Authorship Attribution Based on Specific Vocabulary

JACQUES SAVOY, University of Neuchatel

In this article we propose a technique for computing a standardized Z score capable of defining the specific vocabulary found in a text (or part thereof) compared to that of an entire corpus. Assuming that the term occurrence follows a binomial distribution, this method is then applied to weight terms (words and punctuation symbols in the current study), representing the lexical specificity of the underlying text. In a final stage, to define an author profile we suggest averaging these text representations and then applying them along with a distance measure to derive a simple and efficient authorship attribution scheme. To evaluate this algorithm and demonstrate its effectiveness, we develop two experiments, the first based on 5,408 newspaper articles (*Glasgow Herald*) written in English by 20 distinct authors and the second on 4,326 newspaper articles (*La Stampa*) written in Italian by 20 distinct authors. These experiments demonstrate that the suggested classification scheme tends to perform better than the Delta rule method based on the most frequent words, better than the chi-square distance based on word profiles and punctuation marks, better than the KLD scheme based on a predefined set of words, and better than the naïve Bayes approach.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text Analysis; H.3.1 [**Content Analysis and Indexing**]: Linguistic Processing; H.3.7 [**Digital Libraries**]:

General Terms: Performance, Experimentation

Additional Key Words and Phrases: Authorship attribution, text classifiaction, lexical statistics

ACM Reference Format:
Savoy, J. 2012. Authorship attribution based on specific vocabulary. ACM Trans. Inf. Syst. 30, 2, Article 12 (May 2012), 30 pages.
DOI = 10.1145/2180868.2180874 http://doi.acm.org/10.1145/2180868.2180874

1. INTRODUCTION

Given the extensive amount of textual information now freely available and recent progress made in Natural Language Processing (NLP) [Manning and Schütze 2000], a variety of text categorization tasks and successful solutions have been put forward [Sebastiani 2002; Weiss et al. 2010]. In this study, we consider authorship attribution [Craig and Kinney 2009; Juola 2006; Love 2002] whereby the author of a given text must be determined based on text samples written by known authors. More precisely, we focus on the closed-class attribution method in which the real author is one of several possible candidates. Other pertinent concerns related to this issue include the mining of demographic or psychological information on an author (profiling) [Argamon et al. 2009] or simply determining whether or not a given author did in fact write a given Internet message (chat, email, Wikipedia article) or document (verification) [Koppel et al. 2009].

Author's address: J. Savoy, Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2000 Neuchatel, Switzerland; email: Jacques.Savoy@unine.ch.

DOI 10.1145/2180868.2180874 http://doi.acm.org/10.1145/2180868.2180874

The initial requirement in these categorization problems is to represent the texts by means of numerical vectors comprising relevant features helpful in assigning them to various authors or categories. This process requires the ongoing extraction and selection of such features [Yang and Pedersen 1997], especially those more useful in identifying differences between the authors' writing styles (authorship attribution). More generally we need to seek out differences between topics or categories (e.g., politics, finance, macro-economics, sports) [Finn and Kushmerick 2005; Sebastiani 2002] or between genres (surveys, editorials, research papers, blog, homepages, etc.) [Argamon 2006; Stamatatos et al. 2001]. In a second stage we weight the selected features according to their discriminative power as well as their importance in the underlying textual representation. Finally, through applying classification rules or learning schemes [Bishop 2007; Hastie et al. 2009; Witten and Franck 2005], the system assigns the most appropriate author (or category) to a given input text (single-label categorization problem).

To achieve this objective we propose and evaluate a simple new method for selecting and weighting terms (e.g., *n*-gram of characters, word types, lemmas, *n*-gram of words, noun or verb phrases, Parts-Of-Speech (POS) sequences, etc.) and representing the documents and author styles involved. Our approach mainly relies on differences found between expected and observed term occurrence frequencies within two disjoint subsets. Based on a standardized Z score, we then define overused terms in one subset (defined as its specific vocabulary), terms common to both subsets (common vocabulary), and finally underused terms. In our opinion, a simple categorization rule capable of providing reasonable performance levels is preferable to a more complex approach (*lex parcimoniae* or Occam's razor principle [Bishop 2007]). Although it might not always be the best solution it would at least guarantee a practical system (a single-rule method was suggested and successfully applied in the data mining field [Holte 1993]). Moreover in our opinion, for a given corpus there is limited interest in obtaining better performance levels simply by adjusting various parameter settings without any solid theoretical foundation. Such a practice may lead to overfitting the model to the available data [Bishop 2007; Hastie et al. 2009] on the one hand, and on the other based on past experiments, the appropriate value for a new collection or context cannot usually be determined with the required precision. Finally, rather than relying on a black-box method we believe it is important that resulting decisions be clearly explained.

The rest of this article is divided as follows. Section 2 presents related works, while Section 3 depicts the main characteristics of the corpora used in our experiments. Section 4 briefly describes three classical author attribution approaches: the Delta method [Burrows 2002], the χ^2 statistic [Grieve 2007], and KLD [Zhao and Zobel 2007a, 2007b] to which our suggested scheme will be compared. This section exposes the naïve Bayes method [Mitchell 1997], a well-known approach in the machine learning domain. Finally, this section also describes and evaluates our proposed authorship attribution approach based on the Z score method. Section 5 summarizes our main findings and identifies future perspectives.

2. RELATED WORK

Authorship attribution has a long-standing history and recently various noteworthy literature surveys have been published [Juola 2006; Koppel et al. 2009; Love 2002; Stamatatos 2009; Zheng et al. 2006]. As a first paradigm to solve the authorship attribution problem, various approaches based on unitary invariant values have been proposed [Holmes 1998]. These invariant measures must reflect the particular style of a given author, but they should vary from one to another. Previous studies involving this strategy suggested the use of lexical richness or word distribution factors, including average word length and mean sentence length, as well as Yule's

K measure [Miranda-Garcia and Calle-Martin 2005] and statistics on type-token ratios (e.g., Herdan's C, Guiraud's R or Honoré's H), as well as the proportion of word types occurring once or twice (e.g., [Sichel 1975]), or even the slope of Zipf's empirical distribution [Baayen 2001, 2008, Section 6.5; Tuldava 2004]. To these we could also add a few simple statistics such as letter occurrence frequencies [Ledger and Merriam 1994], mean number of syllables per word, number of *hapax legomena* (words occurring once) and their relative positions in a sentence [Morton 1986], etc. As other possible sources of evidence, we might consider the vocabulary size attributed to a given author [Efron and Thisted 1976; Thisted and Efron 1987]. None of these measures has proved very satisfactory, however [Hoover 2003], due in part to word distributions (including word bigrams or trigrams) ruled by a large number of very low probability elements (Large Number of Rare Events or LNRE) [Baayen 2001].

In a second stage, instead of limiting ourselves to a single value we could apply a multivariate analysis to capture each author's discriminative stylistic features [Holmes 1992; Holmes and Crofts 2010; Holmes and Forsyth 1995]. Some of the main approaches applicable here are Principal Component Analysis (PCA) [Binonga and Smith 1999; Burrows 1992; Craig and Kinney 2009], cluster analysis [Labbé 2007], and discriminant analysis [Jockers and Witten 2010; Ledger and Merriam 1994]. In this case we represent documents (with known authors) as points within a given space, and to determine who might be the author of a new text excerpt we simply search the closest document [Hoover 2006], where the author of this nearest document would probably be the author of the disputed text. For these cluster-based approaches to be effective, however, the distance measure definition is of prime importance, and with this in mind various metrics are suggested. We might, for example, mention standardized word-count frequency values [Binonga and Smith 1999] as well as the more sophisticated intertextual distance [Labbé 2007], where the distance between two documents depends on both their shared vocabulary and occurrence frequencies. Yang et al. [2003] proposed a similar approach where the distance between two texts is based on the weighted sum of the rank order-frequency differences of word types occurring in both texts. This distance measure tends to group the documents into several classes, with each reflecting a distinct style or author.

Other recent studies pay more attention to various categories of topic-independent features that may more closely reflect an author's style, and in this perspective we can identify three main sources. First, at the lexical level are word occurrence frequency (or character n-grams), *hapax legomena*, average word length, letter occurrence frequency [Merriam 1998], and punctuation frequency, along with several other representational marks. Special attention has also been given to function words (e.g., determiners (e.g., *the*, *an*), prepositions (*in*, *of*), conjunctions (*and*), pronouns (I, *he*), and certain auxiliary verbal forms (*is*, *was*, *should*)), features which appear in numerous authorship attribution studies [Burrows 2002]. Certain authors have suggested a wide variety of lists, although the precise definition of these function word lists is questionable. Burrows [2002], for example, lists the top n most frequent word types (with $n = 40$ to 150), Holmes and Forsyth [1995] 49 high-frequency words, Baayen and Halteren [2002] a list of 50 words, while Jockers et al. [2008, p. 491] suggest 110 entries, while the list compiled by Zhao and Zobel [2005] contains 363 words. Finally, Hoover [2006] put forward a list of more than 1000 frequently occurring words, including both function words (determiners, prepositions, conjunctions, pronouns, auxiliary verbs) and lexical words (nouns, adjectives, verbs, adverbs). The interjection category (e.g., *oh*, *ah*) as well as other-than-manner adverbs might also be added to the functional word class [Miranda Garcia and Calle Martin 2007].

Not all studies, however, suggest limiting the possible stylistic features to a reduced set of functional words or very frequent word types. In their study of the 85 *Federalist*

Papers for example, Jockers and Witten [2010] derive 2907 words appearing at least once in texts written by all three possible authors. From this word list, the researchers could extract a reduced set composed of 298 words, after imposing the condition that for each item the relative frequency must be greater than 0.05%.

Secondly, at the syntactic level we could account for Part-Of-Speech (POS) information through measuring their distribution, frequency, patterns, or various combinations. Thirdly, some studies suggest considering structural and layout features including the total number of lines, 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. Additional features considered could be particular orthographic conventions (e.g., British versus U.S. spelling) or the occurrence of certain specific spelling errors, and the resulting number of potential features considered could thus be rather large. Zheng et al. [2006], for example, compiled a list of 270 possible features.

After selecting the most appropriate characteristics for a given document, we then need a classification scheme capable of distinguishing between its various possible authors. Related to this is the problem involving identifying the authors of short online messages for which Zheng et al. [2006] suggest employing decision trees, backpropagation neural networks, and Support Vector Machines (SVM). Based on corpora written in English or Chinese, these experiments analyze various lexical, syntactic, structural, as well as other content-specific features. For English descriptions that are only based on lexical features result in performance levels similar to POS and lexical feature combinations. This finding is confirmed by another recent study [Zhao and Zobel 2007b]. Zheng et al. [2006] also find that SVM and neural networks tend to performance levels significantly better than those achieved by decision trees. Zhao and Zobel [2005], on the other hand, find that when defining the authorship of newspapers articles the Nearest-Neighbor (NN or *k*-NN) approach tends to produce better effectiveness than both the naïve Bayes or decision-tree approaches (five possible authors, 300 training documents per author).

Instead of applying a general-purpose classification method, Burrows [2002] designs a more specific Delta classifier based on the "mean of the absolute difference between the z-scores for a set of word-variables in a given text-group, as well as the z-scores for the same set of word-variables in a target-group". This method was originally based on the 150 most frequently occurring word tokens while Hoover [2004b] suggested this scheme could be improved by considering the top 800 most frequent words. A few Delta method variants have also been put forward [Hoover 2004a, 2007], as well as various other interpretations of this same scheme [Argamon 2008; Stein and Argamon 2006]. In all cases the underlying assumption is that a given author's style is best reflected by identifying the use of function words (or by very frequent words) together with their occurrence frequencies, rather than relying on a single vocabulary measure or more topic-oriented terms. Recently, Jockers and Witten [2010] showed that the Delta method could surpass performance levels achieved by the SVM method. In a related study Kešelj et al. [2003] propose summing the normalized differences of occurrence frequencies, which based on their results and performance levels, proved to be fairly effective methods. To capture the individual style nuances of each author under consideration, these same researchers also suggest applying *n*-gram characters instead of words.

