Published in Information Processing and Management 49, 1, 341-354, 2013
which should be used for any reference to this work

Authorship attribution based on a probabilistic topic model

Jacques Savoy

Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2000 Neuchâtel, Switzerland

A B S T R A C T

This paper describes, evaluates and compares the use of *Latent Dirichlet allocation* (LDA) as an approach to authorship attribution. Based on this generative probabilistic topic model, we can model each document as a mixture of topic distributions with each topic specifying a distribution over words. Based on author profiles (aggregation of all texts written by the same writer) we suggest computing the distance with a disputed text to determine its possible writer. This distance is based on the difference between the two topic distributions. To evaluate different attribution schemes, we carried out an experiment based on 5408 newspaper articles (*Glasgow Herald*) written by 20 distinct authors. To complement this experiment, we used 4326 articles extracted from the Italian newspaper *La Stampa* and written by 20 journalists. This research demonstrates that the LDA-based classification scheme tends to outperform the Delta rule, and the χ^2 distance, two classical approaches in authorship attribution based on a restricted number of terms. Compared to the Kullback–Leibler divergence, the LDA-based scheme can provide better effectiveness when considering a larger number of terms.

Keywords:
Authorship attribution
Text categorization
Machine learning
Lexical statistics

1. Introduction

In order to manage the huge amount of freely available textual information, various text categorization tasks have been proposed. In this study, we address the authorship attribution (AA) problem (Love, 2002; Juola, 2006) whereby the author of a given text must be determined based on text samples written by known authors. Knowing that the real author is one of the candidates, this specific challenge is defined as the *closed-class authorship attribution* problem. In such applications, the query text might correspond to various items such as a romance, a part of a play, an anonymous letter, a web page, or a sequence of paragraphs. Other questions are related to such issues as the mining of demographic or psychological information on an author (*profiling*) (Argamon et al., 2006) or simply determining whether or not a given author did in fact write a given Internet message (chat, e-mail, Wikipedia article) or document (*verification*) (Koppel, Schler, & Argamon, 2009).

As with all text categorization problems, the first step is to represent the texts by means of numerical vectors comprising selected features that help in discriminating among the various authors or categories. In the current context, we must identify pertinent terms depicting differences between the authors' writing styles. In the second stage, we weight the chosen features according to their discriminative power as well as their importance in the text representation. Finally, through computing a distance or applying classification rules, the system assigns the most appropriate author to a given input text (*single-label categorization* problem).

In the classical authorship attribution studies (Juola, 2006), we usually focus on frequent words or on a small number of very frequent terms to represent each text item. We then define a distance measure and determine the probable author of the query text as the one that depicts the smallest distance. As an alternative way, the machine learning paradigm (Sebastiani, 2002) will focus more of the selection process on identifying the most pertinent features according to their

E-mail address: Jacques.Savoy@unine.ch

distribution into the different categories or authors. As classification model, this paradigm may use a larger range of possible strategies (Witten & Franck, 2005). In this paper we propose following a third way. Using a probabilistic generative approach based on a probabilistic topic model (Blei, Ng, & Jordan, 2003) we show how we can determine the possible author of a disputed text.

The rest of this paper is organized as follows: Section 2 exposes the state of the art. Section 3 describes the corpora used in our experiments while Section 4 presents an overview of four authorship attribution models used as baselines. Section 5 presents the idea of latent Dirichlet allocation (LDA) and its application as an authorship attribution method. An evaluation of these classifiers is presented and analyzed in Section 6.

2. Related work

Authorship attribution owns a long-standing history (Mosteller & Wallace, 1964, Juola, 2006, Zheng, Li, Chen, & Huang, 2006, Stamatatos, 2009). As a first solution, past studies have proposed using a unitary invariant measure that must reflect the style of a given author but should vary from one writer to another. The average word length, mean sentence length, as well as Yule's *K* measure and statistics based on type-token ratios (e.g., Herdan's *C*, Guiraud's *R* or Honoré's *H*) (Baayen, 2008) have been suggested as well as the proportion of word types occurring once or twice. However, none of these measures has proved satisfactory (Hoover, 2003).

As a second approach, multivariate analysis (principal component analysis (Craig & Kinney, 2009), cluster analysis, or discriminant analysis (Jockers & Witten, 2010)) has been applied to capture each author's discriminative stylistic features. In this case, we represent documents as points within a given lexical space. In order to determine who might be the author of a new text excerpt we simply search the closest document assuming that the author of this nearest document would probably be the author of the disputed text.

Following the idea of measuring an intertextual distance, some recent studies suggest using more topic-independent features that may reflect an author's style more closely. In this vein, we can limit text representation to function words (e.g., determiners, prepositions, conjunctions, pronouns, and certain auxiliary verbal forms). Since the precise definition of function words is questionable, a wide variety of lists have been proposed. Burrows (2002), for example, lists the top *m* most frequent word types (with $m = 40–150$), while the list compiled by Zhao and Zobel (2005) contains 363 English words. 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. As another example, Hoover (2007) suggests considering up to the top 4000 frequently occurring words, including in this case both function and lexical words (nouns, adjectives, verbs, and adverbs). On the other hand, more sophisticated intertextual distances have been suggested (Labbé, 2007), where the distance between two documents depends on both their shared vocabulary and occurrence frequencies.

As another source of features, we could take into account part-of-speech (POS) tags by measuring their distribution, frequency and patterns or their various combinations. Finally, some studies, usually related to the web, suggest exploiting structural and layout features (the number of lines per sentence or per paragraph, paragraph indentation, presence of greetings, etc.). Additional features that could be considered are particular orthographic conventions (e.g., British vs. US spelling) or the occurrence of certain spelling errors.

As a second main paradigm, we can apply machine learning techniques to determine the probable author of a disputed text. In this vein, we can see each author as one possible category using the set of previously described features. We then need to define a classification model that can distinguish among possible authors. Zheng et al. (2006) suggests employing decision trees, back-propagation neural networks and support vector machines (SVMs). They found that by solely using lexical features, the performance levels obtained are similar to those of POS and lexical feature combinations. This finding is confirmed by another recent study (Zhao & Zobel, 2007). Zheng et al. (2006) also found that SVM and neural networks tend to have significantly better performance levels than those achieved by decision trees. Nanavati, Taylor, Aiello, and Warfield (2011) have obtained good overall performance with a naïve Bayes classifier for deanonymizing referees' reports extracted from two scientific conferences. Zhao and Zobel (2005) on the other hand, found that the Nearest Neighbor (NN or *k*-NN) approach tends to produce better effectiveness levels than both the naïve Bayes and decision-tree approaches.

Instead of applying a machine learning classification method, Burrows (2002) designed a more specific Delta classifier based on the differences of standardized word occurrence frequencies. This method assumes that authors' styles are best reflected by identifying the use of function words (or very frequent words) 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 when using the SVM method.

3. Evaluation corpora

The number of publicly available test corpora related to the authorship attribution domain is quite limited. Thus, making sufficiently precise comparisons between reported performances and general trends regarding the relative merits of various classification approaches is problematic. The relatively small size of the corpora used is a second concern. In various experiments, the number of disputed texts and possible authors are rather limited (e.g., the *Federalist Papers* are composed of 85 texts from which 12 are disputed articles mainly written by two possible authors (Mosteller & Wallace, 1964, Jockers & Witten, 2010)).

