500 word types with the Politics corpus (84.1%) (see Table 10). This selection procedure proposed with the Delta rule is equivalent to the *tf* scoring function with the sum aggregation.

Without any feature selection, the Delta rule produces, with the English corpus, an accuracy of 21.1% (Sports, 6,616 available terms) and 43.0% (Politics, 5,128 terms), as depicted in the line 'All' in Table 9. With the 'La Stampa' newspaper's corpora and considering all terms, the Delta rule achieves an accuracy of 29.8% (Sports, 6,780 terms) and 15.5% (Politics, with 10,644 terms). The performance differences with the first row are rather large, indicating that the Delta rule must be applied with a reduced number of word types. As with the other authorship attribution methods, we have also reported the best success rate according to the eleven selection approaches together with the best parameter setting.

Overall, the evaluations depicted for the Delta rule (see Tables 9 and 10) confirm that high-performance levels are achieved when using the *df*, or the absolute *tf* as feature selection functions. The performance differences are usually statistically significant over the selection based on the most frequent words, but only for the two Sports corpora. As a second choice, we can use the GSS, IG, and chi-square functions. However, the OR, the PMI, and the DIA function tend to produce less pertinent feature sets, at least in an authorship attribution context and especially when using the Delta rule (see Tables 9 and 10).

A few cases are worth a comment. With the English language (Table 9), the high result of the GR function in the Sports subset is not confirmed by the Politics corpus. We also notice that the correlation coefficient (CC) function with the English Sports corpus (see Table 9) achieves a rather low accuracy level, as does the BNS selection function with the Italian Politics subset.

Finally, to give a view of the selection effect of the different functions, we have counted the percentage of selected terms in common between two functions with the English corpora. Using only the sum as the aggregate operator, and varying the number of selected terms between 150 and 3,000, we can see that the functions *df* and *tf* return, on an average, similar sets of terms (overlap degree between 92 to 99%). A similar effect can be detected with the function CC and the chi-square metric (this can be explained by the fact that the function CC is derived from the chi-square) or between the IG and GR. We can also detect a relationship between the set of features defined by the functions GSS and IG. In these cases, the overlay is, on an average, 77%. Finally, it is difficult to find a clear relationship between the BNS, OR, or PMI functions and all the others. These three selection functions tend to propose different sets of features.

7 Practical Considerations

When faced with a new authorship attribution problem, which term selection function must we apply and how many terms must we select? Based on four test collections and three attribution schemes, the experimental results do not show a strong systematic pattern. However, some trends can be detected. First, the absolute occurrence frequency (*tf*) and the *df* tend to produce pertinent and well-performing term sets for the three attribution schemes and the four collections. This is an indication that using frequent words as features to discriminate between different authors is an effective strategy. Such selection approaches also have the advantage to be easy to implement and own a clear interpretation for the end-user. Moreover, we cannot detect significant differences between the evaluations done with the English language and those performed over the Italian collections.

It is worth mentioning that the term selection is not based on the whole vocabulary. As specified in Section 3, we can take into account the domain knowledge. Thus, it is a good practice to ignore word types having a low occurrence frequency or appearing in a single (or a few) document(s). Moreover, we have also removed words used by only a single author. It is known that such terms can be effective to distinguish between authors and most of the selection functions will detect them as effective features for authorship discrimination. However, these terms are vulnerable because they can be easily used to spoof a given identity.

The second important question is to define the number of terms to be selected. Determining a priori such an optimum value is difficult. The various experiments depicted in the previous section indicate that we need a small number of terms (between 150 and 300) to obtain one of the best performance levels with the Delta rule (Tables 9 and 10). The chi-square authorship attribution scheme also requires a relatively small number of terms to produce one of the best accuracy rates (300–500 (*df* function) with the English corpora (Table 7), and 500 terms (*df* function) with the Italian corpora (Table 8)). With the KLD method, a general conclusion is harder to draw. For the English corpora, we need ~500 for the Politics subset, and 1,500–3,000 terms for Sports subset (Table 5). For the Italian language, ~5,000 terms for the Sports subset and 500 for the Politics part (Table 6) are required to achieve the highest performance levels.

