Computational Forensic Authorship Analysis: Promises and Pitfalls

Shlomo Engelson Argamon

Illinois Institute of Technology, USA

Abstract. *The authorship of questioned documents often constitutes important evidence in criminal and civil cases. Linguistic stylistic analysis can often help to determine authorship. Computational methods have been applied to authorship analysis in academia for decades, and in recent years have achieved the levels of reliability needed for application to real-world cases. This article surveys the different types of computational authorship analysis methods and their components in a practical vein—describing the assumptions each makes, the analytic controls they require, and the tests needed to measure and ensure their reliability. Specifically, I discuss many of the potential pitfalls in their application, to guide practitioners in more effectively achieving trustworthy and understandable results. It must always be remembered, though, that there is no substitute for expertise, experience, and careful human judgment.*

Keywords: *Authorship, computational forensic linguistics, computational autorship analysis, reliability.*

Resumo. *A autoria de documentos questionados constitui, muitas vezes, prova importante em casos civis e criminais. A análise linguística estilística ajuda frequentemente a determinar a autoria. Na academia, há várias décadas que os métodos computacionais são aplicados à análise de autoria, tendo, recentemente, alcançado os níveis de fiabilidade necessários para aplicação em casos reais. Este artigo apresenta uma revisão dos diversos tipos de métodos de análise de autoria computacional e os seus diversos componentes numa perspetiva prática—descrevendo os pressupostos de cada um, os controlos analíticos de que necessitam, e os testes necessários para medir e assegurar a sua fiabilidade. Especificamente, discuto muitas das possíveis armadilhas inerentes à sua aplicação, de modo a ajudar os peritos fornecendo-lhes orientações para alcançarem resultados mais fiáveis e compreensíveis. Não podemos esquecer, contudo, que não existe qualquer substituto para a especialização, experiência e cuidadoso julgamento humano.*

Palavras-chave: *Autoria, linguística forense computacional, análise de autoria computacional, fiabilidade.*

Introduction

Computational methods for authorship attribution have grown in importance for forensics as they have become more accurate and more applicable to real-world situations. A well-publicized recent case of computational authorship attribution (if not in a forensic context) was the 2013 computational unmasking of J. K. Rowling as the author of the novel *The Cuckoo's Calling* by (independently) Peter Millican and Patrick Juola (Mostrous, 2013; Zimmer, 2013). They were contacted by London's *Sunday Times* to confirm a tip that Rowling had pseudonymously written the book. The two researchers independently performed computational stylometric analyses that pointed towards Rowling as a more likely author than some other plausible candidates; when shown the evidence, she reluctantly admitted that she was the author.

Of course, it is rare for a forensic authorship question to end with an unequivocal confession, and so the question of the strength and reliability of the evidence adduced is critical. *Daubert's* criterion that a method have "known or potential rate of error" is not a simple question to answer, since performance of any method will depend greatly on the specifics of the case. It can be tricky to ensure that the right analysis method is used for the task, to design the analysis protocol to produce reliable results, and to properly assess the strength of the resulting evidence. There are many parameters that must be determined and set, and there are no simple formulas for doing so that are valid in all cases. Always expert judgment is a key factor.

This article provides guidelines for using computational authorship attribution in the forensic context (and for critiquing such use). Specifically, my aims here are to show (i) how current computational methods can be used for authorship attribution, (ii) the promise they bring to forensic authorship analysis as a complement to traditional linguistic techniques, and (iii) how to recognize and avoid common methodological pitfalls in their application.

Overview of the Process

The process of applying computational authorship attribution starts with three key choices (partly externally constrained):

- Choose an attribution algorithm/method to use;
- Create a corpus comprising two or three subcorpora: the questioned texts (Q) of unknown authorship (provided as part of the case), a set of known texts (K) by candidate authors (usually provided by the attorneys), and possibly (depending on the method to be used) some comparison texts (C);
- Determine what features of the texts to extract and what measurements of their occurrence to use to characterize each text.

While these choices can be made separately—different methods can be applied to the same corpora and features, different feature sets can be used with a single method, etc.—they are significantly interrelated. Some features may work better with some methods than others, and the choice of method may have strong implications for how the corpus is constructed and vice versa.

Given a corpus, method, and features, the attribution process is as follows, in broad outline:

1. Evaluate the chosen method on texts of known authorship to establish the reliability of the method for the given case;

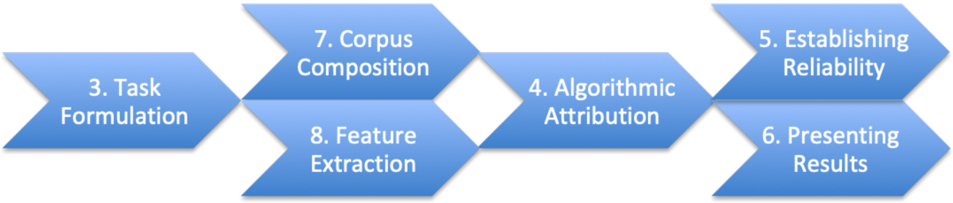

Figure 1. The flow of the overall process of authorship attribution, showing which sections of this article deal with each subtask.