In order to obtain a replicable test-collection containing authors sharing a common culture and having similar language registers, we selected a subset of the stable and publicly available CLEF 2003 test suite (Peters, Braschler, Gonzalo, & Kluck, 2004). More precisely, we extracted articles in the *Glasgow Herald* (GH) and published in 1995. To form a suitable test-collection, we then chose 20 authors (see Table 1), either as well-known columnists (names in italics) or having published numerous papers. This selection process yields a set of 5408 articles.

This corpus covers different subjects with a clear overlap among authors. For example, five authors are listed under the main descriptor *Business*, five others are listed under *Sports*, while only four are listed under *Social* and three under both the *Politics* and *Arts & Film* headings.

The "Number" column in Table 1 lists the number of articles written by each author, showing a minimum of 30 (John Fowler), and a maximum of 433 (Andrew Wilson). This distribution is rather skewed, with a group of eight authors having published more than 350 articles and another group of four journalists writing less than 100 articles (mean: 270, median: 332, standard deviation: 139). An analysis of article length shows that the mean number of word tokens is 725 (minimum: 44, maximum: 4414, median: 668, standard deviation: 393).

As a second evaluation corpus, we selected newspaper articles published in *La Stampa* during the year 1994. This corpus written in Italian language is also part of the CLEF 2003 test suite (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.

Table 1
Distribution of 5408 *Glasgow Herald* articles by author name, subject, and number of articles per author.

	Name	Subjects	Number
1	*Davidson Julie*	Arts & Film	57
2	Douglas Derek	Sports	410
3	*Fowler John*	Arts & Film	30
4	Gallacher Ken	Sports	408
5	Gillon Doug	Sports	368
6	*Johnstone Anne*	Social, Politics	72
7	McConnell Ian	Business	374
8	*McLean Jack*	Social	118
9	Paul Ian	Sports	418
10	Reeves Nicola	Business	370
11	Russell William	Arts & Film	291
12	*Shields Tom*	Politics	173
13	Sims Christopher	Business	390
14	*Smith Ken*	Social	212
15	Smith Graeme	Social, Politics	329
16	Traynor James	Sports	339
17	Trotter Stuart	Politics	336
18	Wilson Andrew	Business	433
19	*Wishart Ruth*	Politics	72
20	*Young Alf*	Business, Economics	208

Table 2
Distribution of 4326 articles from *La Stampa* by author name, subject,and number of articles per author.

	Name	Subjects	Number
1	Ansaldo Marco	Sports	287
2	Battista Pierluigi	Politics	231
3	Beccantini Roberto	Sports	364
4	Beccaria Gabriele	Social	71
5	Benedetto Enrico	Politics	252
6	Del Buono Oreste	Sports	434
7	Comazzi Alessandra	Social	223
8	Conti Angelo	Social	198
9	Gavano Fabio	Politics	347
10	Gramellini Massimo	Politics	118
11	Meli Maria Teresa	Politics	215
12	Miretti Stefania	Social	63
13	Nirenstein Fiama	Politics	52
14	Novazio Emanuele	Politics	249
15	Ormezzano Gian Paolo	Sports	232
16	Pantarelli Franco	Politics	202
17	Passarini Paolo	Politics	303
18	Sacchi Valeria	Business	203
19	Spinelli Barbara	Politics	57
20	Torabuoni Lietta	Social	225

From the set of all possible articles (58,051 documents), we must first select papers with a known author (37,682 articles). From this, we 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 2), 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 4326 articles.

The "Number" column in Table 2 lists the number of articles written by each author, showing a minimum of 52 (Fiama Nirenstein), and a maximum of 434 (Oreste Del Buono). An analysis of article length shows that the mean number of word tokens is 777 (minimum: 60; maximum: 2935; median: 721; standard deviation: 333). In the selected newspapers articles, we automatically removed 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.).

The accuracy rate (percentage of correct answers) will be used as an evaluation measure. As an aggregation function, we consider that all classification decisions share the same importance (micro-average). In authorship attribution domain, this evaluation measure is the most frequently used. As a second approach, we may use the macro-average accuracy rate. In this case, we first compute the accuracy rate obtained for each author then average these 20 values. Both effectiveness measures are strongly correlated and thus we have just reported the micro-average in our experiments.

4. Categorization models

4.1. Preprocessing

Before classifying the newspaper articles according to their possible authors, we first replaced certain punctuation marks with their corresponding ASCII symbols and removed a few diacritics (e.g., *chiché*) when appearing in *Glasgow Herald* articles. For the English language, the absence of the diacritics is usually not seen as a real spelling error. For the Italian corpus, we keep the diacritics (e.g., *città* (city), *società* (society)) because the words are always written with their corresponding diacritics. To standardize spelling forms we expanded contracted forms (e.g., *don't* into *do not*) and replaced uppercase letters with their lowercase equivalents.

After this step, the resulting vocabulary contains 56,447 word types, with 19,221 *hapax legomenon* (words occurring once), and 7530 *dis legomenon* (words occurring exactly twice). Usually the expression *word* is ambiguous. To be precise, in the sentence "the cat sleeps on the table" we count six word tokens (or simply tokens) but only five word types (the determiner *the* appears twice). When considering only those word types having an occurrence frequency of 10 or more, we count 14,890 types, or 9628 types having frequencies equal to or greater than 20. The most frequent token is *the* (219,632 occurrences), followed by the comma (183,338), the full stop (146,590), and ranking fourth is the token *to* (95,350), followed by *of* (92,755), and *a* (78,867).

From the Italian newspaper *La Stampa*, we find 102,887 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 full stop (126,891), and the word type *di* (of) (100,433), ranking fourth is the token *e* (and) (73,818), followed by *il* (the) (63,931), and *che* (that) (59,600).

Finally, to define each author profile, we concatenated all texts written by the same writer. Using this subset, we applied the feature selection procedure, and represented each author profile or disputed text by a set of weighted features. Throughout all the experiments, the text whose author we need to determine was never included in the corresponding author profile.

4.2. Delta rule

To determine the most probable author of a disputed text, Burrows (2002) suggests that the most important aspect to take into account is the frequency of very frequent terms (the top $m = 40$–150 most frequent word types). In this set of very frequent word types we can find many function words (e.g., determiners (e.g., *the*, *a*), prepositions (*in*, *of*), conjunctions (*or*), pronouns (*we*, *she*), and certain auxiliary verbal forms (*has*, *was*, *can*)). Following Burrows' specifications, we ignore punctuation marks and numbers. The main underlying idea is to consider only very frequent words used unconsciously by an author and able to reveal his/her own "fingerprint".

Burrows proposed to not consider directly the absolute frequencies, but rather their standardized scores. This Z score value is computed for each selected term t_i (word type) in a text sample (corpus) by calculating its relative term 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 as depicted in the following equation.

$$Zscore(t_{ij}) = \frac{tfr_{ij} - mean_i}{sd_i} \quad (1)$$

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

$$\Delta(Q, A_j) = 1/m \cdot \sum_{i=1}^{m} |Zscore(t_{iq}) - Zscore(t_{ij})| \quad (2)$$

When computing this distance we attribute the same importance to each term t_i, regardless of their absolute occurrence frequencies. Large differences may occur when, for a given term, both Z scores are large and have opposite signs. In such cases, one author tends to use the underlying term more frequently than the mean while the other employs it very infrequently. On the other hand, when the Z scores are very similar for all terms, the distances between the two texts would be small, indicating that the same author had probably written both texts.