In summary, it seems reasonable to suggest that we make use of vocabulary features, thus allowing us to conclude not only the presence or absence of words but also their occurrence frequencies, allowing us to reveal the underlying and unknown "fingerprint" of a given author during a specified period and relative to a particular genre and topic. It is known, however, that word frequencies tend to change over time and use [Hoover 2006], as do genres or forms (e.g., poetry or romance, drama or comedy, prose or verse) [Burrows 2002; Hoover 2004b; Labbé 2007].

3. EVALUATION

Unlike the information retrieval domain [Manning et al. 2008], the authorship attribution domain does not benefit from a relatively large number of publicly available corpora. As such, making sufficiently precise comparisons between reported performances and general trends regarding the relative merits of various feature selections, weighting schemes, and classification approaches is problematic. Moreover, so that verification and comparison can be done by others, the test collections used to evaluate a proposed scheme must be stable and publicly available. Finally, we are convinced that absolute performance levels cannot be directly compared across the various evaluation studies. As such, only relative rankings between different tested schemes could be reliably utilized, rather than direct comparisons between absolute performance levels obtained from distinct corpora. When employing the same corpus, however, it is not always fully clear how the various processing methods should be implemented (e.g., tokenization, normalization of uppercase letters, etc.).

Another main concern is the size of the available test collection. In various previous studies, the number of disputed texts and the number of possible authors are rather limited. With the well-known *Federalist Papers*, for example, we tackled 85 texts from which 12 are disputed articles written mainly by two possible authors (binary or two-case classification) [Holmes and Forsyth 1995; Jockers and Witten 2010; Mosteller and Wallace 1964]. Various binary classification problems related to Shakespeare's works are discussed in Craig and Kinney [2009], while in Burrows [1992] various experiments are performed on six texts having two possible authors. Moreover, various in-deep studies focus on a single text (book, play, diary) by two or three authors [Holmes and Crofts 2010; Hoover and Hess 2009; Jockers et al. 2008; Ledger and Merriam 1994]. Other studies are, however, based on literary texts where the number of possible authors is greater than three, such as experiments described in Labbé [2007], which focus on 52 text excerpts written by possibly nine distinct authors.

3.1. Corpus Evaluation

To handle these problems and in the interest of promoting test beds comprising more authors and documents, we may consider using literary works available through dedicated Web sites such as the Gutenberg project (see www.gutenberg.org). The number of possible documents is, however, limited, due to the fact that not all works are available and certain recent works are still under copyright. Along this same vein, Zhao and Zobel [2007a] were able to download 634 books written by 55 authors mainly covering English literature. To include comparable styles from different authors, we must, however, consider texts of the same or similar genres, written during the same period. Mixing Twain's works with Shakespeare's plays or even translations from Schiller's works, for example, does not produce a very useful corpus.

An alternative might be downloading Wikipedia articles, although such a corpus would not be stable. At any time and without warning, a given text could be more or less heavily edited, and even worse fully disappear, replaced by another, or written by another person. Moreover, in working with such freely available material, we would have to contend with greater variability in writing quality, expressions, and language registers employed. More variability should also be expected with respect to authors and their own backgrounds, given they could originate from very different cultures, a phenomenon that renders the resulting test collection less challenging and less pertinent.

To build a large and useful test collection, we could employ a corpus of newspaper articles. In this vein, Grieve [2007] downloaded articles from the London *Telegraph* Web site (published from April 2001 to January 2005). The resulting corpus contained works by 40 authors, each having 40 columns (1600 documents in total). In this case, the precise selection of each document is not specified and free access to this corpus is not guaranteed. Zhao and Zobel [2007b] used a similar strategy by considering articles made available by newswire services (Associated Press), comprising about 200,000 articles written by around 2380 authors. These newswire articles usually contain very short documents in which the authors may simply describe an event (or simply translate it) without adding any personal comments reflecting their own style. In the end, having a large number of authors is not always the most pertinent approach. It is known, for example, that in the event of disputed texts, the number of possible authors is usually limited, with only 10 to 20 possible writers covering a large majority of problematic cases, at least in terms of literary analysis. Moreover, in the analysis of political speeches when searching for the name of the actual speechwriter behind each discourse (such as T. Sorensen writing for President Kennedy [Carpenter and Seltzer 1970]), the number of possible authors is also limited, and certainly under the limit of 20. Even when the number of possible writers is limited, the fact that they share a common culture and education could render the task more difficult, as, for example, in the case of Goldsmith, Kelly, and Murphy and their common Anglo-Irish roots [Dixon and Mannion 1993].

In order to obtain a replicable test collection containing 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[1] [Peters et al. 2004]. More precisely, we extract articles published in the *Glasgow Herald* (GH) during 1995, a subset comprising 56,472 documents, of which 28,687 included the name of the author(s). Knowing that an article could be written by two or more authors, or that an author could contribute to only a few texts, we could not simply decide to use all these articles. In order to form a suitable test collection, we thus chose 20 authors (see Table I), either as well-known columnists (names in italics) or having published numerous papers in 1995. This selection process yields a set of 5,408 articles.

As shown in Table I, the *Glasgow Herald* (GH) corpus covers different subjects and a clear overlap among authors evidently exists. Five authors are listed under the main descriptor *Business* and also under *Sports*, while only four are listed under *Social*, and three under both the *Politics* and *Arts and Film* headings. The advantage of this corpus is that it contains articles written in a similar register, targeting the same audience, during the same short period of time (1995), and by authors sharing a common background and culture. Moreover, throughout all articles copy-editors and proofreaders impose respect for the in-house newspaper style, correct orthography (spelling, punctuation, and capitalization) while also reinforcing the use of the same vocabulary and naming conventions (e.g., Beijing or Peking).

The "Number" column in Table I lists the number of articles written by each author, showing a minimum of 30 (Fowler John), and a maximum of 433 (Wilson Andrew). This distribution is rather skewed, with a group of eight authors having published more than 350 articles, and another group of four journalists in this corpus writing less than 100 articles (mean: 270, median: 332, standard deviation: 139). Moreover, an analysis of article length shows that the mean number of word tokens is 725 (minimum: 44, maximum: 4,414, median: 668, standard deviation: 393), an overall value closely reflecting only one of the chosen authors (Gallacher Ken), in terms of the mean tokens length of 727 per article, as reported under the column "Mean Length". This mean

[1] This corpus is available thought the ELRA Web site (www.elra.info).

Table I. Distribution of *Glasgow Herald* Articles by Author, Subject, Number of Articles per Author, and Their Mean Length (in number of word tokens)

	Name	Subjects	Number	Mean Length
1	*Young Alf*	Business, Economics	208	1,013
2	*Davidson Julie*	Arts & Film	57	1,310
3	Douglas Derek	Sports	410	808
4	*Fowler John*	Arts & Film	30	890
5	Gallacher Ken	Sports	408	727
6	Gillon Doug	Sports	368	713
7	*Johnstone Anne*	Social, Politics	72	1,258
8	McConnell Ian	Business	374	455
9	*McLean Jack*	Social, Sports	118	1,008
10	Paul Ian	Sports	418	842
11	Reeves Nicola	Business, Social	370	531
12	Russell William	Arts & Film	291	1,019
13	*Shields Tom*	Politics	173	1,001
14	Sims Christopher	Business	390	471
15	*Smith Ken*	Social, Culture	212	616
16	Smith Graeme	Social, Politics	329	520
17	Traynor James	Sports	339	983
18	Trotter Stuart	Politics	336	666
19	Wilson Andrew	Business	433	452
20	*Wishart Ruth*	Politics	72	1,137

value varies widely across journalists indicating that Davidson writes longer articles, on average, (mean: 1,310) while Wilson has the shortest mean (452).

As a second evaluation corpus, we selected newspaper articles published in *La Stampa* during the year 1994, a subset comprising 58,051 documents, of which 37,682 included the name of the author(s). This corpus is part of the CLEF 2003 test collection [Peters et al. 2004], which is available publicly through the ELRA Web site. In selecting this corpus, our intention was to verify the quality of the different authorship attribution methods using a language other than English.

From the set of all possible articles, we must ignore articles written by more than one author, as well as authors contributing to only a few texts. In order to form a suitable test collection, we thus chose 20 authors (see Table II), either as well-known columnists (names in italics) or as authors having published numerous papers in 1994. This selection process resulted in a set of 4,326 articles.

The "Number" column in Table II lists the number of articles written by each author, showing a minimum of 52 (Nirenstein Fiama), and a maximum of 434 (Del Buono Oreste). An analysis of article length shows that the mean number of word tokens is 777 (minimum: 60; maximum: 2,935; median: 721; standard deviation: 333). As for the *Glasgow Herald* corpus, this mean value varies widely across journalists indicating that Spinelli writes longer articles, on average, (mean: 1,478) while Conti has the shortest mean (612). In the selected newspaper articles, we automatically remove the author name (full name or first name) as well as some recurrent phrases (e.g., *Dal nostro* (or *della nostra*) *corrispondente*, *nostro servizio*, etc.).

3.2. Evaluation Measures

We use the accuracy rate as evaluation measures, meaning the percentage of correct answers that can be computed according to two distinct schemes. As a first method, the micro-averaging principle assumes that one decision corresponds to one vote. When the system is able to correctly identify, for example, the right author for 3,166 articles out of a grand total of 5,408 articles, the resulting accuracy rate (micro-average) is

Table II. Distribution of *La Stampa* Articles by Author, Subject, Number of Articles per Author, and Their Mean Length (in number of word tokens)

	Name	Subjects	Number	Mean Length
1	Ansaldo Marco	Sports	287	812
2	Battista Pierluigi	Politics	231	840
3	*Beccantini Roberto*	Sports	364	831
4	*Beccaria Gabriele*	Social	71	686
5	Benedetto Enrico	Politics	252	732
6	Del Buono Oreste	Sports	434	799
7	Comazzi Alessandra	Social	223	616
8	Conti Angelo	Social	198	612
9	Galvano Fabio	Politics	347	738
10	*Gramellini Massimo*	Politics	118	955
11	Meli Maria Teresa	Politics	215	857
12	*Miretti Stefania*	Social	63	793
13	*Nirenstein Fiama*	Politics	52	1,090
14	Novazio Emanuele	Politics	249	750
15	Ormezzano Gian Paolo	Sports	232	738
16	Pantarelli Franco	Politics	202	692
17	Passarini Paolo	Politics	303	720
18	Sacchi Valeria	Business	203	776
19	*Spinelli Barbara*	Politics	57	1,478
20	Torabuoni Lietta	Social	225	784

3166/5408 = 0.5854 or 58.54%. In authorship attribution this is the method most frequently used to compute mean performance.

As a second method we first compute the accuracy rate obtained for each of the 20 authors (or categories), under the assumption that we attach the same importance to each author (or category). In this case, one author corresponds to one vote (macro-average), and thus the overall accuracy rate is the mean of all categories. For example, if we obtain an accuracy rate of 0.7 for the first author, 0.4 for the second, and 0.8 for the third, then the macro-averaging accuracy rate is (0.7 + 0.4 + 0.8) / 3 = 0.633, or 63.3%. When we have the same number of texts for each author, both measures return the same value but, as depicted in Tables I and II, this is not the case in our evaluation corpora.

Both the micro- or macro-average measures are presented in this study and either can be used. In the machine learning domain, the first one usually tends to produce better results because frequent categories are assigned more importance, and are usually easier to predict. With more data, a frequent category (or author) might be more precisely defined or the underlying classifier would have more training data to distinguish between this particular category and the others.

To determine statistically whether or not a given attribution method would be better than another scheme, we apply the sign test (or s-test) [Conover 1980] in which the null hypothesis H_0 states that both attribute models result in similar performance levels [Yang and Liu 1999]. When applying a two-sided test, n' denotes the number of times that the assignment resulting from each of the two models is different. Moreover, t_+ represents the number of times that the first system proposes a correct assignment while the second system indicates an incorrect decision. Under the H_0 assumption stating that both schemes produce similar performance, t_+ follows a binomial distribution with parameter $p = 0.5$ and n'. Thus at a given significance level α, the expected limit for the t_+ value is

$$t = 0.5 \cdot \left(n' - z_{a/2} \cdot \sqrt{n'}\right).$$

When fixing the significance level α = 5%, the $z_{\alpha/2}$ value is 1.96 (or 2.57 for a significance level at α = 1%). The null hypothesis is rejected if the observed value t_+ is smaller than t or greater than $n' - t$.