2. Apply the chosen method to K, Q, and C to form an analysis of the authorship of the texts in Q;
3. Evaluate the meaning and significance of attribution results in the context of the given case.

There are different ways to implement each step of the process, some of them valid only for certain methods or in certain circumstances; we will discuss these considerations below.

In the remainder of the article I will discuss the considerations that should go into these six elements of the process, and the pitfalls that must be avoided to ensure trustworthy results. Each of the following sections treat one main aspect of the process and method of computational authorship attribution, as depicted in Figure 1. The article does not follow the order of the process, for expository reasons.

It must be emphasized that this article is naturally only a roadmap, and the mere fact that an analysis avoids the pitfalls discussed herein cannot guarantee its validity—expert judgment must always be applied to the specifics of any case.

Task Formulation

Before discussing different computational authorship attribution methods, we must first discuss the variety of attribution tasks that can be addressed. Different formulations of the task will be appropriate for different cases, as we will see.

The simplest task formulation, *author classification* is where a set of specific candidate authors with known writings is given. For example, the *Federalist Papers* are a series of articles published pseudonymously by Alexander Hamilton, James Madison, and John Jay in 1787 and 1788 to promote the ratification of the new United States Constitution. In this case, famously addressed in Mosteller and Wallace's (1964) landmark stylometric study, there are three candidate authors, and the problem is to classify each article to its correct author. Or consider the authorship question of the various sections of the late 16th Century play *The Raigne of King Edward the Third*, whose authorship is widely disputed. Much of the play is attributed to Shakespeare, but many sections are variously attributed to several other period playwrights, mainly Thomas Kyd, Christopher Marlowe, Michael Drayton, and George Peele. The attribution question for a particular section (say, one scene) could be formulated as "Which of these five individuals wrote this section?"

In general, the larger the number of candidates, the harder the task is to solve. Even if a set of candidate authors is known, it is often necessary to consider the possibility that some unknown author outside that set is the actual author (i.e., to allow "unknown" as an answer to the classification question). This setting, *open-set attribution*, is more difficult to solve reliably than *closed-set* attribution, where the candidate set is known (or can be assumed) to contain all possible authors of the questioned text.

An important form of open-set attribution is *author verification*, where there is only one candidate author A and the task is to determine whether or not that individual was the author of the questioned document or not (Koppel *et al.*, 2007; Halteren, 2007; Koppel *et al.*, 2007). One important version of verification is when we are asked whether two documents X and Y were authored by the same person (Koppel *et al.*, 2012b).

As an example of verification, consider the question of whether the book *The Cuckoo's Calling* was written by J. K. Rowling, or not, as mentioned above. To analyze this question, Patrick Juola compared the style of the book with that of one other book by each of J. K. Rowling and three other British mystery authors, Ruth Rendell, P. D. James, and Val McDermid. The question was whether Rowling was a noticeably more likely author than the other three, which would provide some evidence for or against her authorship of *The Cuckoo's Calling*.

A solution to verification can also be used to solve general open-set author classification by comparing Q to the known documents for each candidate and attributing it to the author whose documents are most reliably same-authored with Q, and if none are, giving the result "unknown".

Verification is more difficult than classification, and requires different methods, since the alternatives include everyone in the world other than A.

In cases where a specific set of candidate authors is not available, *authorship profiling* can sometimes be useful, determining demographic and social characteristics of the author based on language use. Such profiling is based on comparing features of Q with features drawn from analysis of large datasets labeled for the profile categories of interest, such as author age, sex, education, linguistic background, and the like. As a general rule, due to its broader conclusions, authorship profiling is more useful for investigations rather than for evidence of specific authorship.

Pitfall 0(a) (Match the task formulation to the case) *Different formulations of authorship attribution make different assumptions about the nature of the data and the question to be answered. Make sure that your formulation of the problem matches the structure and evidential requirements of the case.*

Algorithmic Attribution

Methodological foundations

Most of the methods for solving the above problems rely on a fundamental notion of computing form of *stylometric similarity* and comparing its values for different texts.