In addition to these findings, we must recall that the morphology of the Italian language is more complex than that of English. Thus, we can expect having more functional word types in this language than in English. For example, the translation of the determiner 'the' (definite article) could be 'il, lo, l, i, gli, la, or le' because the variations in gender and number must be specified in the Italian language. This difference in size can be reduced when we consider representing text using the lemmas (headword or dictionary entry) instead of the word types. In this case, we can conflate all inflected forms under the same entry (e.g. 'was, were, is,' etc. are regrouped under the lemma 'be', whereas the pronouns 'I, me' under 'I'). This processing can be done manually, but it is a costly operation. On the other hand, we can apply an automatic part-of-speech tagger. However, such an approach is not error-free, and some recent studies have compared the relative merits of these two text representation schemes for authorship attribution (Savoy, 2012; Miranda García and Calle Martín, 2012). Finally, we must mention that some natural languages may have other linguistic construction than those used in the English language. For example, the definite article appears as a suffix in the Bulgarian or Swedish language and not as a distinct and separate lemma. Moreover, the indefinite article ('an/a') does not exist in the Bulgarian language.

As a possible default parameter setting, we can suggest selecting the first 300 most frequent word types according to the document frequency (or df_{sum}) for the English collections, and the top 500 most frequent terms (df_{sum}) with the Italian corpora. Using the occurrence frequency (*tf*) will produce similar results, and according to our experiments, there is no real reason to prefer one function to the other. We have a slight preference for the document frequency information because this function ignores the variations in document lengths. Finally, to determine the number of terms, it is more efficient to work with a small term set. According to our experiments, a size of 300 terms seems reasonable for the English language. Considering that the morphology of the Italian language is more complex, we suggest adding 200 more terms when working with languages having a more complex morphology (i.e. gender, grammatical cases).

Table 11 depicts the accuracy rates obtained when adopting these suggested default parameter settings. For the 'Glasgow Herald' corpora, we have used the first 300 most frequent terms according to the document frequency (or df_{sum}) and the top 500 most frequent terms (df_{sum}) for 'La Stampa'. These performance levels are then compared with the optimal parameter setting (considering all selection functions and number of terms).

As shown in Table 11, we can see that the performance differences between the proposed default parameter settings and the optimal ones are rather small, particularly with the KLD attribution scheme. On the other hand, the Delta rule is more sensitive to deviation from an optimal number of terms.

These data also show that the *df* selection strategy tends to produce similar term sets when compared with the *tf* function. This relationship can be shown by considering the performances obtained with the Politics corpus of the 'Glasgow Herald' in Table 11. Using the chi-square attribution scheme and the 300 most frequent occurring terms (*tf*), we achieve an accuracy rate of 90.4%. Using the *df* function, the selection of the 300 most frequent terms provides a mean performance of 87.4%, a relative decrease of −3.3%.

Table 11 Evaluation of the proposed parameter setting versus the optimal one, with the English corpora and the Italian corpora

Attribution method	'Glasgow Herald'		'La Stampa'	
	Corpus—Parameter	Accuracy (%)	Corpus—Parameter	Accuracy (%)
KLD	Sports—3,000 *df* max	93.5	Sports—5,000 *df* max	97.7
	300 terms	90.0 (−4%)	500 terms	96.4 (−1%)
Chi-square	1,633 terms—no select.	83.9	3,000 BNS max	93.1
	300 terms	64.5 (−23%)	500 terms	90.2 (−3%)
Delta	500 GR wmean	85.6	300 *tf* max	93.3
	300 terms	78.7 (−8%)	500 terms	91.8 (−2%)
KLD	Politics—500 GR max / max	96.1	Politics—500 *df* max	95.6
	300 terms	94.5 (−2%)	500 terms	95.6 (0%)
Chi-square	300 *tf* sum	90.4	500 *df* wmean	92.8
	300 terms	87.4 (−3%)	500 terms	90.0 (−3%)
Delta	150 *tf* sum	83.0	150 GR sum	88.7
	300 terms	71.4 (−14%)	500 terms	73.7 (−17%)†