When applying the sign test to the macro-averaging method, we compare the two attribution schemes using the 20 means (one per category or author). In the current evaluation, we consider them as equal if the absolute value of the accuracy difference between two authors is smaller than 0.001. Of course, due to the fact that the value of n' (the number of times that the accuracy per author between the two models differs) is much smaller than that of the micro-averaging method, the sign test does not detect many significant differences.

4. TEXT CLASSIFICATION MODELS

To design and implement an automatic authorship attribution system we need to choose a text representation mechanism that is beneficial when classifying the texts, and also a classifier model. Section 4.1 describes the common form of representation used in our experiments. To provide a comparative view of the relative merits of the three attribution models, in Section 4.2 we choose the Delta rule, in Section 4.3 the χ^2 statistic, and in Section 4.4 the KLD approach. Furthermore, the definition of term specificity based on the Z score is described in Section 4.5, while in Section 4.6 we define a distance between text pairs and then evaluate the suggested authorship attribution method and compare it with the best performance levels achieved when applying the three other schemes. In Section 4.7, we present a set of additional experiments using the same set of terms to evaluate the four author attribution schemes while Section 4.8 compares the effectiveness of the Z score method with the naïve Bayes, a well-known approach used in machine learning. Finally Section 4.9 estimates the reliability of the suggested Z score distance.

4.1. Preprocessing and Text Representation

Even though Kešelj et al. [2003] found that character n-gram representation could be effective in authorship attribution as well as in the information retrieval domain [McNamee and Mayfield 2004], we prefer a method capable of clearly verifying text representation generated, and thus our text representations are based on words.

Before trying to classify the newspaper articles, we first need to preprocess them. We begin by replacing certain system punctuation marks (in UTF-8 coding) with their corresponding ASCII symbols, and replacing single (’) or double quotation marks (“”) with the (') or (") symbols. For the English language only, we remove a few diacritics found in certain words (e.g., *naïve*). To standardize spelling forms we also expand contracted forms or expressions (e.g., *don’t* into *do not*) and replace uppercase letters with their corresponding lowercase equivalents, except for certain words written only with capital letters (e.g., *US*).

To break the stream of text into tokens, we apply the tokenization algorithm developed by Grefensette and Tapanainen [1994], and thus consider words such as *soldiers* and *soldier* to be distinct forms, as we do for each of the conjugated verb forms (e.g., *writes*, *wrote*, or *written*). Moreover, we do not distinguish between possible homographs (e.g., the verb *to desert*, and the noun *desert*) by considering their Part-Of-Speech (POS) categories. In the case of high-frequency words, for example, this distinction provides an entry for *to* as the infinitive or another for *to* as preposition.

After this step, the resulting English vocabulary contains 56,447 distinct word types, with 19,221 *hapax legomenon* (words occurring once), and 7,530 *dis legomenon* (words occurring exactly twice). When considering only those types having an occurrence frequency of 10 or more, we count 14,890 types, or 9,628 types having frequencies equal to or greater than 20. The most frequent token is *the* (219,632 occurrences), followed by the comma (183,338 occurrences), the period (146,590), and ranking fourth is the token *to* (95,350), followed by *of* (92,755), and *a* (78,867).

From the newspaper *La Stampa*, we find 102,887 distinct word types, with 41,965 *hapax legomenon*, and 14,944 *dis legomenon*. In this corpus, we can count 19,580 word types having an occurrence frequency of 10 or more, and 11,410 types having frequencies equal to or greater than 20. The most frequent token is the comma (212,736 occurrences), followed by the period (126,891), and the word type *di* (of) (100,433), and ranking fourth is the token *e*(and) (73,818), followed by *il* (the) (63,931), and *che* (that) (59,600).

In order to define the underlying characteristics of each author, we form an author profile by concatenating all texts written by the same person. From this subset, we then apply the feature selection procedure, and represent each author profile or disputed text by a set of weighted features.

In all experiments, the query text is never included in the corresponding author profile. Moreover, not using this test data during the learning stage or when building the author profile is considered as a fair evaluation principle. In our experiments, the preprocessing of the texts was done using Perl [Bilisoly 2008; Nugues 2006] while the classification and the evaluation were performed using the R system [Crawley 2007].

4.2. Delta Rule

To determine the probable author of a given text, Burrows [2002] suggests accounting for the most frequent word types (and particularly function words) without taking punctuation marks or numbers into account. In an original proposition, Burrows suggests considering from 40 to 150 most frequently occurring word types, with 150 words obtaining the best results. Unlike in Burrows' study, we did not distinguish between homographs, as, for example, between *that* as a conjunction or as a relative pronoun. We must admit that this selection criterion is rather simple to apply, and that computational costs are relatively low, particularly when ignoring the ambiguity of the homographs. On the other hand, taking account of these differences would increase underlying manual or computational costs, rendering this authorship attribution method less appealing.

When comparing two texts, Burrows [2002] suggests that the second important aspect is not the use of absolute frequencies, but rather their standardized scores. These values are obtained by subtracting the mean and then dividing by the standard deviation (Z score) [Hoover 2004a], and once these dimensionless quantities are obtained for each selected word, they can be compared to those obtained from other texts or author profiles. We compute the Z score for each term t_i (word type) in a text sample (corpus) by calculating its term relative frequency tfr_{ij} in a particular document D_j, as well as the mean ($mean_i$), and standard deviation (sd_i) of term t_i according to the underlying corpus (see Eq. (1)) [Hoover 2004a].

$$Z\ \ score(t_{ij}) = \frac{tfr_{ij} - mean_i}{sd_i} \tag{1}$$

From the Z score value attached to each term, we can compute a distance between each pair of texts. Then, given the query text Q, and the author profile A_j, and a set of terms t_i, for $i = 1, 2, \ldots, m$, we compute the Delta value (or the distance) by applying Eq. (2). In this formulation we attach the same importance to each term t_i, independently of their absolute occurrence frequencies. Large differences may occur when, for a given term, both Z scores are large and have opposite signs, and in these

Table III. Evaluation of Delta Method (GH corpus, 5,408 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
Delta	40 words	43.53% ‡	45.97% ‡
Delta	150 words	58.54% ‡	60.80%
Delta	200 words	59.91% ‡	62.75%
Delta	400 words	**63.70%**	**66.14%**
Delta	600 words	61.35% ‡	63.52%
Delta	800 words	54.81% ‡	58.00%
Delta	400 words – PP	60.63% ‡	63.43%
Delta	600 words – PP	61.32%	64.15%
Delta	800 words – PP	53.92% ‡	57.30%

cases one author tends to use the underlying term more frequently than the mean while the other employs it very infrequently. On the other hand, when for all terms the Z scores are very similar, the distances between the two texts would be small, indicating the same author had probably written both of them.

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

The Delta method was originally applied in the Restoration poetry corpus [Burrows 2002], and Hoover [2004b] demonstrated that this method could be effective in a prose corpus containing either dialog or more narrative content (American English texts from the end of the 19th century to the beginning of the 20th century). In this case the text excerpts contained 10,000 to 39,000 word tokens, with a mean length value of 27,000.

In a related study, Hoover [2004a] suggests ignoring personal pronouns in the list of high-frequency words (it is not clear whether this suggestion was made in relation to the underlying corpora or should be applied in all cases). The resulting effect might be small, however, given the rather small number of personal pronouns in a list of 600 to 800 entries.

Table III shows the evaluation obtained with the Delta method using the GH corpus while Table IV reports the same information for *La Stampa*. Under the "Parameter" heading we list the number of high-frequency words taken into account, and when personal pronouns are ignored (- PP). In the last two columns, we report the accuracy rate computed with the micro-average rate (one vote per text) and macro-average rate (one vote per author). Even though micro-averages result usually in lower performance levels, the same conclusions could be drawn from both measures and both corpora. The best performance is obtained using 400 words, and accounting for more words tends to diminish the classifier's quality. Removing the personal pronouns ("- PP") tends to reduce performance levels when considering 400 words, but has no real impact when using 600 or 800 word types.

Using the sign test and the best performance (400 words) as baseline, we add a cross (†) to indicate significance performance differences (significance level α = 5%, two-sided) or a double cross (‡) for significance differences having a significance level α = 1% (two-sided). As shown in Tables III and IV, the performance differences with the best parameter settings tend to be statistically significant when considering the micro-average measure. When using the macro-average indicator however performance differences tend to be nonsignificant, mainly due to the fact the sample size is reduced to 20 (authors).

The statistical tests listed in the bottom part of Tables III and IV, compare the performance differences with and without personal pronouns (- PP). In this case, ignoring

Table IV. Evaluation of Delta Method (*La Stampa* corpus, 4,326 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
Delta	40 words	43.44% ‡	43.36% ‡
Delta	150 words	63.62% ‡	63.21% ‡
Delta	200 words	68.70% ‡	68.75% ‡
Delta	400 words	**76.07%**	**75.08%**
Delta	600 words	73.49% ‡	73.61%
Delta	800 words	66.30% ‡	67.20% ‡
Delta	400 words – PP	74.90% ‡	74.43%
Delta	600 words – PP	74.78% ‡	75.10%
Delta	800 words – PP	67.73% ‡	68.84% †

personal pronouns tends to significantly decrease the micro-average performance levels, while considering the macro-average measure removing them tends to have no precise and real effect.

In the Delta method the feature selection criterion is rather simple, given that it is based only on occurrence frequencies, and word distributions across texts or authors are ignored. This strategy favors words with high occurrence frequencies, even when the underlying occurrences appear only in a few but long documents instead of considering words occurring in a large number of texts or author profiles. Moreover, the feature's capacity to discriminate between different authors is not taken into account.

Hoover [2004a] suggests considering occurrence distributions across the different texts by ignoring those word types for which a single text supplies more than 70% of their occurrences (culling process). In the GH corpus, for example, we count 193 word types having occurrence frequencies greater or equal to 10, and for which a single text contains more than 70% of all occurrences. Here the term *Nuremberg* is found to be the most extreme case, having the highest occurrence frequency (47) and with only a single document containing 37 occurrences (or 82%), and thus in this context the culling process has no real effect. From a set of 14,890 words occurring 10 times or more, removing 193 (or 1.3%) of the entries might have no visible impact. Moreover, in our example, a word type having an occurrence frequency of 47 is not ranked among the top 800 most frequently occurring word types (ranking 800 is the term *media* with a frequency of 476, and at 1000 is *conservative,* with a frequency of 381).

4.3. Chi-Square Distance

As a second baseline, we select one of most effective text representations found in an empirical study [Grieve 2007]. This effective text representation is based on the relative frequency of word tokens together with punctuation marks, comprising the eight symbols (. , : ; - ? (’). For feature selection, instead of accounting for all word types, Grieve [2007] considers words in a k-limit profile, where k indicates that each word type must occur, at least, in k articles written by a given author and for every possible author (e.g., a value $k = 5$ imposes the presence of the corresponding term in at least five articles written by every possible author). This selection criterion can also be analyzed as a minimum document frequency value on a per-author basis. As effective values for the parameter k, Grieve [2007] observed that the best performance results were achieved when $k = 2$, $k = 5$, or $k = 10$ (knowing that each author had written exactly 40 texts in a corpus of 1,600 newspapers articles). Although increasing the value of k reduces the number of word types taken into account, a small value for k implies that we consider more words, and particularly more content words.