There are many ways to devise a similarity measure for this purpose, and we will discuss some of the details of doing so below. In the vast majority of approaches, a similarity measure is constructed by first identifying a number of textual features which are presumed to be more-or-less indicative of style and authorship. The collection of the frequencies of these features in a given text then is considered to characterize the style of the text. For example, in one of Mosteller and Wallace's (1964) foundational studies of the *Federalist Papers*, they used a set of 68 function words as features. They thus characterized each of the Federalist Papers by a vector of 68 numbers, each the frequency in the document of one of the function words from their list. In Section ‘Feature Extraction’ below we discuss the choice of features and how this may affect the reliability of authorship attribution.

Given a set of features, measuring the similarity of two texts comes down to measuring the similarity between two numeric vectors representing the frequencies of all the features in each of the documents. The more similar are the corresponding frequencies, the more similar the two texts are, in terms of the features that have been counted. A number of different mathematical formulations have been proposed for calculating a score for measuring similarity—the most commonly used today are:

- *cosine similarity*, commonly used in information retrieval (Salton and Lesk, 1968), computed for two vectors $\langle x_1, x_2, \cdots, x_n\rangle$ and $\langle y_1, y_2, \cdots, y_n\rangle$ as

$$\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

- *min-max similarity* (or *Ruzicka similarity*), which has been recently shown to be particularly effective in authorship attribution applications (Kestemont *et al.*, 2016; Halvani *et al.*, 2018), computed as

$$\frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$

This *feature-vector approach* to computing a measure of stylometric similarity between two texts comprises the steps of:

1. Identify the features of interest in each text;
2. Count the number of occurrences of each feature type in the texts, and normalize them to compute relative frequencies (as a fraction of total tokens in each respective text), giving a numeric feature vector for each text;
3. Compute a similarity score from the two vectors.

The precise character of the resulting similarity measure will depend on what textual features are chosen, how frequencies are normalized, and what similarity scoring function is used. All of these must be taken into account when comparing different methods.

Other document representations have also been used to construct useful similarity measures for authorship attribution. Similarity of graph representations of word type collocations (Arun *et al.*, 2009; Vilariño *et al.*, 2013) in documents can be compared by measuring the similarity of the graphs directly. Sequence-based “string kernel” methods (Lodhi *et al.*, 2001; Cancedda *et al.*, 2003; Xing *et al.*, 2010), developed for general text and genome comparison can also be used. In each case, the correlation of the chosen similarity measure with likelihood of authorship (and, as far as possible, independence of topic and text type) for the relevant texts must be established.

Stylometry and attribution

Now, let us suppose that we have in hand a reliable stylometric similarity measure M, such that we can assume that the likelihood that two texts have the same author is (roughly) proportional to the similarity of the texts under M. (This is of course a strong and unrealistic assumption; we will discuss how to deal with this fact further below.)

Given such an M, we can solve authorship attribution in a relatively straightforward manner.

For author classification, we would compare the questioned text Q to each of the known texts K_1 through K_n, and choose the author whose known texts are most similar to Q. If there is a near-tie, then we might have evidence of co-authorship. And if none of the known documents are sufficiently similar, and we have a large number and variety of known documents, we may conclude with some degree of certainty that Q's author is not one of the candidates.

This intuitive algorithmic schema is not, however, quite sufficient in practice. First, how do we devise a stylometric similarity measure that will have the desired correlation with authorship? Next, even given such a measure, what do we mean by "sufficiently similar"? How similar is similar enough? Third, how reliable can such a similarity measure be anyway? How can we know how reliable it is? Perhaps more importantly, since no similarity measure will be perfectly reliable, can we devise methods that are robust to not-perfectly-reliable similarity measures? How can characteristics of the known and questioned documents, such as number and length of documents and their genres, affect results? Finally, this overall framework does not tell us how to directly address the verification problem (we have no alternative candidates) or the profiling problem. We will now turn to outlining different specific algorithmic approaches which deal with these questions in a variety of ways.

Classifier learning

Perhaps the most straightforward approach is *classifier learning*, in which the set of known documents, each labeled with its correct author, is used as input to a classifier learning algorithm, whose output is a *classification model* m which outputs a predicted author for any input text it is given. A great variety of different classifer learning algorithms have been developed, many of which can be meaningfully applied to authorship attribution. Each has its own strengths and weaknesses, and the reliability of any particular method needs to be established for the particular task at hand. While a method should be chosen for its *plausibility* on the given problem based on the research literature, its reliability must also be evaluated on the available data, since the specifics of the scenario (the number of candidates and texts per candidate, the lengths, genres, diversity, etc. of the texts, and so forth) will affect accuracy, sometimes significantly (see Section 'Establishing Reliability' below).

The specific choice of classification algorithm, however, is less important than the composition of the corpus of known documents relative to the questioned document, and the choice of linguistic features by which to represent the character of a text. I briefly give an overview of classification learning here; for more detail about machine learning and how to use it, see (Domingos, 2012).