8 Conclusion

To design an effective authorship attribution scheme, we need to select the most appropriate features (word types and punctuation symbols in the current study) that can discriminate between the different categories or authors. To evaluate the different selection strategies, we compared nine feature-scoring functions and two well-known selection approaches used in authorship attribution studies (based on the absolute term frequency (*tf*), or the *df*). To combine the scores computed for various terms, we have evaluated three aggregation operators. As a classifier, we used the KLD measure (Zhao and Zobel, 2007), the chi-square metric (Grieve, 2007), and the Delta rule (Burrows, 2002).

Using four corpora extracted from the newspaper 'Glasgow Herald' and 'La Stampa' about Sports (1,948 and 1,321 articles, respectively) and Politics (987 and 2,036 articles, respectively), our evaluations show that using the *df* or the absolute term frequency (*tf*) tends to provide good overall performances. In a second class of performance level, we can place the GR, the chi-square, the GSS function, and the IG. The use of the PMI, the OR, or the DIA function does not provide comparable results, at least in the authorship attribution context. Finally, the BNS function presents an erratic behavior, working well in some cases, and moderately in others.

The overall good results achieved by the absolute term frequency (*tf*) are clearly an indication that the suggested selection procedure proposed by Burrows (2002) for the Delta rule is an effective one. Similarly, the *df* selection function tends to propose discriminative term sets, and this study confirms, in part, the *k*-limit selection strategy suggested by Grieve (2007). However, this latter procedure imposes that the selected terms be used by all possible authors, a constraint not imposed by the *df* selection function.

Unlike the Yang and Pedersen's study (1997) based on topical text classification, the IG or the chi-square is not always the best performing method. According to Sebastiani (2002), good selections can be achieved by applying the OR_{sum} or the GSS_{max}. The current study, which is based on authorship attribution, indicates that this choice, adequate for topical text classification, is not the best when dealing with authorship attribution.

Finally, as an aggregation function, this study tends to indicate that applying the maximum operator seems to be a good default choice. On the other hand, our evaluations based on four distinct corpora are unable to clearly define a specific number of terms to be used for a new collection. As a default range of values, the evaluations reported in this study indicate that considering between 300 and 500 terms seems a good starting point before further investigations, when possible.

References

Baayen, H. R. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Press.

Baayen, H. R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen, H. R. and Halteren, H. V. (2002). An Experiment in Authorship Attribution. In *Proceedings of the 6th Journées d'Analyse des Données Textuelles 2002*. St-Malo, pp. 69–75.

Burrows, J. F. (2002). Delta: a measure of stylistic difference and a guide to likely authorship. *Literary & Linguistic Computing*, **17**: 267–87.

Church, K. W. and Hanks, P. (1989). Word Association Norms, Mutual Information and Lexicography. In *Proceedings Association for Computational Linguistics (ACL)*. pp. 76–83.

Conover, W. J. (1980). *Practical Nonparametric Statistics*, 2nd edn. New York: John Wiley & Sons.

Damerau, F. J. (1975). The use of function word frequencies as indicators of style. *Computers and the Humanities*, **9**: 271–80.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning*, **3**: 1289–305.

Fuhr, N., Hartmann, S., Knorz, G., Lustig, G., Schwantner, M., and Tzeras, K. (1991). IR/X a Rule-Based Multi-Stage Indexing System for Large Subject Fields. In: *Proceedings Recherche d'Information Assistée par Ordinateur (RIAO)*. pp. 606–23.

Gavalotti, L., Sebastiani, F., and Simi, M. (2000). Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization. In: *Proceedings European Conference in Digital Libraries (ECDL)*. Berlin: Springer, pp. 59–68.

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

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

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, **13**: 111–7.

Holmes, D. I. and Forsyth, R. S. (1995). The *federalist* revisited: new directions in authorship attribution. *Literary and Linguistic Computing*, **10**: 111–27.