4.3. Chi-square distance

As a second baseline, we selected one of the best text representations found in an empirical study done by Grieve (2007) which is based on word tokens together with eight punctuation marks (. , : ; - ? (‘). 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 in at least k articles written by every possible author. For example, when $k = 2$, the selected terms must appear in at least two articles written by every journalist. This selection strategy implies the principle that all authors must use terms used to discriminate between them. In this case, we cannot base an attribution decision on word types used by a single author or more generally a subset of authors. At the limit, when only a single author uses a given term, we can assign this text to this author as soon as the target term occurs. However, another author may easily play a masquerade by using the same term.

Increasing the value of k will reduce the number of word types to be taken into account while a small value for k implies that we consider more words. As for the effective values of the parameter k, Grieve (2007) observed that the best performance results were achieved when $k = 2$, $k = 5$ or $k = 10$.

To compare a given query text Q with an author profile A_j, Grieve (2007) uses the χ^2 statistic defined by Eq. (3), in which $fq(t_i)$ represents the relative frequency of the ith feature in the query text Q, $fa_j(t_i)$ the corresponding frequency in the jth author profile A_j, and m the number of selected terms t_i in a k-limit.

$$\chi(Q, A_j) = \sum_{i=1}^{m} (fq(t_i) - fa_j(t_i))^2 / fa_j(t_i) \quad (3)$$

When comparing a disputed text with different author profiles, we simply select the lowest χ^2 value to determine the most probable author.

4.4. Kullback–Leibler divergence (KLD)

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

In our experiments with the *Glasgow Herald* newspaper, we found 19 words that do not appear. For nine of them, their absence can be attributed to the fact that during the preprocessing we expanded the contracted forms (e.g., *aren*, *isn*, *wasn*, and *weren*), and the other absences are caused by rare forms (e.g., *hereupon*, *inasmuch*, and *whereafter*). Thus our final list will contain 344 words ($363 - 19$).

To obtain a similar list for the *La Stampa* corpus, we select an Italian stopword list provided by a search system achieving high retrieval performance in CLEF evaluation campaigns for that language (Savoy, 2001). This Italian list contains 399 words.

After defining the feature set, the probability of occurrence of each item associated with a given author or a disputed text must be estimated. Based on these estimations, we can measure the degree of disagreement between the two probabilistic distributions. To do so Zhao and Zobel (2007) suggest using the Kullback–Leibler divergence (KLD) formula, also known as *relative entropy* (Manning & Schütze, 2000). 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.

$$KLD(Q||A_j) = \sum_{i=1}^{m} \text{Prob}_q[t_i] \cdot log_2 \left[\frac{\text{Prob}_q[t_i]}{\text{Prob}_j[t_i]} \right] \quad (4)$$

where $\text{Prob}_q[t_i]$ and $\text{Prob}_j[t_i]$ indicate the occurrence probability of the term t_i in the query text Q or in the jth author profile A_j, respectively. In the underlying computation, we state that $0.\log_2[0/p] = 0$, and $p.\log_2[p/0] = \infty$. With this definition, and when the two distributions are identical, the resulting value is zero, while in all other cases the returned value is positive. As a decision rule, we assign the query text to the author whose profile shows the smallest KLD value.

To estimate the underlying probabilities, we may consider the term occurrence frequency (denoted tf_i) and the size of the corresponding text (n) (e.g., Prob[t_i] = tf_i/n). This solution tends, however, to overestimate the occurrence probability of terms appearing in the sample at the expense of the missing terms. To resolve this difficulty, we suggest smoothing the probability estimates using the Lidstone's technique (Manning & Schütze, 2000) based on the following estimation: Prob[t_i] = ($tf_i + \lambda$) / ($n + \lambda|V|$), with $|V|$ indicating the vocabulary size. Based on some experiments, we suggest fixing this λ value to 0.01 which yields a slightly improved performance over other choices.

4.5. Naïve Bayes

With the three previously described authorship attribution methods we follow the classical paradigm. In this case, these approaches propose a selection procedure focusing on frequent words. Then, based on a distance measure between the query text and author profiles, they suggest determining the probable author as the one that depicts the smallest distance.

In the machine learning domain, other approaches have been suggested for text categorization (Sebastiani, 2002). Following this paradigm, we first need to define a selection criterion to reduce the number of possible features (term space reduction). This step is useful to decrease the computational cost and to reduce the over-fitting of the learning scheme to the training data. In a second step, we use the training data to let the classifier learn the characteristics discriminating the most across the different categories or authors in our context.

As a learning scheme, we selected the naïve Bayes model (Mitchell, 1997) to determine the probable writer between the set of twenty possible journalists (or hypotheses), denoted by A_i for $i = 1, 2, \ldots, r$. To define the most probable author of a query text Q, the naïve Bayes model selects the one maximizing Eq. (5), in which t_j represents the jth term included in the query text Q, and n_q indicates the size of the query text.

$$\text{Arg max}_{A_i} \text{Prob}[A_i|Q] = \text{Prob}[A_i] \cdot \prod_{j=1}^{n_q} \text{Prob}[t_j|A_i] \tag{5}$$

To estimate the prior probabilities (Prob[A_i]), we simply take into account the proportion of articles written by each author (see Table 1 for the GH corpus, and Table 2 for *La Stampa*). To determine the term probabilities Prob[$t_j \mid A_i$], we regroup all texts belonging to the same author to define the author profile. For each term t_j, we then compute the ratio between its occurrence frequency in the author profile A_i (tf_{ji}) and the size of this sample (n_i) as Prob[$t_j \mid A_i$] = tf_{ij}/n_i.

As mentioned previously this definition tends to over-estimate the probabilities of terms occurring in the text with respect to terms that still never occur. As with the previous method, we will apply Lidstone's law through smoothing each estimate as Prob[$t_j \mid A_i$] =($tf_{ji} + \lambda$) / ($n_i + \lambda|V|$), 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. As a first approach, we have selected the odds ratio (OR), a selection function found historically effective (Sebastiani, 2002). 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. (6). In this formulation, Prob[$t_j \mid A_i$] indicates the probability that, for a random document, the term t_j appears knowing that this text was written by author A_i. Similarly, Prob[$t_j \mid -A_i$] indicates the same probability except that the underlying document was not written by author A_i.

$$OR(t_j, A_i) = \frac{\text{Prob}[t_j|A_i] \cdot (1 - \text{Prob}[t_j| - A_i])}{(1 - \text{Prob}[t_j|A_i]) \cdot \text{Prob}[t_j| - A_i]} \tag{6}$$

When a given term t_j appears more frequently in the author profile A_i, then it can be used to discriminate the ith author from the rest. In this case, its probability Prob[$t_j \mid A_i$] will be relatively high. On the other hand, the probability Prob[$t_j \mid -A_i$] will be relatively small because the term t_j will occur relatively less often in texts not written by the ith author. As shown in Eq. (6), this phenomenon will assign a relatively high value for the numerator compared to the denominator. The resulting OR(t_j, A_i) 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.