A classifier learning system C (see Figure 2) takes as input a set of known documents, each with a label (collectively the *training set*), represented as a set of document/label pairs, $\{\langle d_i, L_i\rangle\}$ – in authorship attribution, each label L_i is the known author of the corresponding document d_i. The output of C is a classification model m, which itself takes as input a document d and outputs a predicted label L. The goal is that m should classify new documents (not in the training set) with high accuracy. A key question therefore, which we will discuss in Section 'Establishing Reliability' below, is how to

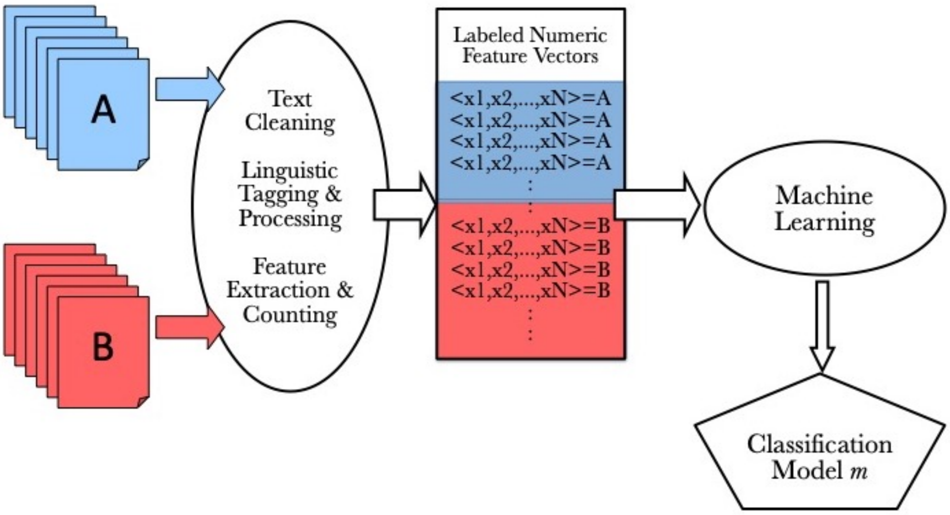

Figure 2. Flowchart showing the classification process (for two candidate authors). See text for details.

effectively evaluate C's ability to produce a high-accuracy m. Such evaluation is done using a separate *test set* of document/label pairs, where the labels assigned by m are to be compared to the 'correct' labels given in the test set.

Pitfall 0(b) (Evaluate methods on case documents) *Do not assume that a particular classification algorithm will work well for a given case, just because it has been shown to work in published research. If the texts used in that research differ qualitatively or quantitatively from those in the case, or the features used differ, results can be noticeably different. Whenever possible, you should evaluate the chosen method on the given documents in the case as well.*

Keep in mind that a classification model m will always give *some* answer for any text, so it is helpful (if possible) to use a method that can also give a (validated) measure of m's confidence in its answer. Such a measure, if reliable, can give a clearer picture of the strength of evidence provided.

Authorship Verification

In authorship verification, we seek to determine if a particular individual, A, wrote the questioned document Q. We are provided some known documents by A, but have no other candidates for the authorship of Q—the other candidates are everyone in the world other than A. A naive approach to verification would select some number of plausible alternative authors and show that A is more likely than any of them, using a classification approach. This will neither be reliable, nor convincing, since it is always possible that even if A's documents are closer to Q than any of these alternates, an even closer candidate B may be lurking just around the corner. (This is not an issue per se when a candidate author set is constructed based on the facts of the case.) So more sophisticated methods are needed.

Pitfall 0(c) (Verification $\neq$ 'more likely than known alternates') *If the question in the case is verification—whether or not a specific individual authored Q—it is not enough to just show that Q is a more likely author than an arbitrary set of alternate possibilities,* *since there is no guarantee that it is sufficiently broad to characterize the near-infinite alternatives.*

Verification by classification

One scenario in which verification might be approached using classification techniques is when there is a known closed candidate set, but just one of the candidates is of interest. That is, the question is whether or not A wrote the document, and it is known that the author was either A or one of a small set of other candidates $B_1, ..., B_n$. Given known documents for all of these candidates, a straightforward approach would be to build a two-way classifier, deciding between similarity to A's known documents and the collected known documents of $B_1, ..., B_n$. If Q looks more like the former than the latter, than there is evidence to verify A. However, the fact that A's known documents will be less stylistically varied than those of $B_1, ..., B_n$ taken together can bias the process. This can be evaluated by also running n other similar tests, verifying authorship of B_1 versus $A, B_2, ..., B_n$, of B_2 versus $A, B_1, B_3, ..., B_n$, and so on. If the results are consistent, i.e., only one of the candidates is verified, the method can be considered potentially reliable in this case. But if many of them appear verified, the method has been shown to be unreliable in the given case.