To compare the representation of a given text Q with an author profile A_j, Grieve [2007] uses the χ^2 statistic defined by Eq. (3) in which $q(t_i)$ represents the ith feature in the query text, and $a_j(t_i)$ the corresponding ith feature in the jth author profile, for the set of terms t_i, for $i = 1, 2, \ldots, m,$. In the current case, the values of $q(t_i)$ and $a_j(t_i)$ become the relative frequencies of a given word or punctuation symbol.

$$\chi(Q, A_j) = \sum_{i=1}^{m} \left(q(t_i) - a_j(t_i)\right)^2 \Big/ a_j(t_i) \tag{3}$$

When comparing a text with different author profiles, we simply select the lowest χ^2 value to determine the most probable author. Admittedly, when computing this metric, many small values for either $q(t_i)$ or $a_j(t_i)$ could be problematic [Knuth 1981]. Grieve [2007] did not, however, specify any special treatment, and thus we strictly followed the described procedure. When applying the 2-limit of course, all $a_j(t_i)$ values would be greater than zero, and thus the divisor shown in Eq. (3) would never be zero. The 2-limit does in fact impose that each word or punctuation mark must appear in at least two documents. At the limit, the author profile minus the query text would contain one occurrence of the given term, and the corresponding $a_j(t_i)$ would therefore always be greater than zero.

In this scheme, feature selection is based on the document frequency (*df*), considered in information retrieval to be a useful relevance indicator [Manning et al. 2008]. The *df* value is nonetheless not computed for the entire corpus, but rather on a per-author basis. Using document frequency as selection feature has also been found effective in other text categorization problems, as mentioned by Yang and Pedersen [1997].

> This suggests that DF *(document frequency)* thresholding, the simplest method with the lowest cost in computation, can be reliably used instead of IG *(information gain)* or CHI *(χ^2-test)*.

With the GH corpus, the 30-limit is chosen as the maximum because we only have 30 articles written by Fowler John. In this case, the system can select 15 terms, being {*a and as but from in is it of that the to with , .* }. When using the corpus *La Stampa*, the system may select up to the limit of 52 (corresponding to the maximum number of articles written by one author, F. Nirenstein in this case). Appearing in all texts, we find the following 20 word types and punctuation marks {*a al che da del della di e è i il in l la non per un . , '* }.

The accuracy rates analysis reported in Table V (GH corpus) or Table VI (*La Stampa*) indicates that the best performance under micro-average measure is achieved when considering the 2-limit constraint, involving more words and punctuation symbols than with the other solutions. For the GH corpus, the 5-limit produces the best accuracy rate when considering the macro-average metric. For both corpora, however, performance differences between the 2-limit or 5-limit schemes are rather small, but when compared to other parameter settings, the performance differences are relatively important. Using the best performance as baseline and applying a two-sided sign test, a double cross (‡) indicates a significant performance (significance level α = 1%) while a single cross (†) is associated with a significance level of 5%. As shown in Tables V and VI, the performance differences with the best parameter setting are always statistically significant when analyzing micro-average measure. Using the macro-average measure, the performance differences with the best parameter setting tend not to be significant, except with the GH corpus where the sign test detects significant differences with the 20- and 30-limits.

4.4. Kullback-Leibler Divergence

Zhao and Zobel [2007a, 2007b] suggest considering a limited number of predefined word types to discriminate between different author profiles. Their proposed English

Table V. Evaluation of χ^2 Statistic on Words and Punctuation Marks (GH corpus, 5,408 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
χ^2 measure	2-limit (653 terms)	**65.26%**	63.57%
χ^2 measure	5-limit (289 terms)	62.39% ‡	**65.26%**
χ^2 measure	10-limit (149 terms)	59.39% ‡	62.84%
χ^2 measure	20-limit (52 terms)	52.27% ‡	52.48% †
χ^2 measure	30-limit (15 terms)	40.03% ‡	40.36% ‡

Table VI. Evaluation of χ^2 Statistic on Words and Punctuation Marks (*La Stampa* corpus, 4,326 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
χ^2 measure	2-limit (720 terms)	**68.28%**	**65.78%**
χ^2 measure	5-limit (333 terms)	65.49% ‡	65.40%
χ^2 measure	10-limit (203 terms)	66.07% ‡	66.99%
χ^2 measure	20-limit (106 terms)	62.83% ‡	62.97%
χ^2 measure	30-limit (71 terms)	62.51% ‡	61.58%
χ^2 measure	40-limit (42 terms)	59.78% ‡	59.10%
χ^2 measure	50-limit (30 terms)	56.26% ‡	56.01%
χ^2 measure	52-limit (20 terms)	49.24% ‡	48.74%

list contains 363 terms, mainly function words (e.g., *the*, *in*, *but*, *not*, *am*, *of*, *can*) and also certain frequently occurring forms (e.g., *became*, *nothing*). Other entries are not very frequent (e.g., *howbeit*, *whereafter, whereupon*), while 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*). Zhao and Zobel's [2007a, 2007b] study is limited to the English language, and thus for the Italian language we select an Italian stopword list provided by a search system achieving high retrieval performance in CLEF evaluation campaigns for that language [Savoy 2001]. After defining the feature set, the probability of occurrence of each item associated with a given author or a disputed text then has to be estimated.

Based on these estimations, we can measure the degree of disagreement between two probabilistic distributions. To do so Zhao and Zobel [2007a, 2007b] suggest using the Kullback-Leibler Divergence (KLD) formula, also called *relative entropy* [Manning and Schütze 2000], a choice that has proven effective in the information retrieval domain [Zhai and Lafferty 2004]. The KLD value expressed in Eq. (4) indicates how far the feature distribution derived from the query text Q diverges from the jth author profile distribution A_j. We have

$$KLD(Q||A_j) = \sum_{i=1}^{m} p_q(t_i) \cdot \log_2 \left[\frac{p_q(t_i)}{p_j(t_i)} \right], \tag{4}$$

where $p_q(t_i)$ and $p_j(t_i)$ indicate the occurrence probability of the term t_i in the query text or in the jth author profile, respectively. In the underlying computation, we state that $0 \cdot \log_2[0/p] = 0$, and $p \cdot \log_2[p/0] = \infty$.

With this definition and when the two distributions are identical, the resulting value is zero, while in all other cases the returned value is greater than zero. With this approach the main concern is accurately estimating the different probabilities. As a first estimate for the occurrence probability of term t_i (namely $p_q(t_i)$ or $p_j(t_i)$), we apply the maximum likelihood principle and estimate it as

$$p(t_i) = {}^{tf_i}/_{n}, \tag{5}$$

where tf_i indicates the term frequency (or the number of occurrences) of term t_i in the underlying text or sample, and n the sample size (number of tokens). This first solution tends to overestimate the occurrence probability of terms appearing in the sample, at the expense of the missing terms. Since the occurrence frequency for the latter is 0, its probability would also be 0, as, for example, when an author does not use a given term. We know, however, that the word distribution follows the LNRE law (Large Number of Rare Events [Baayen 2001]), whereby new words always tend to appear. To correct this problem we apply a smoothing technique that also has the advantage of eliminating any special processing resulting from an occurrence probability of 0. This kind of problem could, for example, occur with the Delta formulation [Hoover 2007], or in Eq. (3) (χ^2statistic) when $a_j(t_i)$ equals zero.

As a first approach, Laplace suggests adding one to the numerator in Eq. (5) and likewise adding the vocabulary size to the denominator [Manning and Schütze 2000]. This approach could then be generalized by using a λ parameter (Lidstone's law [Lidstone 1920]), resulting in the following probability estimates: $p = (tf_i + \lambda)/(n + \lambda \cdot |V|)$, with |V| indicating the vocabulary size. In our experimentsge suggest fixing this λ value to 0.1, a choice that avoids assigning a relatively higher probability to rare words, since in authorship attribution rare words are usually not of prime importance. Moreover, in certain circumstances maximum likelihood estimation would be better [Gale and Church 1994], thus justifying a smaller value for the parameter λ. Finally, when compared to the Good-Turing approach [Sampson 2001], this smoothing technique is rather easy to implement.

As an alternative, Zhao and Zobel [2007a, 2007b] suggest using the Dirichlet smoothing method, which estimates occurrence probabilities by applying the following equation. We have

$$p(t_i) = \frac{tf_i}{\mu + n} + \frac{\mu}{\mu + n} \cdot p_B(t_i), \tag{6}$$

where $p_B(t_i)$ is the probability of term t_i in the background model, and μ a parameter applied to adjust the importance of direct estimation versus that of the background model.

With this approach, the resulting estimation relies on a mixture of direct estimation ($tf_i/\mu + n$) and probability provided by the background model B. This latter model is useful when the corresponding frequency tf_i equals 0, or when the size n of the underlying sample is small, often resulting in inaccurate estimates. In such cases, the background model may provide better estimates of the underlying probabilities. To generate the background model used in our experiments we considered all 56,472 articles published in the *Glasgow Herald* or the 58,051 articles in *La Stampa*. The value for the parameter μ was set at $1000^*\sqrt{10}$, because this value achieved the best performance in Zhao and Zobel's [2007a, 2007b] experiments. Assigning a high value to this parameter usually gives more importance to the background model, with the possible μ values typically falling within the range of 0.001 to 10,000 [Zhao 2007].

In our experiments with the English language, we found 19 words in Zhao's [2007] list that could not be found in our corpus. For nine of them, their absence was attributed to the fact that during the preprocessing we expanded the contracted forms (e.g., *aren, isn*, *wasn*, *weren*). The other absences are caused by rare forms (e.g., *hereupon*, *inasmuch*, *whereafter*) not appearing in the GH corpus. As such, our experiments are based on 344 words (363 − 19), and for the Italian language we used a stopword list containing 399 terms.

Using the GH corpus, Table VII compares performances achieved by the KLD approach after applying two different smoothing techniques (Lidstone or Dirichlet) while for the Italian language Table VIII shows the same information. For both corpora

Table VII. Evaluation of KLD Approach with Predefined List of 344 Words (GH corpus, 5,408 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
KLD	Lidstone, λ = 0.1	60.23% ‡	64.14%
KLD	Lidstone, λ = 0.01	**70.80%**	**70.87%**
KLD	Lidstone, λ = 0.001	70.51%	70.27%
KLD	Dirichlet, μ = 0.1	69.75% ‡	68.96% †
KLD	Dirichlet, μ = 10	70.36% †	70.07%
KLD	Dirichlet, μ = 100	67.88% ‡	68.70%
KLD	Dirichlet, μ = 300	68.23% ‡	67.84% †
KLD	Dirichlet, μ = 1000*$\sqrt{}$10	27.27% ‡	23.13% ‡

Table VIII. Evaluation of KLD Approach with Predefined List of 399 Words (*La Stampa* corpus, 4,326 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
KLD	Lidstone, λ = 0.1	75.98% ‡	75.87% †
KLD	Lidstone, λ = 0.01	**84.84%**	**82.84%**
KLD	Lidstone, λ = 0.001	84.51%	82.64%
KLD	Dirichlet, μ = 0.1	83.03% ‡	80.37% ‡
KLD	Dirichlet, μ = 10	84.10% ‡	82.12% ‡
KLD	Dirichlet, μ = 100	84.56%	82.68%
KLD	Dirichlet, μ = 300	83.80% ‡	81.04%
KLD	Dirichlet, μ = 1000*$\sqrt{}$10	34.56% ‡	24.75% ‡

Lidstone's smoothing scheme (λ = 0.01) provides the best performances, although differences resulting from the Dirichlet method (μ = 100) are rather small and not significant. Due to the additional computational costs required in the latter technique (e.g., in estimating background probabilities), we prefer using the Lidstone's approach.

As for the other evaluations, using the best performances as baseline and applying a two-sided sign test, a double cross (‡) indicates a significant performance difference with a significance level α = 1%, while a single cross (†) specifies it at a significance level of 5%. These tests indicate that when using the Dirichlet smoothing method, the best value associated with the parameter μ must be around 100 and this scheme produces performance level similar to the Lidtsone's method (with λ = 0.01).