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, **37**: 151–78.

Hoover, D. L. (2007). Corpus stylistics, stylometry, and the styles of *Henry James*. *Style*, **41**: 160–89.

Hughes, J. M., Foti, N. J., Krakauer, D. C., and Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences United States of America*, **109**(20): 7682–6.

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**: 215–23.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, **37**: 77–96.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**: 233–334.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability. *Journal of Machine Learning Research*, **8**: 1261–76.

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

Liu, H. and Motoda, H. (2008). *Computational Methods of Feature Selection*. Boca Raton: Chapman & Hall / CRC.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Miranda García, M. and Calle Martín, J. (2007). Functions words in authorship attribution studies. *Literary & Linguistic Computing*, **22**: 27–47.

Miranda García, M. and Calle Martín, J. (2012). The authorship of the disputed federalist papers with an annotated corpus. *English Studies*, **93**: 371–90.

Mosteller, F. and Wallace, D. L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading: Addison-Wesley.

Ng, H. T., Goh, W. B., and Low, K. L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In: *Proceedings*

ACM-Special Interest Group in Information Retrieval (SIGIR). New York: The ACM Press, pp. 67–73.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns. What Our Words Say about Us*. New York: Bloomsbury Press.

Savoy, J. (2001). Report on CLEF-2001 Experiments. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (eds), *Cross-Language Information Retrieval and Evaluation,* LNCS #2069. Berlin: Springer, pp. 27–43.

Savoy, J. (2012). Authorship attribution: a comparative study of three text corpora and three languages. *Journal of Quantitative Linguistics*, **19**: 132–61.

Sebastiani, F. (2002). Machine learning in automatic text categorization. *ACM Computing Survey*, **14**: 1–27.

Singhi, S. and Liu, H. (2006). Feature subset selection bias for classification learning. In *Proceedings International Conference on Machine Learning*. New York: The ACM Press, pp. 849–56.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal American Society for Information Science and Technology*, **60**: 538–56.

Yang, Y. and Liu, X. (1999). A Re-Examination of Text Categorization Methods. In *Proceedings ACM-SIGIR*. New York: The ACM Press, pp. 42–9.

Yang, Y. and Pedersen, J. O. (1997). A Comparative Study of Feature Selection in Text Categorization. In *Proceedings International Conference on Machine Learning*. pp. 412–20.

Zhao, Y. and Zobel, J. (2007). Entropy-Based Authorship Search in Large Document Collection. In *Proceedings European Conference on Information Retrieval (ECIR)*. Berlin: Springer, pp. 381–92.