Pitfall 0(d) (Test author vs. group classification for all candidates) *Even given a closed candidate set $A, B_1, ..., B_n$, verifying A's authorship by classifying A versus the other candidates is not a prima facie reliable procedure. You also need to probe the reliability of such binary classification by similarly verifying each of B_1, B_2, and so forth; unless all results are consistent, the original result cannot be considered reliable.*

Unmasking

One important type of scenario for which verification is the appropriate paradigm is when the potential author A is suspected of attempting to disguise their authorship. If A is at all competent at doing so, simple classification will likely fail, since they will include features that are highly uncharacteristic of their own writing, which will tend to confuse classification. This can also happen without deception, in some cases where known and questioned documents differ in extraneous ways such as genre or time of composition, that can introduce irrelevant but distinguishing features.

A method that has been shown to work well for such cases, despite this difficulty, is *unmasking* (Koppel *et al.*, 2007; Kestemont *et al.*, 2012). Suppose we have two sets S_1 and S_2 of documents (or sections of documents), where we know that each set has a single author, and we want to know if S_1 and S_2 have the same author. If there is no deception, then we could try to learn a classifier to distinguish S_1 documents from S_2 documents; if an accurate classifier can be learned, then the author is likely different, but if a learner cannot learn an accurate classifier, the sets are stylometrically indistinguishable, and so are likely by the same author. Obviously, this method will not work in the case of deception, since the (lying) author will have added artificial features to distinguish the document's style from their own, and the classifier will use them and get high accuracy.

The *unmasking* method unmasks these features by learning a sequence of classification models. After learning the first, cross-validation accuracy is measured (see below), and the features that contributed most to determining the classifications are removed from consideration. (Deception-based features are likely to be such strong features by their nature.) Then learning is repeated with a reduced feature set, and accuracy measured. Again, strong features are removed, and learning with accuracy measurement repeated. This process is repeated a number of times, giving a sequence of generally declining accuracy values (an *unmasking sequence*), as more and more features are removed. However, if the case is one of deception, and the two document sets have the same author, we expect accuracy to dip sharply after a small number of iterations, once nearly all the deceptive features have been removed. This will not occur if the sets of documents do have different authors, rather accuracy will slowly decline over the entire range. By comparing the unmasking sequence of interest to others known to be for different authors, the existence of a significant dip can be verified directly.

The impostors method

Another method that addresses author verification is the *Impostors Method* (Koppel *et al.*, 2012a; Seidman, 2013; Stover *et al.*, 2016; Potha and Stamatatos, 2017). A key advantage of this method is that it does not rely on cross-validation like unmasking, and so requires much less data to work. The impostors method takes the questioned document Q and a known document K authored by the suspect author A, and determines the strength of evidence that Q and K share an author. The procedure works by analogy to a police lineup: In addition to Q and K, a set of *impostors* I_i is put together comprising documents by authors other than A which are as similar as possible in other ways to K and Q. The idea is that if the similarity between Q and K is more than that between Q and the impostors, then it is likely that Q and K share authorship. The impostors thus serve to normalize the similarity measure, telling us how similar we expect random pairs of documents to appear. The greater the number of independent impostors, the stronger the evidence is.

Pitfall 0(e) (Use enough impostors, similar to Q and K) *Use a sufficient number of impostors, and use impostors that are as similar as possible to both Q and K in all ways other than authorship.*

It is still possible, however, that Q and K are most similar by coincidence. Hence the full impostors method runs a large number (usually 100) trials, in each of which only a random subset of features is used for computing similarity. This way if the similarity of Q and K is only a coincidence, it will not often recur. So if Q and K are more similar than Q and any impostor in a large number k of these trials ($k > n$ for some threshold n), the evidence of coauthorship can be considered to be reliable. (Note that if $k < n$ that is not evidence against coauthorship, just the failure to make a positive attribution.) The choice of n will determine the false-positive and false-negative rates of the method—higher n will mean fewer erroneous attributions, but more missed attributions, and a lower n the reverse.

In published tests under laboratory conditions, and attribution threshold of $n = 20$ (out of 100 trials) gives false positive rates of below 10% (Koppel and Winter, 2014), but as for all methods, it is always advisable to test the impostor method on the texts of the case at hand. This can be done given sets of known documents by the suspect and other similar authors, considering how many same-author pairs are properly attributed and how many missed, and how many different-author pairs are improperly attributed and how many are not. This will give estimates of the false-positive and false-negative rates, which can be calibrated for how conservative a result is desired. For evidence, a conservative result, which has a low false-positive rate, is desirable, so that if an attribution is made, it can be considered reliable. For investigations, a higher false-positive rate may be acceptable, if it lowers the likelihood that the actual author will slip through the net.