Using Eq. (6) we obtain a value for each pair (term, author), and to rank the different terms, we need to define one selective value per feature reflecting its discriminative capability over all categories (or authors in the current context). To achieve this, we need to aggregate the r values, one for each author. Sebastiani (2002) indicates that the sum operator (see Eq. (7)) tends to produce the best results with the OR used as term selection function.

$$OR_{sum}(t_j) = \sum_{i=1}^{r} OR(t_j, A_i) \tag{7}$$

Based on this criterion, we can then select the m most discriminative terms. As an alternative to this selection procedure, we will also choose terms according to their document frequency (df) values, sorted by decreasing values. In this case, terms appearing in a larger number of articles will be selected first. This selection criterion owns the advantage of being simple to implement and has also been proved to be effective in other text categorization problems, as mentioned by Yang and Pedersen (1997). Finally, in this scheme as well as for the LDA (see next section), we consider as possible terms the following seven

punctuation symbols {. , ; : ! ? '}. Such symbols could provide useful information to discriminate between authors. For example, if a given author tends to write longer sentences, the number of full stop symbols will be less than for another author who usually writes shorter sentences (thus producing more full stops).

5. Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) (Blei, 2011; Blei, 2012; Blei & Lafferty, 2009; Blei et al., 2003) is an extended version of the probabilistic topic model suggested by Hofmann (1999). In this framework, documents are viewed as composed of a mixture of topics. Of course, a given document may cover only a single topic, but this is more the exception than the norm. For example, the document D_1 may cover intensively Topic 1, present some aspects related to Topic 2, and marginally to Topic 3. An equal mixture of Topic 1 and 3 may generate the second article D_2. Thus each document D_j, for $j = 1,\ldots,n$, is generated based on a distribution over the k topics, where k defines the maximum number of topics. This value is fixed and defined *a priori.* In the LDA model, each document in the corpus may be generated according to all k topics with different intensities.

Each topic does not correspond to a symbolic subject heading such as ''Politics'' or ''Sports'' but it is defined as a specific word distribution. In this model, the word position in the document or in the sentence is not taken into account (bag-of-word assumption), as well as the document order in the corpus. When inspecting the vocabulary, we can find a very specific word strongly related to a single topic that may appear only in a particular distribution. Usually, however, a word tends to cover different semantic fields such as the word *Java* (*Java* as a drink, an island, a programming language, a dance, etc.). In such cases, the word may appear in different topics to model its polysemy. Finally, when a word is a function word or simply when it occurs in almost all texts (e.g., *the*, *of*, *in*, *was*, *would*, etc.), each topic will include it with usually a different occurrence probability.

Some examples can illustrate this view. Based on the *Glasgow Herald* corpus, some topic examples with their most probable words are given in Table 3 (after removing the 400 most frequent words). In this table we can see that Topic #3 is related to rugby competitions in Edinburg (Murrayfield stadium) while Topic #4 is related to macro-economic measures in the US. Topic #9 is about budget reduction(s) under the Chancellor of the Exchequer K. Clarke.

For the newspaper *La Stampa*, the same information is depicted in Table 4. For the Italian corpus, we can see that soccer is an important topic appearing in Topic #1 with the Milan AC and its stars (R. Gullit, D. Savicevic, M. Desailly, G. Lentini, F. Baresi) and its trainer (F. Capello). The national soccer team appears with Topic #12 (Arrigo Sacchi, Roberto Baggio (*il Divin Codino*), G. Signori). Politics is present under Topic #2 with the party PDS (*Partito Democratico della Sinistra*) and its well-known figures as M. D'Alema, the successor of A. Occhetto, or W. Veltroni. With Topic #6, we see the financial world with terms such as *borsa* (stock exchange), *lira* (lire), *dollaro*, *mercato* (market), *mercati* (markets), *tassi* (rates) as well as important figures of this world (e.g., Marco Biagi).

In this generative view and for a given document, we first randomly select a topic according to the topic distribution for the underlying text. Then we randomly select a word according to the word distribution associated with the selected topic. Finally we repeat this process for all words in the target document. Denoting by $\text{Prob}[z^i = j]$ the fact that the distribution of the jth topic was used to generate the word at the ith position, and by $\text{Prob}[t^i \mid z^i = j]$ the probability that the term t^i occurring at the ith position was generated according to the distribution of topic j, we can compute the probability of the occurrence of the term t^i in the document as:

$$\text{Prob}[t^i] = \sum_{j=1}^{k} \text{Prob}[t^i | z^i = j] \cdot \text{Prob}[z^i = j] \tag{8}$$

For a given document containing n_d tokens, we can then multiply the $\text{Prob}[t^i]$, for $i = 1, 2, \ldots, n_d$ to define the probability of observing this document with its corresponding sequence of n_d words (assuming the independence assumption between words, or unigram assumption).

In practice, however, we are usually more interested to solve the dual view (posterior probabilistic inference). Given a fixed number of k topics and the observed words (with their occurrence frequencies) in a set of n documents, we need to determine the most likely topic model corresponding to these data. In this formulation, the hidden structure is the topics and the word distributions. Thus we need to estimate the probability distributions over words associated with each topic (see examples given in Tables 3 and 4), and the distribution of topics over documents. Of course, we are interested in the hidden structure that best explains the observed words and documents. Based on the Gibbs sampling procedure, these estimates can be computed under an iterative procedure. As an alternative way to estimate the underlying distribution, some authors have suggested using variational EM methods (Blei, 2011).

A few implementations are freely available (e.g., we used a C version written by Blei, and other possible packages such as lda are available with the R system (Crawley, 2007)). In our case, given as input the 20 documents (author profiles) represented by their term frequencies, the system will return the estimated distribution of words over the topics together with the distribution of topics over documents. This simple LDA approach was chosen due to its simplicity and because previous studies have shown that such simple topic models may provide better effectiveness than more complex ones (Carman, Crestani, Harvey, & Braillie, 2010).

Finally, using these distribution estimates, the system can infer the topic distributions of a new and unseen document (the disputed text in our case). To define the possible author of this query text, we suggest computing the distance between

Table 3
Example of four topics generated from the *Glasgow Herald* (fixing $k = 100$, after removing function words).

Topic #3	Topic #4	Topic #6	Topic #9
scotland	more	pay	tax
against	economic	directors	said
rugby	US	year	chancellor
scottish	currency	shareholders	budget
game	economy	salary	government
murrayfield	inflation	company	clarke
played	rates	annual	cut

Table 4
Example of four topics generated from the *La Stampa* (fixing $k = 100$, after removing function words).