4.5. Z-Score and Specific Vocabulary

As a new authorship attribution approach, we suggest representing each text based on selected terms (word tokens and punctuation symbols in this study) corresponding to its specific vocabulary, as proposed by Muller [1992]. To define and measure a word's specificity, we need to split the entire corpus into two disjoint parts denoted P_0 and P_1. For a given term t_i, we compute its occurrence frequency both in the set P_0 (value denoted tf_{i0}) and in the second part P_1 (denoted tf_{i1}). In our authorship attribution context, the set P_0 would be the disputed text, while P_1 the rest of the corpus. Thus, for the entire corpus the occurrence frequency of the term t_i becomes $tf_{i0} + tf_{i1}$. The total number of word tokens in part P_0 (or its size) is denoted n_0, similarly with P_1 and n_1, and the size of the entire corpus is defined by $n = n_0 + n_1$.

For any given term t_i the distribution is assumed binomial, with parameters n_0 and $p(t_i)$ representing the probability of the term t_i being randomly selected from the entire corpus. Based on the maximum likelihood principle, this probability would be estimated as follows.

$$p(t_i) = \frac{tf_{i0} + tf_{i1}}{n} \tag{7}$$

As explained in the previous section, a good practice is to smooth the probability estimates [Manning and Schütze 2000]. In this study we applied the Lidstone's technique (with λ = 0.1), simple to implement, and producing reasonably good results [Savoy 2010].

Through repeating this drawing n_0 times we are able to estimate the expected number of occurrences of term t_i in part P_0 using the expression $n_0 \cdot p(t_i)$. We can then compare this expected number to the observed number (namely tf_{i0}), where any large differences between these two values indicate a deviation from the expected behavior. To obtain a more precise definition of *large* we account for variances in the underlying binomial process (defined as $n_0 \cdot p(t_i) \cdot (1 - p(t_i))$). Eq. (8) defines the final standardized Z score (or standard normal distribution N(0,1)) for term t_i, using the partition P_0 and P_1.

$$\text{Z score}(t_{i0}) = \frac{tf_{i0} - n_0 \cdot p(t_i)}{\sqrt{n_0 \cdot p(t_i) \cdot (1 - p(t_i))}} \tag{8}$$

For each selected term, we apply this procedure to weight its specificity according to the underlying text excerpt P_0. Based on the Z score value, we then verify whether this term is used proportionally with roughly the same frequency in both parts (Z score value close to 0). On the other hand, when a term is assigned a positive Z score larger than δ (e.g., 2), we consider it overused or belonging to the specific vocabulary of P_0. A large negative Z score (less than $-\delta$) indicates than the corresponding term is underused in P_0 (or similarly overused in P_1). To illustrate this computation, we have created an small example with six documents written by three authors in the Appendix.

Using this technique, Savoy [2010] was able to determine, for example, the specificity of the vocabulary used by J. McCain and B. Obama during a past U.S. presidential campaign. In these speeches, for example, the terms *jobs*, *health* or *Bush* characterized the Democrat candidate while *nuclear*, *government*, and *judicial* appeared in the specific vocabulary of J. McCain.

Although it might be possible to compute the Z score for all terms, we would suggest ignoring words having a small occurrence frequency (e.g., smaller than 4) or appearing in a limited number of texts (df). In the current context, our English vocabulary is composed of 56,447 distinct word types. When ignoring all words having a term frequency less than 10, having a document frequency (df) less than 3 [Yang and Pedersen 1997], or used by at a single author, we obtain a reduced set of 2,511 types (or 4.4% of the initial vocabulary size). During this selection, we thus remove terms having a small occurrence frequency or appearing in a very limited number of articles. Moreover we also ignore terms used by a single author. This resulting set constitutes the vocabulary (words and punctuation symbols) used in our Z score approach. A similar approach is applied for the Italian corpus. Starting with 102,887 word types, we ignore terms whose term frequency is less than 10 or having a document frequency less than 3. In addition, we also impose that each term must be used by at least two distinct authors. As a result, we obtain a set of 9,825 terms (or 9.5% of the initial vocabulary size).

Given that each author wrote more than one article, we generate an author profile by computing the average term Z scores over all articles corresponding to that author (see Appendix for an example).

When considering two GH columnists sharing certain common subjects (e.g., business) such as Sims and McConnell, the computed Z scores attached to their respective profiles reveal some of their lexical affinities and divergences. According to the Z scores, Sims's ten most significant words are {*profits, group, shares, investment, its, market, income, insurance, though, shareholders*} while for McConnell they are {*trust, company, its, bank, investment, during, value, assets, companies, fund*}. These terms are clearly distinct from the most significant words used by Russell, whose main topics are related to *Arts and Film* ({*film, she, her,* ",", *william, war, he, love, story, is*}).

When inspecting the most significant words in these three author profiles, we are able to find very frequently occurring words (e.g., *its* with an occurrence frequency of 8,251 or *is* with 42,588) as well as words having medium occurrence frequencies, such as *profit* with a term frequency of 577, or *insurance* with 375. When applied to define the most important features in each author profile, the Z score approach does not employ term frequency directly but rather the fact that the occurrence frequency is, in mean, higher or lower in articles written by that given author compared to all other texts. This does not mean, however, that words specific to an author could not appear in another profile (e.g., both *its* and *investment* appear among the most significant terms used by Sims and McConnell).

4.6. Z-Score Distance and Evaluation

The previously defined Z score is assigned to each word (or punctuation symbol) found in a text or an author profile. From these values we define the distance between a query text Q and a given author profile A_j as defined by Eq. (9) and based on a set of terms t_i, for $i = 1, 2, \ldots m$. We have

$$\mathrm{Dist}(Q, A_j) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\mathrm{Z\ score}(t_{iq}) - \mathrm{Z\ score}(t_{ij})\right)^2, \tag{9}$$

where t_{iq} indicates the ith term in the query text, and t_{ij} indicates the ith term in the jth author profile.

When both Z scores are very similar for all terms, the resulting distance is small, meaning that the query text Q was probably written by the jth author. Moreover, the squared difference tends to deduce the impact of any differences less than 1.0, which would mainly occur in the common vocabulary. On the other hand, large differences could occur when both Z scores for a given term are large and have opposite signs. In this case the query text tends, for example, to use the underlying term more frequently than the mean (term specific to the disputed text) while for the jth author, this term is underused. To present this computation, an example is given in the Appendix.

The evaluation of the Z-score-based approach is given in Table IX for the GH corpus and Table X for *La Stampa*. In these tables, we also added the best solutions found with the Delta, χ^2measure, or KLD schemes. Varying the value for the parameter λ (1.0 or 0.1 in the current study) seems to have no real impact on both corpora, yet when compared to the other models, the Z score method produces better performance levels both for the document-based (micro-average) and author-based (macro-average) measures.

Applying the sign test while using best performances as the baseline, we add a cross (†) when detecting a significance difference at a significance level α = 5% (two-sided) or a double cross (‡) when the significance level α = 1%. As shown in Tables IX and X, the Z score approach performs significantly better than the χ^2measure when considering both measures and corpora. Using the micro-average indicator, the performance differences are statistically significant between the other approaches and the Z score model. With the macro-averaging scheme, the performance difference is significantly different with the Delta model for both corpora, and when using the *Glasgow Herald*, the performance difference is also significant with the KLD model.

Unlike the Z score, the other three authorship attribution methods rely mainly on function words or very frequent word types. In the Delta approach [Burrows 2002], the selection criterion is based on term frequency information. When considering only terms occurring with high frequencies, in both in English or Italian languages, we

Table IX. Evaluation of Z Score Approach together with Best Solutions Obtained by Other Authorship Attribution Schemes (GH corpus, 5,408 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
Z score	Lidstone, λ = 1	**81.73%**	**79.28%**
Z score	Lidstone, λ = 0.1	81.71%	79.26%
Delta	400 words	63.70% ‡	66.14% †
χ^2 measure	2-limit	65.26% ‡	63.57% ‡
KLD	Lidstone, λ = 0.01	70.80% ‡	70.87% †

Table X. Evaluation of Z Score Approach together with Best Solutions Obtained by Other Authorship Attribution Schemes (*La Stampa* corpus, 4,326 articles, 20 authors)

Method	Parameter	Micro-average	Macro-average
Z score	Lidstone, λ = 1	**89.71%**	**88.06%**
Z score	Lidstone, λ = 0.1	**89.71%**	**88.06%**
Delta	400 words	76.07% ‡	75.08% †
χ^2 measure	2-limit	68.28% ‡	65.78% ‡
KLD	Lidstone, λ = 0.01	84.84% ‡	82.84%

mainly extract determiners, prepositions, conjunctions, pronouns, and auxiliary verb forms, all belonging to parts-of-speech defining functional words as stated by Miranda Garcia and Calle Martin [2007]. As a second authorship attribution method, we also evaluated the χ^2 measure [Grieve 2007], based on word types and punctuation symbols respecting a minimal document frequency. In this case, one of the best performances is achieved when considering all words and punctuation symbols appearing in at least two of every possible author's texts. As a third baseline Zhao and Zobel [2007a, 2007b] suggest using the KLD scheme with a predefined feature list (containing 363 English terms or 399 Italian words). This type of list corresponds to a stopword list in the IR domain [Fox 1990], often applied to identify very frequently appearing forms having no clear and important meaning. It is known, however, that for a given language different stopword lists might be suggested with possibly different retrieval effectiveness [Dolamic and Savoy 2010].

Within the Z score approach and like the χ^2measure, we do not apply a predefined selection strategy. Using words as they appear in the underlying texts would provide the information needed to more or less weight each selected feature. When some word types are not used (e.g., *hereafter*, *hereupon*), we could simply ignore them, and this could also apply to word types having a small term frequency (*tf*) or having a small document frequency (*df*) as suggested by Yang and Pedersen [1997] and applied in this study. On the other hand, word forms (e.g., acronyms) occurring frequently in a corpus and capable of discriminating between authors must be selected in a manner causing them to improve the overall quality of the authorship attribution scheme (e.g., *SNP* (Scottish National Party) or *MPs* (Member of Parliament) in the current study). Simply considering more terms is not the best strategy, however, as demonstrated by the results shown in Tables III and IV (Delta method), where 600 or 800 words produced a lower performance level than 400 words.

4.7. Additional Experiments

So far we have used all authors and articles occurring in our corpora without distinguishing them according to the main topics. We can argue that considering only authors on a given subject will render the authorship attribution more difficult. To evaluate this argument, we have extracted from the *Glasgow Herald* (see Table I) the five authors who wrote on *Business* (namely Young, McConnell, Reeves, Sims, and

Table XI. Evaluation of Z Score Approach Using Two Subsets of the GH Corpus (on the left on business, on the right on sports)

	Business		Sports	
Method, Parameter	Micro-average	Macro-average	Micro-average	Macro-average
Delta, 400	69.58% ‡	66.14%	80.85% ‡	80.74%
χ^2, 2−limit	61.80% ‡	64.78%	79.98% ‡	80.27%
KLD, λg 0.01	80.62% ‡	80.92%	83.38% ‡	83.57%
Z score, λg 0.1	**87.21%**	**86.66%**	**92.38%**	**92.25%**

Table XII. Evaluation of Z Score Approach Using Two Subsets of the *La Stampa* Corpus (on the left on politics, on the right on sports)

	Politics		Sports	
Method, Parameter	Micro-average	Macro-average	Micro-average	Macro-average
Delta, 400	77.34% ‡	77.77% †	67.20% ‡	64.73%
χ^2, 2−limit	74.73% ‡	75.10% ‡	77.45% ‡	77.24%
KLD, λg 0.01	88.60% ‡	89.50%	95.06% ‡	94.41%
Z score, λg 0.1	**92.15%**	**91.31%**	**97.72%**	**97.67%**

Wilson), and the five journalists who wrote on *Sports* (Douglas, Gallacher, Gillon, Paul, and Traynor). Under the *Business* subject, we can find 1775 articles, and 1943 under the *Sports* headline.

With the newspaper *La Stampa* (see Table II), we have also extracted two subcorpora. The first one is composed by four journalists who wrote on *Sports* (Ansaldo, Beccantini, Del Buono, and Ormezzano) while the second contains *Political* articles written by ten columnists (Battista, Benedetto, Galvano, Gramellini, Meli, Nirenstein, Novazio, Pantarelli, Passarini, and Spinelli). The subset covering *Sports* contains 1317 articles while the *Politics* headline occurs in 2026 papers.