Pitfall 0(f) (Consider false-positive/false-negative tradeoff) *Consider whether your case requires conservative (only-if-very-sure) attribution, and thus a higher threshold for attribution.*

Pitfall 0(g) (Determine thresholds before testing) *Determine the threshold based on the literature and based on calibration tests on known documents* before *score attributions for Q, to avoid choosing a threshold that fits the results rather than interpreting the results based on a threshold.*

Finally, since the impostors method relies on statistics of large numbers, texts must be relatively long; the overall feature set must also be large to support many trials with random subsets.

Pitfall 0(h) (Use long documents and many features for the impostors method) *A rule of thumb based on research investigations is that texts should be around 2000 words long or longer to ensure reliability. (Shorter texts can be used if necessary, but reliability will degrade as shorter texts are used.) As well, since the impostors method performs many trials with different random subsets of an overall feature set, the full feature set must be relatively large (over 1000 features in general) to ensure sufficient variability among the subsets.*

Visual attribution

Since, as noted above, all computational attribution methods rely in some fashion on measuring some kind of similarity between documents, we might think of dispensing with fancy algorithmic attribution methods such as those we just discussed, and instead producing a visual representation of the stylometric relationships between documents and then visually determining which candidate author Q is most similar to, if any. This would be straightforward if, say, there were only two relevant linguistic features, so that every document would be represented as a pair of numbers (x, y) corresponding to the relative frequencies of the two features. Then we could plot all known documents on a graph, as in Figure 3, and determine the location of Q (also as a pair of numbers) compared to the known documents for each candidate. If Q's point is clearly in the 'cloud' of points for a particular candidate (as is the black circle in the figure), that gives good evidence for an attribution, and if it is far from any candidate documents' points (as is the black triangle in the figure), no attribution can be made (and possibly we have prima facie evidence to deny any candidate authorship if we can show that the known documents cover the full range of all the candidates' writings).

The trick, of course, is that textual style cannot be adequately represented by two numbers, however computed. Any plausible set of stylometric features will have dozens, if not hundreds, and so if we are to plot the known and questioned documents in two dimensions, we need to somehow boil a large number of dimensions down to two. Fortunately, there are standard statistical methods for doing so, which have been applied to authorship attribution.

The oldest and most standard such technique is *principal component analysis* (PCA). This rotates a set of numeric vectors in a multidimensional space to find axes such that the data distribution along each axis is statistically (linearly) uncorrelated with those

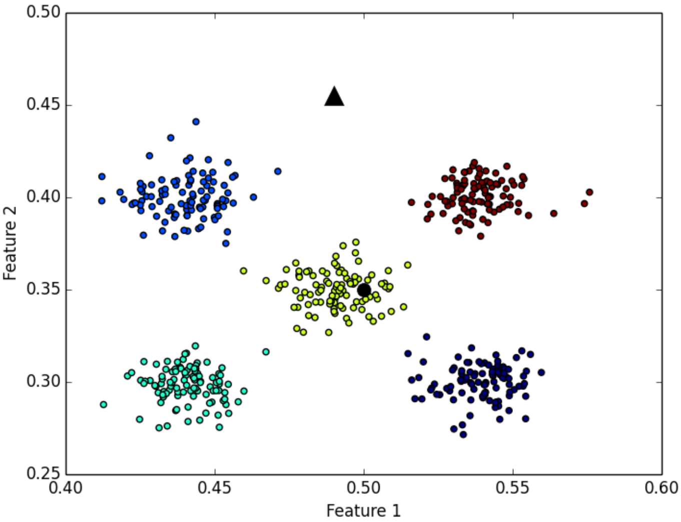

Figure 3. Simulated two-dimensional visualization of known documents from five clearly separated candidate authors shown in different colors, with two hypothetical questioned documents shown as a black circle and triangle.

along other axes (Abdi and Williams, 2010). These axes, called principal components, are ordered in descending order of how much data variability each contains. Hence, the first two components will give the data the widest spread of any two dimensions we could choose, and thus provide arguably the best two-dimensional representation of the data. For example, in his analysis of the authorship of the 15th Book of Oz, José Binongo (2003) plots known segments of Oz books known to be authored by the two candidate authors, L. Frank Baum and Ruth Plumly Thompson, per their first two principal components; we reproduce his figure in Figure 4. In this case, the known documents can be separated between the candidates perfectly using just the first principal component; we note that this level of clarity is very rare in practice.

Another technique for plotting high-dimension data in two dimensions is *multidimensional scaling* (MDS), which seeks to find an embedding of data points in two dimensions which maintains the relative distances between points, as much as possible (some distortion is inevitable, of course). MDS has been used similarly to PCA in authorship attribution research (López-Escobedo *et al.*, 2016). The techniques will give somewhat different results, as they are based on different definitions of what constitutes a 'good' reduction of the data to two dimensions, but the considerations and caveats for properly using them are similar.