Topic #1	Topic #2	Topic #6	Topic #12
milan	pds	borsa	baggio
capello	alema	lira	sacchi
gullit	occhetto	dollaro	arrigo
savicevic	segretario	mercati	italia
gol	leader	mercato	signori
desailly	sinistra	tassi	roberto
lentini	veltroni	marco	mondiale
baresi	quercia	affari	codino

the topic distribution of the query text and those corresponding to the 20 author profiles. To achieve this and following the suggestion of Steyvers and Griffiths (2007), we propose using the symmetric Kullback–Leibler distance (Abdi, 2007; Lin, 1991) based on two discrete probability distributions.

$$sKLD(Q||A_j) = 1/2\left(\sum_{i=1}^{m} p_q(s_i)\cdot log_2\left[\frac{p_q(s_i)}{p_j(s_i)}\right] + \sum_{i=1}^{m} p_j(s_i)\cdot log_2\left[\frac{p_j(s_i)}{p_q(s_i)}\right]\right) \quad (9)$$

in which Q indicates the distribution of topics over the query text, A_j the same distribution for the jth author profile, and $p_q(s_i)$, $p_j(s_i)$ are the probabilities that the texts Q and A_j, respectively, were generated according to the ith topic or subject. As a decision rule, we define the author of the query text as the author profile showing the smallest sKLD value.

Various applications in different contexts (unsupervised learning) have been suggested, and good overviews are presented in (Blei et al., 2003, Steyvers & Griffiths, 2007, Blei, 2012) As a first example, we can mention the study done by Sugimoto, Li, Russell, Finlay, and Ding (2011). In this study, we can see the topics evolution between 1930 and 2009 in doctoral dissertations related to library science. Xing and Croft (2006) suggest using the topic modeling techniques as a search model depicting a reasonable overall performance. Carman et al. (2010) propose using the LDA approach to build personalized search models and evaluate them using query logs and their click-through data. More closely related to our context, Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) and Griffiths, Smyth, Rosen-Zvi, and Steyvers (2004) proposed to include the authors' information by the mean of a distribution over the topics. Such a view can be useful to see the topic variation for a given author, or relationship between authors based on shared topics, but the authorship attribution was only mentioned as possible future directions.

In a more closely related study, Arun, Saradha, Suresh, Narsimha Murty, and Veni Madhavan (2009) suggest using the LDA approach to represent documents as a mixture of topics. In a first experiment, these authors used only 555 stopwords while all available word types (93,980) were employed in a second. The main idea is to reduce the document representation to a smaller set of topics (from 5 to 50). Such a compact representation is then used in conjunction with a Support Vector Machine (SVM) classifier (Cristianini & Shawe-Taylor, 2000; Joachims, 2001) to determine the probable author of a literary works (excerpts of 5000 words). Using only words appearing a stopword list or all possible words, this study shows that we can achieve similar performance levels.

In a recent study, Seroussi, Zukerman, and Bonhert (2011) suggest also to employ the LDA approach to represent the document as a mixture of topics. In order to reduce the number of terms, they have applied different filters (e.g., selecting only word types having an occurrence frequency larger than a given threshold). They found that using all available terms tends to produce the highest performance levels. After representing each document as a mixture of topics, they first suggest applying a SVM classifier (solution denoted LDA + SVM). As a second AA scheme, they propose to compute a distance between a query text and two possible author profiles using the Hellinger metric. Their experiments demonstrate that LDA + Hellinger tends to produce significantly better performance (micro-average) than the LDA + SVM classifier. When considering more authors (from 1000 to 19,320, using a blog collection), the achieved accuracy rate was rather low (between 1% and 14%), too low to be practically useful.

6. Evaluation

To evaluate and compare the different authorship attribution methods, we used the 5408 articles of the *Glasgow Herald* written by 20 well-known columnists. As a second corpus, we used the Italian newspaper *La Stampa* (4326 articles written by 20 distinct journalists). As the evaluation measure we will use the accuracy rate based on the assumption that each decision owns the same importance (micro-average). As a second possible performance measurement, we can use the macro-average principle. In this case, the mean performance obtained for each author corresponds to one vote and the overall accuracy rate is the mean over all authors. We have decided to limit the performance measure to the micro-average because this measure is commonly used in authorship attribution studies on the one hand, and, on the other, both measures are strongly correlated in our context.

To statistically determine whether or not one strategy is better than another, we applied the sign test (Conover, 1980), (Yang & Liu, 1999). In this test, the null hypothesis H_0 states that both categorization schemes produce similar performance levels. Such a null hypothesis would be accepted if two authorship attribution schemes returned statistically similar accuracy rates, otherwise, it would be rejected (two-tailed test, significance level $\alpha = 1\%$). Statistically significant performance differences will be indicated by a double cross (‡) in the following tables.

7. Classical authorship attribution

Using the Delta method (Burrows, 2002), we first need to select the top m most frequent word types. Based on previous experiments, the highest accuracy rate was achieved with the top 400 most frequent types (with small differences as more or less terms are taken into account, e.g., 200 or 600). 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 on a per-author basis. In this case, the best performance was achieved when considering all words and punctuation symbols appearing in at least two texts for each author (2-limit). As a third baseline, we used the KLD scheme proposed by Zhao and Zobel (2007) based on a predefined set of 363 English words. For the *La Stampa* corpus, we have selected an Italian stopword list containing 399 terms.

In order to compare the proposed LDA-based scheme with the other three classical approaches, we will use the same set of terms. In Table 5, we have reported the accuracy rate of the Delta method, χ^2 approach, and KLD method using both the *Glasgow Herald* (GH) and *La Stampa* (IT) corpora. Using the same terms, we have also reported the accuracy rate achieved by the LDA-based approach, fixing the number of possible topics k between 20 and 80. In the current context, we may assume that each topic corresponds to the style of a given author and thus fixing $k = 20$ represents a reasonable choice. Of course, having a value for k smaller than 20 does not really make sense. The value for k could however be larger than 20 reflecting the fact that two or more distributions of words (topics) are needed to well describe the various styles of a given columnist.

In Table 5 we have reported the performances obtained using the three classical authorship approaches together with the LDA model. As statistical testing, we used the sign test (significance level $\alpha = 1\%$, two-sided), using the performance achieved by the best performance (depicted in bold) as a baseline. After the accuracy rate value, we added a double cross (‡) to denote a statistically significant performance difference.

When applying the Delta rule with the 400 most frequent terms, the accuracy rate (micro-average) is 63.7% with the English corpus, and 76.07% with the Italian collection. When computing this performance using the macro-average principle (mean over all author performances), we obtain an accuracy rate of 66.14% with the *Glasgow Herald*, and 75.08% with *La Stampa*. The two performance measurements are clearly correlated. Thus we will limit our analysis to the micro-average principle.

Table 5
Evaluation of various classical authorship attribution approaches using the *Glasgow Herald* (GH) or *La Stampa* (IT) newspaper.

Method	Parameter	GH Accuracy (%)	IT Accuracy (%)
Delta	400 terms	63.70 ‡	**76.07**
LDA	400 terms, $k = 20$	67.80 ‡	72.83 ‡
LDA	400 terms, $k = 50$	73.32 ‡	75.64
LDA	400 terms, $k = 60$	**74.11**	73.00 ‡
LDA	400 terms, $k = 80$	73.06 ‡	67.67 ‡
χ^2	653 (GH)/720 (IT) terms	65.26 ‡	68.28 ‡
LDA	653 (GH)/720 (IT) terms, $k = 20$	66.44 ‡	73.04
LDA	653 (GH)/720 (IT) terms, $k = 40$	72.24	**73.25**
LDA	653 (GH)/720 (IT) terms, $k = 50$	**72.51**	72.37 ‡
LDA	653 (GH)/720 (IT) terms, $k = 60$	71.54 ‡	70.73 ‡
KLD	363 (GH)/399 (IT) terms	**70.80**	**84.84**
LDA	363 (GH)/399 (IT) terms, $k = 20$	47.57 ‡	55.88 ‡
LDA	363 (GH)/399 (IT) terms, $k = 50$	40.93 ‡	48.65 ‡

The evaluations reported in this table indicate that the LDA-based authorship attribution method performs significantly better than the Delta model or the χ^2 measure for the English corpora. With the *La Stampa* newspaper, the LDA scheme performs better than the χ^2 metric but at a lower performance level than the Delta model. In this latter case, the performance differences between the two schemes are not always statistically significant.