Appendix

Table A1 List of the functions used for feature selection

DIA(t_k, c_j)	$\text{Prob}[\mathrm{c_j} \mid \mathrm{t_k}]$
PMI(t_k, c_j)	$\log_2\left(\text{Prob}[t_k,c_j] \big/ \text{Prob}[t_k] \cdot \text{Prob}[c_j]\right) = \log_2\left(\text{Prob}[t_k \mid c_j]\right) - \log_2(\text{Prob}[t_k])$
OR(t_k, c_j)	$\text{Prob}[t_k \mid c_j] \cdot \left(1 - \text{Prob}[t_k \mid -c_j]\right) \big/ \left(1 - \text{Prob}[t_k \mid c_j]\right) \cdot \text{Prob}[t_k \mid -c_j]$
IG(t_k, c_j)	$\sum_{c \in \{c_j, -c_j\}} \sum_{t \in \{t_k, -t_k\}} \text{Prob}[t,c] \cdot \log_2\left[\text{Prob}[t,c] \big/ \text{Prob}[t] \cdot \text{Prob}[c]\right]$
GR(t_k, c_j)	$\text{Prob}[t,c] \cdot \log_2\left[\text{Prob}[t,c] / \text{Prob}[t] \cdot \text{Prob}[c]\right] + \text{Prob}[-t,c] \cdot \log_2\left[\text{Prob}[-t,c] / \text{Prob}[-t] \cdot \text{Prob}[c]\right]$
χ^2(t_k, c_j)	$\dfrac{n \cdot \left[\left(\text{Prob}[t_k,c_j] \cdot \text{Prob}[-t_k,-c_j]\right) - \left(\text{Prob}[t_k,-c_j] \cdot \text{Prob}[-t_k,c_j]\right)\right]^2}{\text{Prob}[t_k] \cdot \text{Prob}[-t_k] \cdot \text{Prob}[c_j] \cdot \text{Prob}[-c_j]}$
CC(t_k, c_j)	$\dfrac{\sqrt{n} \cdot \left[\left(\text{Prob}[t_k,c_j] \cdot \text{Prob}[-t_k,-c_j]\right) - \left(\text{Prob}[t_k,-c_j] \cdot \text{Prob}[-t_k,c_j]\right)\right]}{\sqrt{\text{Prob}[t_k] \cdot \text{Prob}[-t_k] \cdot \text{Prob}[c_j] \cdot \text{Prob}[-c_j]}}$
GSS(t_k, c_j)	$\left(\text{Prob}[t_k,c_j] \cdot \text{Prob}[-t_k,-c_j]\right) - \left(\text{Prob}[t_k,-c_j] \cdot \text{Prob}[-t_k,c_j]\right)$
BNS(t_k, c_j)	$\left\| F^{-1}(\text{Prob}[t_k \mid c_j]) - F^{-1}(\text{Prob}[t_k \mid -c_j]) \right\|$[a]

[a] F^{-1} represents the inverse of the normal cumulative distribution function.

Table A2 Estimation for the selection functions and possible values for a positive association or independence

Function	Estimation	Positive	Independence
$\mathrm{DIA}(t_k, c_j)$	$a \,/\, (a+b)$		
$\mathrm{PMI}(t_k, c_j)$	$\log_2[a \cdot n \,/\, (a+b) \cdot (a+c)]$	$\gg 0$	≈ 0
$\mathrm{OR}(t_k, c_j)$	$(a \cdot d) \,/\, (c \cdot b)$	$\gg 1$	≈ 1
$\mathrm{IG}(t_k, c_j)$	$a/n \cdot \log_2[a{\cdot}n \,/\, (a+b)(a+c)] +$ $b/n \cdot \log_2[b{\cdot}n \,/\, (a+b)(b+d)] +$ $c/n \cdot \log_2[c{\cdot}n \,/\, (a+c)(c+d)] +$ $d/n \cdot \log_2[d{\cdot}n \,/\, (b+d)(c+d)]$	$\gg 0$	≈ 0
$\mathrm{GR}(t_k, c_j)$	$a/n \cdot \log_2[a{\cdot}n \,/\, (a+b)(a+c)] +$ $c/n \cdot \log_2[c{\cdot}n \,/\, (a+c)(c+d)]$	$\gg 0$	≈ 0
$\chi^2(t_k, c_j)$	$n \cdot (a{\cdot}d - c{\cdot}b)^2 \,/$ $[(a+c){\cdot}(b+d){\cdot}(a+b){\cdot}(c+d)]$	$\gg 1$	≈ 0
$\mathrm{CC}(t_k, c_j)$	$\mathrm{sqrt}(n) \cdot (a{\cdot}d - c{\cdot}b) \,/$ $\mathrm{sqrt}[(a+c){\cdot}(b+d){\cdot}(a+b){\cdot}(c+d)]$	$\gg 0$	≈ 0
$\mathrm{GSS}(t_k, c_j)$	$[(a{\cdot}d) - (c{\cdot}d)] \,/n^2$	$\gg 0$	≈ 0
$\mathrm{BNS}(t_k, c_j)$	$\lvert\, \mathrm{F}^{-1}(a/(a+c)) - \mathrm{F}^{-1}(b/(b+d)) \,\rvert$ [a]	$\gg 0$	≈ 0

[a] F^{-1} represents the inverse of the normal cumulative distribution function.