The key such consideration is the deceptive simplicity of a scatter plot such as that in Figure 4. Visual inspection gives a clear answer—if Q falls on one side it was written by A, and if on the other side, B. The figure hides the complexity and statistical assumptions behind the result. The same procedure carried out on a slightly different set of documents, or on the same documents with different features, can give significantly different results. Using a different set of relevant documents also might. These possibilities need to be considered and ruled out, instead of simply relying on the force of visual clarity that the figure provides.

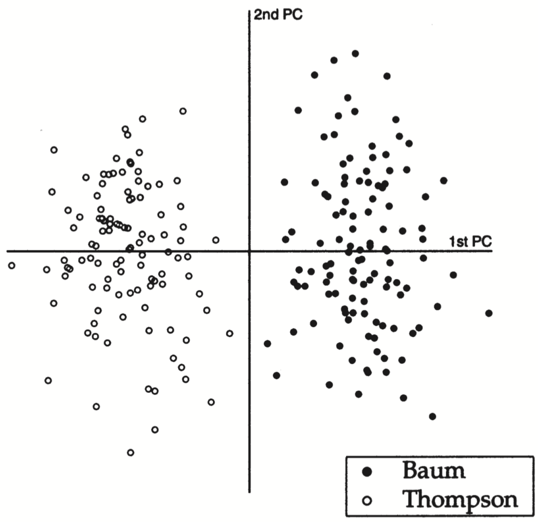

Figure 5. Baum vs. Thompson.

Figure 4. Texts by Baum vs. Thompson, plotted by first two principal components (Figure 5 of Binongo, 2003).

Pitfall 0(i) (Sensitivity testing for visualizations) *Dimensionality-reduced visualizations rest on complex assumptions and algorithms—don't just rely on visual clarity. You should probe how sensitive plots are to changes of features, similarity measures, and document sets before relying on them.*

That said, if a reliable two-dimensional plot can be constructed that gives a meaningful and useful answer, it can be very useful in making analysis results comprehensible to the judge and jury.

Clustering

One of the main goals of these visualization-based techniques is show how (or whether) the known documents divide into clear clusters by authorship, so that Q's authorship can be attributed by ascertaining which cluster it best belongs to. This idea can be implemented directly by using one of a number of *clustering algorithms* (Han *et al.*, 2011: Ch. 10) and see Berry and Castellanos (2004) and Xu and Tian (2015), which automatically divides known documents into a set of clusters, according to some criterion for the quality of such a division. Clustering is an *unsupervised* method of analysis, which does not use information about the authorship of the known documents to divide them into sets of stylistically similar documents. The idea is that if such sets correspond to specific authors, then the clustering has captured the stylistic correlates of authorship in the corpus, and the cluster identity of the questioned document is a likely indicator for its authorship.

Clustering has a long history of use in authorship analysis, as in Holmes and Forsyth's pioneering study of the Federalist Papers (1995) and Burrows's later application of the method to literary analysis of poetry and prose (2004). Cluster analysis for forensic authorship analysis may be less reliable, though, due to shorter text lengths and smaller corpora; the reliability of cluster analysis for literary texts has also been questioned (Hoover, 2003).

Even when considered effective, the results of clustering are highly sensitive to experimental parameters, such as the number and types of features, the distance measure used to compare feature vectors, and the way distances are aggregated to compare clusters with each other (Jain *et al.*, 1999; Halkidi *et al.*, 2001; Zaïane *et al.*, 2002). This difficulty can be somewhat mitigated through recent techniques that build consensus clusterings, combining information derived from many different parameter settings (Eder, 2017), but without a sensitivity analysis results cannot be considered reliable, just as noted above for visualizations.

Pitfall 0(j) (Sensitivity testing for clustering) *Clustering results can vary greatly depending on system parameters. Probe how sensitive results are to changes of features and other parameters before relying on them.*

Establishing Reliability

To evaluate an attribution method's reliability, we need to run it using some known documents for training and then test the result on new data for which we also know the correct answers (a *test set*). Note that the testing data must comprise different texts than the training, since it would be trivial (and meaningless) to get perfect accuracy on the training, simply by memorizing it.