The KLD method tends, however, to display a better effectiveness than the corresponding LDA-based scheme. This feature set defined *a priori* seems to be not well adapted for the LDA approach. For the Italian corpus for example, this term set containing 399 entries is able to produce an accuracy of 55.88% ($k = 20$). Using a similar amount of terms (400 most frequent terms) as proposed by the Delta method, we obtain a higher effectiveness of 72.83% ($k = 20$). When more terms are selected (as shown in the top and the middle of Table 5), the LDA-based scheme may produce better performance levels than those achieved with the small set of terms selected under the KLD model.

When considering the possible values of the parameter k (number of topics), we cannot see a clear and precise prescription for this parameter. Using the top 400 most frequent terms or terms belonging to a 2-limit, the effectiveness achieved with $k = 40$–80 reveals some small performance differences that are however usually statistically significant. With the feature set defined with the KLD model, the accuracy rate variations are more important between different k values as shown in the bottom part of Table 5.

7.1. Naïve Bayes

In order to have another point of view, we decided to compare the LDA-based scheme with a machine learning approach. As a typical text categorization model used in this domain, we have selected the naïve Bayes technique. Unlike the three classical attribution schemes described previously, the selection of the feature set is not usually specified within a machine learning model. Therefore, we have conducted a set of experiments using the odds ratio (OR sum) criterion, and a second set using the document frequency (*df*).

As shown in Table 6, we have reported the evaluation of the naïve Bayes model together with the suggested LDA-based model using either the *Glasgow Herald* (GH) or *La Stampa* (IT) newspaper. In these evaluations, we have used the same terms for both the naïve Bayes and LDA experiments. In a first evaluation, we have considered the naïve Bayes with terms selected according to the OR sum operator (Column 3 and 4). 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 (Column 5 and 6). 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 & Pedersen, 1997). After the accuracy rate value, we added a double cross (‡) to denote a statistically significant performance difference (using the sign test and compared to the best performance depicted in bold).

These evaluations tend to indicate that when using the odd ratio (OR) as the selection function, the achieved performance levels are lower than those obtained with the document frequency (*df*) measure using the same number of terms. When the number of word types is increased, the accuracy rates are improved, but the marginal enhancement tends to decrease. When analyzing the best number of topics, it is difficult to specify any rule. The optimal value for this parameter depends on the corpus, the number of terms and the underlying selection function. From data reported in Table 6, we can simply state that fixing $k = 20$ is usually not a good solution.

Table 6
Evaluation of LDA and naïve Bayes approaches for authorship attribution using the *Glasgow Herald* (GH) or *La Stampa* (IT) corpus.

Method	Parameter	GH OR sum (%)	IT OR sum (%)	GH *df* (%)	IT *df* (%)
NB	500 terms	**47.26**	**69.73**	69.88 ‡	78.16 ‡
LDA	500 terms, $k = 20$	30.35 ‡	61.87 ‡	67.53 ‡	79.14 ‡
LDA	500 terms, $k = 40$	35.34 ‡	64.91 ‡	72.93 ‡	**81.40**
LDA	500 terms, $k = 60$	34.56 ‡	66.23 ‡	**74.18**	80.61 ‡
LDA	500 terms, $k = 80$	35.46 ‡	65.63 ‡	73.73 ‡	78.53 ‡
NB	1000 terms	**57.78**	**76.40**	**79.40**	85.71 ‡
LDA	1000 terms, $k = 20$	33.19 ‡	65.42 ‡	76.67 ‡	86.14 ‡
LDA	1000 terms, $k = 40$	35.43 ‡	69.20 ‡	78.58 ‡	88.67
LDA	1000 terms, $k = 60$	35.42 ‡	68.12 ‡	79.23	88.58 ‡
LDA	1000 terms, $k = 80$	35.34 ‡	67.07 ‡	79.18	**89.09**
NB	2000 terms	65.35 ‡	**78.64**	**83.27**	90.08
LDA	2000 terms, $k = 20$	76.43 ‡	68.72 ‡	75.01 ‡	86.17 ‡
LDA	2000 terms, $k = 40$	80.75 ‡	71.61 ‡	81.27 ‡	89.75 ‡
LDA	2000 terms, $k = 60$	81.27	69.71 ‡	80.75 ‡	90.10
LDA	2000 terms, $k = 80$	**81.38**	69.02 ‡	81.27 ‡	**90.32**
NB	5000 terms	75.24 ‡	**83.01**	**84.28**	**91.70**
LDA	5000 terms, $k = 20$	76.67 ‡	71.57 ‡	77.31 ‡	86.51 ‡
LDA	5000 terms, $k = 40$	82.28	75.41 ‡	82.62 ‡	90.59 ‡
LDA	5000 terms, $k = 60$	82.25	75.47 ‡	82.84 ‡	90.69
LDA	5000 terms, $k = 80$	**82.36**	75.01 ‡	82.19 ‡	90.45 ‡

7.2. Additional experiments

In a final set of experiments, we want to study the best performance that can be achieved with the LDA-based authorship attribution model. As a term selection scheme, we first want to ignore words having a small occurrence frequency. Such words are clearly less important in defining the particular style of a given author. Moreover, they represent a large proportion of the vocabulary due to the Zipfian distribution of word occurrence. For example, with the *Glasgow Herald* we can find 56,447 distinct word types, with 19,221 *hapax legomenon* (words occurring once). When considering only those types having an occurrence frequency of 10 or more, we count only 14,890 types. During the term selection, we also add the constraint that each selected word appears in at least three distinct articles written by at least two authors. Thus we want to ignore words appearing only in a few papers or used only by a single author. Taking into account this second constraint, the English vocabulary is reduced to 2511 word types (or 4.4% of the initial vocabulary size).

From the newspaper *La Stampa*, we find 102,887 distinct word types, with 41,965 *hapax legomenon*. In this corpus, we can count 19,580 word types having an occurrence frequency of 10 or more. Adding the constraint that each word type must appear in at least three articles written by two distinct journalists, we obtain a vocabulary of 9825 terms corresponding to 9.5% of the initial size. The accuracy rates achieved using these feature sets are depicted in the top part of Table 7.

As a second feature selection, we have applied a less strict scheme, considering all words occurring in at least three different articles ($df \geqslant 3$). In this case, we reduced the English vocabulary size from 56,447 word types to 26,005 (or 46.1% of the initial size), but the resulting feature set is clearly larger than the previous one. For the Italian corpus, the vocabulary decreases from 102,887 word types to 36,928 representing around 35.9% of the initial size.