When applying the four authorship attribution methods on these subsets, we obtained the accuracy rates reported in Table XI for the *Glasgow Herald*, and in Table XII for the *La Stampa*. The evaluations done on these subsets reveal similar conclusions to those obtained on the whole corpus. The Z score method shows the best performance, that is also statistically significant when compared using the micro-averaging method (a significance level of 5% is indicated by †, and 1% with ‡). Under the macro-average measure, the number of authors is too small to detect any significant performance differences when using the *sports* or *business* subsets.

From the results reported in Tables XI and XII, we can conclude that limiting our corpus to articles written in a given domain does not change our previous conclusions. The Z score scheme tends to produce the best overall accuracy rate. The performance differences are statistically significant under the micro-average measure. When applying the macro-averaging evaluation technique, the number of authors is rather limited and thus the statistical test cannot usually detect any significant differences.

As a second additional set of experiments, we can evaluate the four authorship attribution schemes using exactly the same set of terms instead of applying their own selection method. To achieve this, we have considered choosing all terms having a document frequency (*df*) larger than or equal to a given threshold δ, for δ = 400, 200, 100, and 50. Using this criterion, we tend to favor terms appearing in many different articles. A high threshold value limits the number of terms used in the evaluations, and, by decreasing this threshold, we will consider more terms. We also applied a similar selection procedure using the term frequency (*tf*, the occurrence frequency in the underlying corpus) with different threshold values. The effectiveness achieved with the *Glasgow Herald* under these two selection procedures is depicted

Table XIII. Accuracy Rate (micro-average) of Four Authorship Attribution Schemes Using the Same Terms According to Different Document Frequency (*df*) or Term Frequency (*tf*) Thresholds (GH corpus 5,408 articles, 20 authors)

Selection \ Number of terms	$df \geq 400$ 715	$df \geq 200$ 1,511	$df \geq 100$ 2,827	$df \geq 50$ 4,710
Delta	45.75% ‡	25.57% ‡	9.36% ‡	6.56% ‡
χ^2	63.50% ‡	49.43% ‡	45.67% ‡	47.69% ‡
KLD	**81.82%**	78.20% ‡	66.48% ‡	52.98% ‡
Z score	81.03%	**83.43%**	**85.80%**	**88.05%**
Selection Number of terms	$tf \geq 500$ 784	$tf \geq 300$ 1,297	$tf \geq 150$ 2,434	$tf \geq 50$ 5,433
Delta	48.89% ‡	30.51% ‡	13.81% ‡	7.71% ‡
χ^2	56.07% ‡	47.98% ‡	45.69% ‡	47.89% ‡
KLD	**81.36%**	80.05% ‡	70.23% ‡	51.16% ‡
Z score	80.57%	**83.15%**	**84.80%**	**87.44%**

Table XIV. Accuracy Rate (micro-average) of Four Authorship Attribution Schemes Using the Same Terms According to Different Document Frequency (*df*) or Term Frequency (*tf*) Thresholds (*La Stampa* corpus, 4,326 articles, 20 authors)

Selection \ Number of terms	$df \geq 400$ 516	$df \geq 200$ 1,171	$df \geq 100$ 2,406	$df \geq 50$ 4,470
Delta	61.60% ‡	44.27% ‡	23.37% ‡	19.56% ‡
χ^2	78.09% ‡	67.85% ‡	57.72% ‡	59.18% ‡
KLD	**91.93%**	90.98% ‡	82.22% ‡	62.88% ‡
Z score	91.59%	**92.70%**	**93.09%**	**93.99%**
Selection Number of terms	$tf \geq 400$ 689	$tf \geq 200$ 1,482	$tf \geq 100$ 2,832	$tf \geq 50$ 5,183
Delta	62.88% ‡	48.73% ‡	24.18% ‡	21.71% ‡
χ^2	69.39% ‡	67.24% ‡	63.04% ‡	64.93% ‡
KLD	**91.26%**	88.74% ‡	79.15% ‡	65.74% ‡
Z score	89.00%	**90.75%**	**91.70%**	**94.17%**

in Table XIII, and Table XIV shows the same information using the Italian corpus. In both tables, only the micro-average measure was computed.

In Tables XIII and XIV, we added a double cross (‡) to indicate a significant performance difference based on the sign test (significance level α = 1%, two-sided), using the performance achieved by the Z score as baseline. The data depicted in these tables indicate that the Z score scheme usually achieves the best accuracy rate. When comparing the Z score to other strategies, the performance differences are usually statistically significant. Only when the number of terms is limited (between 500 to 800) are the performance differences not statistically significant between the Z score and the KLD scheme.

Tables XIII and XIV also show that when the number of terms increases, the performance tends to decrease for all schemes except for the Z score. This decrease is clearly marked for the Delta approach, less so for the χ^2 and KLD approaches. For the Z score scheme, increasing the number of terms leads to a slightly improved performance. Overall, the performance of the Z score method seems to be more stable with a different number of terms used to represent the texts and author profiles.

4.8. Naïve Bayes

Until now, we have presented authorship attribution methods following the classical paradigm. In this vein, we have first selected a set of relevant terms. Then, based on a distance measure between the query text representation and author profiles, we have defined the probable author as the one that depicts the smallest distance.

As another paradigm, we can apply a machine learning approach [Sebastiani 2002]. In this case, we first need to define a selection criterion to reduce the number of possible terms (term space reduction). This step is useful to reduce the computational cost and to reduce the overfitting of the learning scheme to the training data. In a second step, we use the training data to let the classifier learn from positive and negative examples. In the current study, the training data will be formed by the whole corpus minus the query text (leaving-one-out).

As an effective approach to text classification, we may use the Support Vector Machine (SVM) model [Cristianini and Shawe-Taylor 2000; Joachims 2002]. This is an adapted solution for binary classification problems where the SVM determines the hyperplane that best separates the examples belonging to the two categories. In this case *best* hyperplane refers to having the largest separation (or margin) between the two classes (together with the reduction of the number of incorrect classifications). However, in our context of applying the SVM approach on 20 categories, it requires a combination of several binary SVM classifiers (with different possible variants [Duan and Keerthi 2005]). Moreover, predicting the most effective text representation is a rather difficult task (e.g., various stemmers weighting schemes, normalizations, and kernel functions). Finally, as mentioned in the Introduction, the effectiveness is not our main objective and we rather focus on a simple learning scheme able to explain its decisions. This last requirement is not fully achieved by an SVM approach.

As another typical and simpler text classifier derived from the machine learning paradigm, we choose the naïve Bayes model [Mitchell 1997] to determine the possible author between the set of 20 possible journalists (or hypotheses), denoted by A_i for $i = 1, 2, \ldots r$. To define the probable author of a query text Q, the naïve Bayes model selects the one maximizing Eq. (10), in which tq_j represents the jth term included in the query text Q, and n_q indicates the size of the query text.

$$\text{Arg max}_{A_i} \ \text{Prob}[A_i|Q] = \text{Prob}[A_i] \cdot \prod_{j=1}^{n_q} \text{Prob}[tq_j|A_i] \tag{10}$$

To estimate the prior probabilities (Prob$[A_i]$), we simply take into account the proportion of articles written by each author. To determine the term probabilities we regroup all texts belonging to the same author to form the author profile. For each term t_j, we then compute the ratio between its occurrence frequency in the corresponding author profile A_i (tf_{ji}) and the size of this sample (n_i).

$$\text{Prob}[t_j|A_i] = {}^{tf_{ij}}/_{n} \tag{11}$$

This definition (see Eq. (11)) tends to overestimate the probabilities of terms occurring in the text with respect to missing terms. For the latter, the occurrence frequency (and probability) was 0, so a smoothing approach had to be applied to correct this. As for the other methods, we will apply Lidstone's law through smoothing each estimate as $\text{Prob}[tq_j|A_i] = (tf_{ji} + \lambda)/(n_i + \lambda \cdot |V|)$, with λ as a parameter (set to 0.1), and |V| indicating the vocabulary size.

As a selection criterion, various measures have been suggested and evaluated. Following Sebastiani [2002], we have selected the Odds Ratio (OR), a selection function found historically effective. For each term t_j, for $j = 1, 2, \ldots m$, and each author A_i for $i = 1, 2, \ldots r$, we can compute the odds ratio defined by Eq. (12). In this formulation, Prob$[t_j|A_i]$ indicates the probability that, for a random document, the term t_j appears

Table XV. Accuracy Rate (micro-average) of the Naïve Bayes and Z Score According to Different Number of Terms Selected (GH corpus, 5,408 articles, 20 authors)

\ Method Nb terms \ Selection	Naïve Bayes OR SUM	Naïve Bayes *df*	Z score *df*
500	46.26% ‡	69.88% ‡	**78.53%**
1,000	57.78% ‡	79.40% ‡	**82.13%**
2,000	65.34% ‡	83.27% ‡	**84.54%**
4,000	73.32% ‡	84.78% ‡	**87.37%**

Table XVI. Accuracy Rate (micro-average) of the Naïve Bayes and Z Score According to Different Number of Terms Selected (*La Stampa* corpus, 4,326 articles, 20 authors)

\ Method Nb terms \ Selection	Naïve Bayes OR SUM	Naïve Bayes *df*	Z score *df*
500	69.37% ‡	78.16% ‡	**91.12%**
1,000	76.40% ‡	85.71% ‡	**92.16%**
2,000	78.64% ‡	90.08% ‡	**93.00%**
4,000	81.88% ‡	91.59% ‡	**93.57%**

knowing that this text was written by author A_i. Similarly, Prob$[t_j|\neg A_i]$ indicates the same probability except that the underlying document was not written by author A_i.

$$OR(t_j, A_i) = \frac{\text{Prob}[t_j|A_i] \cdot (1 - \text{Prob}[t_j|\neg A_i])}{(1 - \text{Prob}[t_j|A_i]) \cdot (\text{Prob}[t_j|\neg A_i]} \tag{12}$$

If a given term t_j appears mainly in the author profile A_i, the probability Prob$[t_j|A_i]$ will be relatively high and, in contrast, the probability Prob$[t_j|\neg A_i]$ will be relatively small. As shown in Eq. (12), this phenomenon will assign a relatively high value for the numerator compared to the denominator. The resulting OR value will be high. The corresponding term t_j is then viewed as able to discriminate between the author A_i and the other possible writers.

Eq. (12) returns a value for each pair (term, author). In order to compare and rank each term, we need a single value able to consider the term's discriminative capability over all categories (or authors in the current context). To aggregate the r values, one for each author, Sebastiani [2002] indicates that the SUM operator (see Eq. (13)) tends to produce the best results with the OR used as term selection function.

$$OR_{sum}(t_j) = \sum_{i=1}^{r} OR(t_j, A_i) \tag{13}$$

Using this machine learning scheme with our corpora, we achieved the micro-average performances depicted in Table XV for the *Glasgow Herald*, and in Table XVI for *La Stampa*. In a first evaluation, we have considered the naïve Bayes with the OR SUM as selection procedure. In a second experiment, we used the document frequency (*df*) as a selection function to rank all possible features, from the highest to the lowest. In this case, we favor terms appearing in many articles over those occurring in a limited number of documents. Such a selection function is simple and efficient to apply and has been found effective in text classification applications [Yang and Pedersen 1997]. The same selection procedure was applied to define terms used with the Z score method (performances reported in the last column). In these tables, we also added a double cross (‡) to indicate a significant performance difference based on the sign test (significance level α = 1%, two-sided), using the performance achieved by the Z score as a baseline.

The performances shown in these tables indicate that the Z score scheme achieves the best accuracy rate. The performance differences with the naïve Bayes model tend to be statistically significant. Under the naïve Bayes method, the performance differences between the two selection procedures are relatively large, indicating that the term selection stage represents an important choice to achieving high performance. Finally, when the number of terms selected increases, the performance differences between the naïve Bayes and the Z score tend to be reduced.

4.9. Assignment Reliability

In Eq. (9), we define the Z score distance between two texts, or in our context between a disputed text and an author profile. When handling several possible authors, the suggested strategy is to assign the article to the author having the minimal Z score distance. If this resulting minimum value is small, we are more confident that the corresponding author is the real author of the disputed document. On the other hand, if the minimum mean squared difference is large, the author assignment must be viewed as more doubtful.