The evaluations reported in Table 7 complement the performance values shown in Table 6. Using the LDA-based authorship attribution method, we can achieve an accuracy rate around 82% with the *Glasgow Herald* corpus and around 91.5% with the *La Stampa* newspaper. For the English corpus, the overall best performance level of 82.84% was obtained when selecting the top 5000 words according to the *df* criteria (see Table 6). With the Italian collection, the best performance was obtained when considering 9825 terms (see Table 7) with an accuracy rate of 91.94%. These values are similar to those achieved by the naïve Bayes, maybe slightly lower for the *Glasgow Herald*, and slightly higher when considering the *La Stampa* corpus.

When considering the number of features selected, having more word types is not always a guarantee to obtain a higher accuracy rate. In Table 6 with the *df* selection scheme, the performance level with 2000 or 5000 terms were similar with the Italian collection. Increasing the number of word types as shown in Table 7 may slightly increase the overall accuracy rate with the *Glasgow Herald* corpus. For the *La Stampa* however, the increase from 9825 to 36,928 terms tends to hurt the overall quality of the assignment.

The computational cost is however not the same when faced with a larger number of terms. In our experiments using a C version of the LDA, we need around 1 min and 30 s to build the author profiles for the smallest feature sets (e.g., those presented in Table 5) and with $k = 20$ (English corpus). Based on our implementation of the Delta rule, χ^2 distance or KLD approach using the interpreter language Perl, the computational time needed to build the underlying data structures was around one minute. Using the larger set of 26,005 terms (see Table 7), this task requires an average of 21 min and 54 s for the LDA model ($k = 20$).

7.3. Further analysis of LDA results

Until now we have focused our evaluation on the accuracy rate and compared it to other possible authorship attribution schemes. In addition to this assignment evaluation, the LDA-based model allows us to analyze the relationships between the different topics (or word distributions) and the authors. To achieve this, the LDA scheme returns k topic distributions showing the occurrence probability of each word. Each topic may cover one or more subjects and two topics may reflect a similar domain but using different word frequencies. Some topic examples are given in Table 3 (*Glasgow Herald*) and in Table 4 (*La Stampa*). However, these examples were obtained after removing very frequent word types. As another example, we can fetch some topics (or word distributions) obtained with the *Glasgow Herald* newspaper based on 2511 words and with the $k = 20$ topics (accuracy rate depicted in Table 7). Table 8 depicts a subset of six of these 20 topics.

Table 7
Evaluation of LDA approach for authorship attribution.

Method	Parameter	GH Accuracy (%)	IT Accuracy (%)
LDA	2511/9825 terms, $k = 20$	76.79 ‡	89.13 ‡
LDA	2511/9825 terms, $k = 40$	81.17 ‡	91.23 ‡
LDA	2511/9825 terms, $k = 50$	81.16 ‡	**91.94**
LDA	2511/9825 terms, $k = 60$	**82.01**	91.30
LDA	2511/9825 terms, $k = 80$	81.31	91.17
LDA	26,005/36,928 terms, $k = 20$	76.83 ‡	86.98 ‡
LDA	26,005/36,928 terms, $k = 40$	81.58 ‡	**90.00**
LDA	26,005/36,928 terms, $k = 60$	**82.25**	89.76
LDA	26,005/36,928 terms, $k = 80$	81.67 ‡	89.80

Table 8
Example of six topics with their top 12 most frequent words.

Prob.	Topic#17	Prob.	Topic#3	Prob.	Topic#10	Prob.	Topic#4	Prob.	Topic#9	Prob.	Topic #20
0.101	the	0.087	,	0.087	he	0.124	the	0.062	.	0.079	the
0.088	,	0.054	.	0.066	,	0.064	.	0.050	to	0.041	.
0.055	of	0.047	she	0.065	his	0.061	,	0.043	the	0.039	,
0.054	.	0.042	to	0.043	to	0.045	was	0.033	and	0.028	of
0.038	and	0.042	her	0.039	.	0.041	and	0.032	of	0.026	to
0.036	a	0.040	a	0.039	the	0.037	a	0.024	a	0.026	a
0.026	is	0.035	and	0.037	a	0.032	in	0.023	in	0.024	scotland
0.022	in	0.029	in	0.032	and	0.024	to	0.020	,	0.022	and
0.017	‘	0.026	the	0.030	in	0.023	had	0.018	are	0.021	that
0.014	s	0.015	is	0.025	of	0.022	of	0.018	we	0.019	in
0.013	it	0.012	with	0.022	is	0.017	were	0.016	not	0.016	rugby
0.011	with	0.011	film	0.021	has	0.012	for	0.016	they	0.015	is

Table 9
Topic distribution according to three authors (GH corpus).

	T #3	T #4	T #5	T #9	T #10	T #13	T #15	T #17	Others
Davidson	0.047	0.054	0.055	**0.157**	0.009	0.109	0.047	**0.347**	0.175
Fowler	0.062	**0.104**	0.062	0.048	0.060	0.057	0.050	**0.390**	0.168
Russell	**0.163**	0.066	0.067	0.038	**0.100**	0.070	0.080	**0.212**	0.204

In this table, we clearly see that the function words dominate the top part of these distributions. We can see mainly determiners (*the*, *a*), prepositions (*in*, *of*, *to*, etc.), conjunctions (*and*, etc.), punctuation symbols, the possessive form (*‘s*), pronouns (*she*, *we*, etc.), and auxiliary verb forms (*is*, *are*, *has*, etc.). In Topic #20 we can see the content words *Scotland* and *rugby* while under Topic #3 we can find the word *film*.

These topics reflect the different word distributions used in the documents composing the underlying corpus. We can use the topic distribution generated for each of the 5408 *Glasgow Herald* articles to define a topic distribution per author. To achieve this, we compute an average topic distribution for each journalist based on all articles written by that person.

After doing this computation, we obtain a topic distribution for each author and an example is provided in Table 9. In this case, we have limited our analysis to the three journalists related to the “Arts & Film” domain (see Table 1). Moreover, we have ignored all topics representing less than 5% for all three chosen journalists (these 12 remaining percentages are regrouped under the label “Others” in Table 9).

As we can see from data depicted in Table 9, Julie Davidson’s profile is mainly related to Topic #17 (34.7%), and Topic #9 (15.7%). With John Fowler, we can see that his articles are mainly related to Topic #17 (39%), and Topic #4 (10.4%) while William Russell is first related to Topic #17 (21.2%), then to Topic #3 (16.3%) and Topic #10 (10%). Clearly, these journalists belonging to the “Arts & Film” domain have in common a word usage reflected by Topic #17 for which the corresponding word distribution is depicted in Table 8. In this distribution, the most probable word type is the determiner *the* (10.1%) followed by the comma (8.8%), and the preposition *of* (5.5%). In the fifth rank, we can see the conjunction *and* (3.8%). On the other hand, the full stop appears only in the fourth rank. These elements seem to indicate that these three authors tend to write longer sentences than the mean (other topics have the full stop closer to the top).

Topic #17 represents the common vocabulary they share. But Topic #3 and #10 are useful to discriminate Russell from the others, while Topic #4 does the same with Fowler. Referring to word frequencies shown in Table 8, we can see that Russell’s style can be characterized by the frequent use of pronouns (*she*, *her*, *his*) or by the content word *film*. According to word distribution of Topic #4 (see Table 8), Fowler uses the verb forms *was*, *were*, or *had* more frequently, showing maybe a preference for the past tense and/or the passive voice. Finally, the word usage reflecting Julie Davidson’s style is also related to Topic #9.