In order to verify this assumption, we need a mean to predict the probability of a correct assignment according to the minimum Z score distance computed from a set of possible author profiles. To achieve this objective, we suggest using the logistic regression approach [Hosmer and Lemeshow 2001], a statistical methodology used to predict the probability of a binary outcome variable according to a set of explanatory variables. In our context, we need to predict the probability of a correct assignment based on a single explanatory variable, namely the minimum Z score distance. The resulting model is defined according the following equation. We have

$$\text{Prob}\left[\text{Assignement}_j \text{ is correct} \mid \text{Dist}_j\right] = \pi\ (\text{Dist}_j) = \frac{e^{\alpha+\beta\cdot\text{Dist}_j}}{1 + e^{\alpha+\beta\cdot\text{Dist}_j}} \tag{14}$$

within which Dist_j is the minimum Z score distance corresponding to author profile A_j.

In this equation, the coefficients α (intercept) and β (slope) are unknown parameters which fit the S-curve shown in Figure 1. The value of these coefficients is estimated according the principle of maximum likelihood (the required computations are done using the R package).

When using the *Glasgow Herald* corpus, the estimations return α = 2.31 and β = −0.499. To examine the fit adequacy, we can use a single overall goodness-of-fit statistic (Wald test [Hosmer and Lemeshow 2001]), as well as a test to assess the significance of each coefficient. In our study, the entire logistic model is significant and, for each coefficient, the null hypothesis stating that the corresponding value is equal to zero is always rejected (significance level α = 1%). Using these estimates, the probability that the assignment is correct when obtaining a minimum Z score distance of 1 is 85.98% (see Eq. (14)). As depicted in Figure 1, this probability decreases when the minimum Z score distance increases, as, for example, with a distance of 4, the resulting probability is 57.86%, or only 33.6% when faced with a distance of 6 between an author profile and a disputed text.

5. CONCLUSION

Text classification tasks involve numerous interesting challenges, particularly when applied to authorship attribution. This article suggests a simple method based on word usage in texts written by different authors. To evaluate and compare our suggested

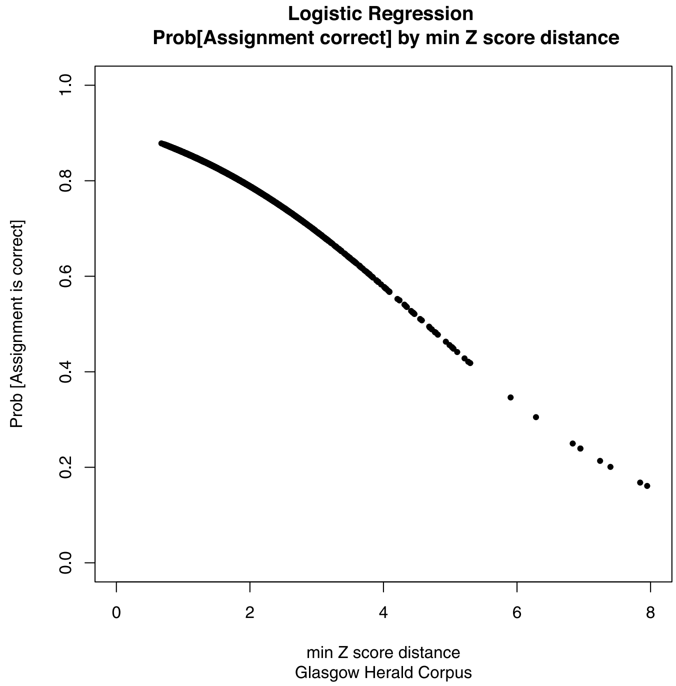

Fig. 1. Logistic regression given the probability that the assignment is correct according to the minimum Z score distance.

scheme with other approaches, we used articles contained in a freely available newspaper corpus written in English (*Glasgow Herald,* published in 1995). To complement this first experiment, a second corpus written in the Italian language (*La Stampa,* published in 1994) is also used. From these corpora we extract all articles written by 20 well-known columnists or journalists having published numerous articles.

For comparison purposes, we used the Delta method [Burrows 2002; Hoover 2004a] based on the 40 to 800 most frequent word types, where for both languages the highest accuracy rate was obtained with the top 400 most frequent types. As a second authorship attribution method we also evaluated the χ^2measure [Grieve 2007], based on word types and punctuation symbols respecting a minimal document frequency on a per-author basis. In this case, one of the best performances was achieved when considering all words and punctuation symbols appearing in at least two texts for each author. As a third baseline, we used the KLD scheme proposed by Zhao and Zobel [2007a, 2007b] and based on a predefined set of 344 words in English, or 399 Italian terms. This last approach results in better performance levels that the Delta and χ^2 measure schemes. These three baselines do, however, produce accuracy rates that are inferior to those obtained by the suggested Z scores. Finally, when comparing with the naïve Bayes model, we show that the performances achieved by the Z score method are better than those obtained with this well-known machine learning approach.

Using frequent word types as well as function words might be useful in authorship attribution, but the proposed Z score method selects features (word types and punctuation symbols in our study) according to their distinct distributions in the underlying texts. Our work focuses on a simple approach producing results that can be easily interpreted and require only certain easy to understand parameter settings (e.g.,

Table XVII. Frequency of Occurrence of Six Word Types over the Seven Documents

	Q	A1	A2	B1	B2	C1	C2	Sum	Prob.
the	85	97	106	171	185	246	254	1059	0.554
of	48	48	56	89	98	157	145	593	0.310
from	5	4	6	12	13	28	27	90	0.046
year	0	0	0	2	3	7	9	21	0.010
we	5	7	4	21	30	0	1	63	0.033
I	8	9	10	32	37	1	0	89	0.047
Sum	151	165	182	327	366	439	436	1915	

ignoring word types below a given document frequency (*df*) or the value of the smoothing parameter).

It is our opinion that these computer-based methods should not be viewed as the only devices capable of recognizing the real or ghost author behind a text. They should rather be viewed as complementary methods, especially given that none of them is able to determine the right author with absolute certainty in all cases. Such computational linguistic approaches could be reserved as signals that complement additional evidence obtained from other useful sources of external information (incipits, titles, diaries, correspondence, publishers' records), biographical information, classical stylistic methods (synonyms, prosody, metre), along with earlier attribution studies [Love 2002].

APPENDIX

In order to illustrate the computation of the Z score approach, we have built a small example composed of six documents written by three authors denoted as A, B, and C. To indicate the corresponding author of each paper, we add the letter A, B, or C in each document's identifier. As we can see in Table XVII, we have first the query text (denoted by Q) followed by two documents for each possible author. For each paper, we count the occurrence frequency of six word types. In the last line, we indicate the size of each paper as the sum of these frequencies. Based on this information, we can see that the longest paper is document C1, while the shortest is Q. In the column "Sum," we indicate the number of occurrences of each word type in the corpus formed by papers A1 to C2. The most frequent word type is the determinant *the*, followed by the preposition *of*. Finally, the size of this corpus is 1,915.

Now we want to determine the possible author of query text Q. According to the explanation given in Section 4.5, we consider two parts in our corpus; the first, denoted as P_0, corresponds to the single document Q, and P_1 regroups the six documents (A1, A2, B1, B2, C1, and C2). According to Eq. (7), we can estimate the occurrence probability of each word type as its occurrence frequency in parts P_0 and P_1 divided by the size of the corpus (n = 1915 + 151 = 2066). For the determinant *the*, this estimate is (85 + 1059) / 2066 = 1144 / 2066 = 0.554. In Table XVII, we have added these estimations in the last column under the label "Prob".

To compute the Z score of each word type and for each document, we applied Eq. (8). For the word type *the* and document Q, we obtain

$$\text{Z score}(the, Q) = \frac{85 - 151 \cdot 0.554}{\sqrt{151 \cdot 0.554 \cdot (1 - 0.544)}} = \frac{85 - 83.613}{\sqrt{37.314}} = 0.227$$

We repeat this computation for all remaining word types and documents to get the Z score values depicted in Table XVIII.

To determine the possible author of document Q, we will compare the Z score values obtained from the query document to the different author profiles. To define an

Table XVIII. Z Score Values of Each Word Type According to the Seven Papers

	Q	A1	A2	B1	B2	C1	C2
the	0.227	0.882	0.779	−1.120	−1.857	0.280	1.211
of	0.202	−0.537	−0.075	−1.489	−1.758	2.145	1.007
from	−0.755	−1.333	−0.838	−0.802	−0.956	1.781	1.590
year	−1.245	−1.302	−1.367	−0.730	−0.375	1.208	2.181
we	0.014	0.685	−0.827	3.173	5.260	−3.865	−3.584
I	0.350	0.461	0.510	4.352	4.897	−4.425	−4.635

Table XIX. Z Score Values of the Query Text and the Three Author Profiles

	Q	A	B	C
the	0.227	0.831	−1.489	0.746
of	0.202	−0.306	−1.623	1.576
from	−0.755	−1.086	−0.879	1.685
year	−1.245	−1.334	−0.553	1.694
we	0.014	−0.071	4.217	−3.725
I	0.350	0.486	4.624	−4.530

Table XX. Details of the Computation of the Distance between the Query Text and the Three Author Profiles

	A	B	C
the	0.364	2.944	0.269
of	0.259	3.333	1.887
from	0.109	0.015	5.954
year	0.008	0.480	8.641
we	0.007	17.664	13.974
I	0.018	18.268	23.814
Distance	0.128	7.117	9.090

author profile, we simply compute the average of the Z score values for each word type obtained for all papers written by that author. For example, for the preposition *of* and the author C, the resulting Z score is (2.145 + 1.007) / 2 = 1.576. Table XIX shows the corresponding Z score values for the other word types and authors.

Finally, we need to compute the Z score distance between the query text Q and the three profiles according to Eq. (9). For the word type *the* and author A, we calculate the Z scores difference (0.227 − 0.831), and take the power of two of this difference ($-0.604^2 = 0.364$). These intermediate values are depicted in Table XX for the other word types and author profiles.

The overall distance between the query text and a given author profile is the average over all word types. In our example, this average is 0.128 with the author profile A, 7.117 with B, and 9.09 with the last possible writer. The Z score scheme suggests that the probable author of document Q is author A, the one depicting the smallest distance.

REFERENCES

ARGAMON, S. 2006. Introduction to the special topic selection on the computational analysis of style. *J. Amer. Soc. Inf. Sci. Technol.* 57, 11, 1503–1505.

ARGAMON, S. 2008. Interpreting Burrows's delta: Geometric and probabilistic foundations. *Liter. Linguist. Comput. 23*, 2, 131–147.

ARGAMON, S., KOPPEL, M., PENNEBAKER, J. W., AND SCHLER, J. 2009. Automatically profiling the author of an anonymous text. *Comm. ACM 52*, 2, 119–123.

BAAYEN, H. R. 2001. *Word Frequency Distributions*. Kluwer Academic Press, Dordrecht.

BAAYEN, H. R. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

BAAYEN, H. R. AND HALTEREN, H. V. 2002. An experiment in authorship attribution. In *Proceedings of the 6th International Conference on Statistical Analysis of Texttual Data (JADT'2002)*. 69–75.

BILISOLY, R. 2008. *Practical Text Mining with Perl*. John Wiley Sons, Hoboken, NJ.

BINONGA, J. N. G. AND SMITH, M. W. 1999. The application of principal component analysis to stylometry. *Liter. Linguist. Comput. 14*, 4, 445–465.

BISHOP, C. M. 2007. *Pattern Recognition and Machine Learning.* Springer.

BRILL, E. 1995. Transformation-Based error driven learning and natural language processing: A case study in part-of-speech tagging. *Comput. Linguist. 21*, 4, 543–565.

BURROWS, J. F. 1992. Not unless you ask nicely: The interpretative nexus between analysis and information. *Liter. Linguist. Comput. 7*, 1, 91–109.

BURROWS, J. F. 2002. Delta: A measure of stylistic difference and a guide to likely authorship. *Liter. Linguist. Comput. 17*, 3, 267–287.