Even if for the “Arts & Film” subject we have found a common word distribution between the journalists writing in this domain, we are not able to find a similar pattern for the five journalists working in the sports columns. Each of them is more strongly related to a distinct topic, as for example, Derek Douglas who is related mainly to Topic #20 (32.8%), and the second most important word distribution is Topic #4 (10.4%). Based on word distributions shown in Table 8, we can see that words related to *rugby* or *Scotland* are mainly listed under Topic #20 while the usage of functional words appears under Topic #4, a word distribution shared with John Fowler.

8. Conclusion

In this paper we described the authorship attribution problem in the context of the closed-class problem in which the correct author is one of possible known candidates. As an attribution scheme, we have described the Delta rule method (Burrows, 2002) based on the top m (=400) most frequent word types. As a second authorship attribution method we

evaluated the χ^2 measure (Grieve, 2007), based on word types and punctuation symbols having a minimal document frequency of two on a per-author basis (generating 653 terms with the GH corpus or 720 Italian words). As a third baseline, we used the KLD scheme proposed by Zhao and Zobel (2007) and based on a predefined set of 363 English words. We have adapted this approach to the Italian language by using a known stopword list (containing 399 terms). These first three solutions represent classical approaches to the authorship attribution problem. As a forth baseline we have used the naïve Bayes model (Mitchell, 1997), representing a well-known model in the machine learning domain. In this case, we have applied both the odd ratio (OR) and the document frequency (*df*) as selection function.

As a new authorship attribution scheme, we suggest considering the LDA (Blei et al., 2003) paradigm that was applied for different purposes as, for instance, to produce an overview of *k* topics appearing in a given corpus (as shown in this paper). We explained how we could adapt this scheme for authorship attribution applications.

To evaluate and compare these different authorship attribution approaches, we have extracted 5408 articles of the *Glasgow Herald* written by 20 well-known columnists. This corpus is stable and freely available (CLEF-2003 test suite). Since these articles are written during the same short period (year 1995), sharing the same culture, having similar language registers and targeting the same audience, this corpus presents a pertinent test bed for authorship attribution empirical studies. As a second test collection, we have extracted 4326 articles from the Italian newspaper *La Stampa* written by 20 different journalists. This corpus covers the year 1994 and is also available in the CLEF test suite.

Using the same feature sets, the adapted LDA scheme produces significantly better performance levels than the Delta rule and the χ^2-based method. On the other hand, the suggested approach achieves lower performance levels than the KLD model. Based on more terms, the LDA-based scheme may perform better than the KLD model. Using the same feature sets, and with an appropriate number of selected terms, the LDA-based scheme may show better performance than the naïve Bayes model. For other parameter values, the naïve Bayes may demonstrate a better effectiveness. Finally, the underlying computational cost of the LDA-based model is higher compared to other solutions.

References

Abdi, H. (2007). Distance. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 280–284). Thousand Oaks: Sage.

Argamon, S. (2006). Introduction to the special topic section on the computational analysis of style. *Journal of the American Society for Information Science & Technology, 57*, 1503–1505.

Arun, R., Saradha, R., Suresh, V., Narsimha Murty, M., & Veni Madhavan, C.E. 2009. Stopwords and stylometry: a latent Dirichlet allocation approach. In: Proceedings of the NIPS workshop on applications for topic models. Whistler (BC).

Baayen, H. R. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Blei, D.M. 2011. Introduction to probabilistic topic models. Working paper. <www.cs.princeton.edu/~blei/papers/Blei2011.pdf>.

Blei, D. M. (2012). Probabilistic topic models. *Communication of the ACM, 55*, 77–84.

Blei, D. M., & Lafferty, J. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), *Topic models, text mining: classification, clustering, and applications* (pp. 71–94). London: Taylor & Francis.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Machine Learning Research, 3*, 993–1022.

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

Carman, M. J., Crestani, F., Harvey, M., & Braillie, M. (2010). Towards query log based personalization using topic models. In *Proceedings ACM-CIKM* (pp. 1849–1852). New York: The ACM Press.

Conover, W. J. (1980). *Practical nonparametric statistics. 2nd ed*. New York: John Wiley & Sons.

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

Crawley, M. J. (2007). *The R book*. Chichester: John Wiley & Sons.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge: Cambridge University Press.

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

Griffiths, T., Smyth, P., Rosen-Zvi, M., & Steyvers, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings ACM-KDD* (pp. 306–315). New York: The ACM Press.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings ACM-SIGIR* (pp. 50–57). New York: The ACM Press.

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities, 37*, 151–178.

Hoover, D. L. (2007). Corpus stylistics, stylometry and the styles of Henry James. *Style, 41*, 160–189.

Joachims, T. (2001). A statistical learning model of text categorization for support vector machine. In *Proceedings of the ACM-SIGIR* (pp. 128–136). New York: The ACM Press.

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

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval, 1*, 1–145.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science & Technology, 60*, 9–26.

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

Lin, S. (1991). Divergence measures based on Shannon entropy. *IEEE Transactions on Information Systems, 37*, 145–151.

Love, H. (2002). *Attributing authorship: An introduction*. Cambridge: Cambridge University Press.

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge: The MIT Press.

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

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

Nanavati, M., Taylor, N., Aiello, W., & Warfield, A. 2011. Herbert Wets – Deanonymizer. In: Proceedings USENIX workshop on hot topics in security (HotSec'11), San Francisco.

Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds.). (2004). *Comparative evaluation of multilingual information access systems*. Berlin: Springer (LNCS #3237).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the uncertainty in artificial intelligence* (pp. 487–494). Arlington: The AUAI Press.

Savoy, J. (2001). Report on CLEF-2001 experiments. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Cross-language information retrieval and evaluation* (pp. 27–43). Berlin: Springer (LNCS #2069).

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

Seroussi, Y., Zukerman, I., & Bonhert, F. (2011). Authorship attribution with latent Dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning* (pp. 181–189). Madison (WI): The ACL Press.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science & Technology, 60*, 214–433.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis: a road to meaning* (pp. 427–448). Mahwah: Laurence Erlbaum.

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: analyzing North American library and information science dissertations using latent Dirichlet allocation. *Journal of the American Society for Information Science & Technology, 62*, 185–204.

Witten, I. H., & Franck, E. (2005). *Data mining. Practical machine learning tools and techniques*. Amsterdam: Elsevier.

Xing, W., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings ACM-SIGIR* (pp. 178–185). New York: The ACM Press.

Yang, Y., & Pedersen, J.O. 1997. A comparative study of feature selection in text categorization. In: Proceedings conference on machine learning ICML (pp. 412–420).

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings ACM-SIGIR* (pp. 42–49). New York: The ACM Press.

Zhao, Y., & Zobel, J. 2005. Effective and scalable authorship attribution using function words. In: Proceedings AIRS Asian information retrieval symposium (pp. 174–189).

Zhao, Y., & Zobel, J. (2007). Entropy-based authorship search in large document collection. In *Proceedings ECIR* (pp. 381–392). Berlin: Springer (LNCS #4425).

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