CARPENTER, R. H. AND SELTZER, R. V. 1970. On Nixon's Kennedy style. *Speaker and Gavel 7*, 41–43.

CONOVER, W. J. 1980. *Practical Nonparametric Statistics* 2nd Ed. John Wiley and Sons, New York.

CRAIG, H. AND KINNEY, A. F. Eds. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press.

CRAWLEY, M. J. 2007. *The R Book*. John Wiley and Sons, Chichester.

CRISTIANINI, N. AND SHAWE-TAYLOR, J. 2000. *An Introduction to Support Vector Machines*. Cambridge University Press.

DIXON, P. AND MANNION, D. 1993. Goldsmith's periodical essays: A statistical analysis. *Liter. Linguist. Comput. 8*, 1, 1–19.

DOLAMIC, L. AND SAVOY, J. 2010. When stopword lists make the difference. *J. Amer. Soc. Inf. Sci. Technol. 61*, 1, 200–203.

DUAN, K.-B. AND KEERTHI, S. S. 2005. Which is the best multiclass SVM method? An empirical study. In *Proceedings of the 6th International Workshop on Multiple Classifier System*. 278–285.

EFRON, B. AND THISTED, R. 1976. Estimating the number of unseen species: How many words did Shakespeare know? *Biomerika 63*, 3, 435–447.

FAUTSCH, C. AND SAVOY, J. 2009. Algorithmic stemmers or morphological analysis: An evaluation. *J. Amer. Soc. Inf. Sci. Technol. 60*, 8, 1616–1624.

FINN, A. AND KUSHMERICK, N. 2005. Learning to classify documents according to genre. *J. Amer. Soc. Inf. Sci. Technol. 57*, 11, 1506–1518.

FOX, C. 1990. A stop list for general text. *ACM SIGIR Forum 24*, 19–35.

FRANCIS, W. N. AND KUC̀ERA, H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, MA.

GALE, W. A. AND CHURCH, K. W. 1994. What is wrong with adding one? In *Corpus-Based Research into Language*, N. Oostdijk and P. de Hann Eds., Harcourt Brace.

GREENACRE, M. 2007. *Correspondence Analysis in Practice* 2nd Ed. Chapman and Hall/CRC, Boca Raton, FL.

GREFENSETTE, G. AND TAPANAINEN, P. 1994. What is a word? What is a sentence? Problems of tokenization. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*.

GRIEVE, J. 2007. Quantitative authorship attribution: An evaluation of techniques. *Liter. Linguist. Comput. 22*, 3, 251–270.

HARMAN, D. 1991. How effective is suffixing? *J. Amer. Soc. Inf. Sci. 42*, 1, 7–15.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2009. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction* 2nd Ed. Springer, New York.

HOLMES, D. I. 1992. A stylometric analysis of Mormon scripture and related texts. *J Roy. Statist. Soc. A155*, 1, 91–120.

HOLMES, D. I. 1998. The evolution of stylometry in humanities scholarship. *Liter. Linguist. Comput. 13* , 3, 111–117.

HOLMES, D. I. AND FORSYTH, R. S. 1995. *The Federalist* revisited: New directions in authorship attribution. *Liter. Linguist. Comput. 10*, 2, 111–127.

HOLMES, D. I. AND CROFTS, D. W. 2010. *The Diary of a Public Man*: A Case Study in Traditional and Non-Traditional Authorship Attribution. *Liter. Linguist. Comput. 25*, 2, 179–197.

HOLTE, R. C. 1993. Very simple classification rules perform well on most commonly used datasets. *Mach. Learn. 11*, 1, 63–90.

HOOVER, D. L. 2003. Another perspective on vocabulary richness. *Comput. Humanit. 37*, 151–178.

HOOVER, D. L. 2004a. Delta prime? *Liter. Linguist. Comput. 19*, 4, 477–495.

HOOVER, D. L. 2004b. Testing Burrows's delta. *Liter. Linguist. Comput. 19*, 4, 453–475.

HOOVER, D. L. 2006. Stylometry, chronology and the styles of Henry James. In *Proceedings of the Digital Humanities Conference*. 78–80.

HOOVER, D. L. 2007. Updating delta and delta prime. Graduate School of Library and Information Science, University of Illinois, 79–80.

HOOVER, D. L. AND HESS, S. 2009. An exercise in non-ideal authorship attribution: The mysterious Maria Ward. *Liter. Linguist. Comput. 24*, 4, 467–489.

HOSMER, D. AND LEMESHOW, S. 2001. *Applied Logistic Regression* 2nd Ed. John Wiley and Sons, New York.

JOACHIMS, T. 2002. *Learning to Classify Text Using Support Vector Machines. Methods, Theory, and Algorithms*. Kluwer, Boston.

JOCKERS, M. L. AND WITTEN, D. M. 2010. A comparative study of machine learning methods for authorship attribution. *Liter. Linguist. Comput. 25*, 2, 215–223.

JOCKERS, M. L., WITTEN, D. M., AND CRIDDLE, C. S. 2008. Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification. *Liter. Linguist. Comput. 23*, 4, 465–491.

JOHNSON, K. 2008. *Quantitative Methods in Linguistics*. Blackwell, Malden, MA.

JUOLA, P. 2006. Authorship attribution. *Found. Trends Inf. Retriev. 1*, 3.

KEŠELJ, V., PENG, F., CERCONE, N., AND THOMAS C. 2003. N-Gram-Based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING'03)*. 255–264.

KNUTH, D. E. 1981. *The Art of Computer Programming*, *Vol. 2 Seminumerical Algorithms*. Addison-Wesley, Reading, MA.

KOPPEL, M., SCHLER, J., AND ARGAMON, S. 2009. Computational methods in authorship attribution. *J. Amer. Soc. Inf. Sci. Technol. 60*, 1, 9–26.

LABBÉ, D. 2001. Normalisation et lemmatisation d'une question ouverte. *J. Soc. Franc. Statist. 142*, 4, 37–57.

LABBÉ, D. 2007. Experiments on authorship attribution by intertextual distance in English. *J. Quant. Linguist. 14*, 1, 33–80.

LEDGER, G. AND MERRIAM, R. 1994. Shakespeare, Fletcher, and *The Two Noble Kinsmen*. *Liter. Linguist. Comput. 9*, 3, 235–248.

LIDSTONE, G. J. 1920. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Trans. Faculty Actuar. 8*, 182–192.

LOVE, H. 2002. *Attributing Authorship: An Introduction*. Cambridge University Press.

MANNING, C. D. AND SCHÜTZE, H. 2000. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

MARCUS, M. P., SANTORINI, B., AND MARCINKIEWICZ, M. A. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist. 19*, 2, 313–330.

MCNAMEE, P. AND MAYFIELD, J. 2004. Character n-gram tokenization for European language text retrieval. *Inform. Retrieval 7*, 1–2, 73–97.

MERRIAM, T. 1998. Heterogeneous authorship in early Shakespeare and the problem of Henry V. *Liter. Linguist. Comput. 13*, 15–28.

MIRANDA-GARCIA, A. AND CALLE-MARTIN, J. 2005. Yule's characteristic K revisited. *Lang. Resour. Eval. 39*, 4, 287–294.

MIRANDA GARCIA, A. AND CALLE MARTIN, J. 2007. Function words in authorship attribution studies. *Liter. Linguist. Comput. 22*, 1, 49–66.

MITCHELL, T. M. 1997. *Machine Learning*. McGraw-Hill, New York.

MORTON, A. Q. 1986. Once. A test of authorship based on words which are not repeated in the sample. *Liter. Linguist. Comput. 1*, 1, 1–8.

MOSTELLER, F. AND WALLACE, D. L. 1964. *Inference and Disputed Authorship, The Federalist.* Addison-Wesley, Reading, MA. Reprint 2007.

MULLER, C. 1992. *Principes et Méthodes de Statistique Lexicale*. Honoré Champion, Paris.

MURTAGH, F. 2005. *Correspondence Analysis and Data Coding with Java and R*. Chapman and Hall/CRC, Boca Raton, FL.

NUGUES, P. 2006. *An Introduction to Language Processing with Perl and Prolog*. Springer, Berlin.

PETERS, C. 2001. *Cross-Language Information Retrieval and Evaluation*. Lectures Notes in Computer Science, vol. 2069, Springer.

PETERS, C., GONZALO, J., BRASCHLER, M. AND KLUCK, M. 2004. *Comparative Evaluation of Multilingual Information Access Systems*. Lectures Notes in Computer Science, vol. 3237, Springer.

PORTER, M. F. 1980. An algorithm for suffix stripping. *Program 14*, 3, 130–137.

SAMPSON, G. 2001. *Empirical Linguistics*. Continuum, London, UK.

SAVOY, J. 2001. Report on CLEF-2001 experiments. In *Cross-Language Information Retrieval and Evaluation*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck Eds., Lectures Notes in Computer Science, vol. 2069, Springer, 27–43.

SAVOY, J. 2010. Lexical analysis of US political speeches. *J. Quant. Linguist. 17*, 2, 123–141.

SEBASTIANI, F. 2002. Machine learning in automatic text categorization. *ACM Comput. Surv. 14*, 1, 1–27.

SICHEL, H. S. 1975. On a distribution law for word frequencies. *J. Amer. Statist. Assoc. 70*, 351, 542–547.

STAMATATOS, E. 2009. A survey of modern authorship attribution methods. *J. Amer. Soc. Inf. Sci. Technol. 60*, 3.

STAMATATOS, E., FAKOTAKIS, N., AND KOKKINAKIS, G. 2001. Automatic text categorization in terms of genre and author. *Comput. Linguist. 26*, 4, 471–495.

STEIN, S. AND ARGAMON, S. 2006. A Mathematical explanation of Burrows's delta. In *Proceedings of the Digital Humanities Conference*.

THISTED, R. AND EFRON, B. 1987. Did Shakespeare write a newly-discovered poem? *Biomerika 74*, 3, 445–455.

TULDAVA, J. 2004. The development of statistical stylistics a survey. *J. Quant. Linguist. 11*, 1–2, 141–151.

WEISS, S. M., INDURKHYA, N. AND ZHANG, T. 2010. *Fundamentals of Predictive Text Mining*. Springer, London.

WITTEN, I. H. AND FRANCK, E. 2005. *Data Mining. Practical Machine Learning Tools and Techniques*. Elsevier, Amsterdam.

YANG, Y. AND LIU, J. X. 1999. A re-examination of text categorization methods. In *Proceedings of the ACM SIGIR Conference and Development in Information Retrieval*. 42–49.

YANG, Y. AND PEDERSEN, J. O. 1997. A comparative study of feature selection in text categorization. In *Proceedings of the 14th Conference on Machine Learning (ICML'97)*. 412–420.

YANG, A. C.-C., PENG, C.-K., YIEN, H.-W. AND GOLDBERGER, A. L. 2003. Information categorization approach to literary authorship disputes. *Physica A, 329*, 473–483.

ZHAI, C. X. AND LAFFERTY, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst. 22*, 2, 179–214.

ZHAO, Y. 2007. Effective authorship attribution in large document collections. Ph.D. thesis, RMIT Melbourne.

ZHAO, Y. AND ZOBEL, J. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of the 2nd AIRS Asian Information Retrieval Symposium*. 174–189.

ZHAO, Y. AND ZOBEL, J. 2007a. Searching with style: Authorship attribution in classic literature. In *Proceedings of the 30th Australasian Computer Science Conference (ACSC'07)*. 59–68.

ZHAO, Y. AND ZOBEL, J. 2007b. Entropy-Based authorship search in large document collection. In *Proceedings of the European Conference on IR Research (ECIR2007)*. Lecture Notes in Computer Science, vol. 4425, Springer, 381–392.

ZHENG, R., LI, J., CHEN, H., AND HUANG, Z. 2006. A framework for authorship identification of online messages: Writing-Style features and classification techniques. *J. Amer. Soc. Inf. Sci. Technol. 57*, 3, 378–393.

Received July 2011; revised November 2011; accepted March 2012