Pitfall 0(k) (Ensure disjoint train and test sets) *If you test a model on the same documents used for training it, estimated accuracy will be considerably higher than you can expect for the questioned document. Make sure that training and testing are done on* different *documents.*

A difficulty, of course, is that to get an accurate model, we need as much training data as possible, but the available labeled data is usually limited. In experimental research, we gather as large a set of documents with known authors as possible, so that some can be used for training and some for testing, while in typical operational scenarios, the number of known documents is more limited. Regardless, only occasionally, even in research, do we have a truly enormous number of texts, and so we need to use those we have efficiently. The standard method to do this is *cross-validation* (Alpaydin, 2009), in which the available labeled data is divided randomly into a number (k) of equal-sized subsets $S_1, ..., S_k$, called *folds*, and k train/test evaluations are carried out. First, we apply a learning method C to build a model, training on $S_1, ..., S_{k-1}$ and test its accuracy on the last fold S_k. Then, we train on $S_1, ..., S_{k-2}, S_k$ and testing on the remaining fold S_{k-1}, and so forth, repeating the process a total of k times. The average of the k accuracy figures is then used as an estimate of the expected accuracy of C's learned model for future data. Note that cross-validation thus is able to use all of our labeled data for testing, while ensuring that at no time does it test a learned model on any of the data that was used for training it.

Even with cross-validation, however, you may have very few known documents in a given case, perhaps only two or three (or even just one) from each candidate author. In such a case, if the documents are long, one might consider increasing the number of training texts by splitting each document into sections. (See below for a discussion of text length.) Since the style within a particular document may vary slightly between sections of the document, this strategy can lead to more accurate models being constructed. However, cross-validation needs to be modified so that a model trained on part of a document is never tested on other parts of the same document. If it were, we could not know if accuracy was due to detecting the authorship of the test text or due to the simple fact that they are from one document—about the exact same topic, in the exact same register and genre, for the exact same audience, etc. Hence, in this case, the split of the known texts must be done such that all sections of a single document are in the same fold, to avoid this problem.

Pitfall 0(l) (Don't train and test on sections of the same document) *If known documents are split into multiple sections, increasing the number of training texts, a model trained on some sections of one document cannot be tested on other sections from the same document. Thus all sections from a given document must be in the same fold when doing cross-validation.*

Another possible way to overcome the paucity of data would be to use other documents with known authors, other than the known documents in the case, to evaluate the method or to supplement those documents. The danger here is that if these documents are stylistically different from the documents in the case, whether in terms of register, genre, sociolect, discourse community, etc., the comparison may be invalid. Using results on other datasets can be used to argue for the plausibility of the method for the given case, but attention must be paid to the question of how similar the kinds of documents are to each other, and appropriate caveats attached. Best in such a case is to be able to point to multiple such tests that give consistent results. However, simply adding a number of unrelated documents to known documents from the case, to construct a larger training set, is likely to lead to results that cannot be trusted.

Pitfall 0(m) (Keep training set internally consistent) *Attribution accuracy can depend on the other influences on document style for training and test texts, and thus:*

- *Evaluations on documents not from the current case must be considered relative to the similarities and differences of the provenance of those documents to those available in the case, and*
- *External documents should not be mixed together with case documents to make a larger training set. The differences will lead to unreliable evaluation results.*

A subtle, and surprisingly important, question is raised when feature selection is done. In feature selection, a very large number of potential features, such as wordforms or part-of-speech n-grams, is whittled down to a manageable size by computing some measure of each feature's usefulness for classification and keeping the 'best' k features, or all those that pass a threshold. Such measures evaluate how well an individual feature can distinguish authors from each other; a variety of statistical measures exist such as information gain (Quinlan, 2014), chi-squared statistics (Moh'd A Mesleh, 2007), etc. Any such measure must be computed over labeled training data, wherein lies the danger. If features are selected based on an entire labeled corpus, and then learning (on the reduced feature set) evaluated through cross-validation, the test documents have actually been used in the training process, since feature selection is part of training. This is a very common error that is easy to fall into, but one which can lead to surprisingly misleading results. If this is done, evaluation results often greatly overestimate the accuracy of the classification method, which may appear accurate but turn out to be useless on new data.

Pitfall 0(n) (Don't use test data in feature selection) *If you use feature selection, make sure that selection measures are computing only over training data during evaluation, and not on test data. Otherwise you will overestimate the accuracy of your method.*

It is important also to note that accuracy itself is not a simple and unproblematic notion. If many more documents are available for one candidate author X than for others, high accuracy might be obtained simply by predicting that *all* documents were written by X. For example, if 70% of the known documents are by X and only 30% by other authors, this would give 70% accuracy. However, the method is clearly useless and meaningless, though it seems somewhat accurate based on the numbers. A more fine-grained evaluation is obtained by using two different "accuracy" measures—*precision* and *recall*. Precision measures, for each candidate author A, what fraction of the documents that the model m predicts are written by A were actually written by A. Recall, on the other hand, measures what fraction of the documents actually written by A were predicted by m to have been written by A. In our example of a dumb attribution method above, the precision for author X would be 70%, but for other authors would be undefined (since no predictions are made for them); recall for X would be 100%, but for other authors would be 0% (since they are never predicted). Thus we see how by looking at both precision and recall we get a better picture of the actual performance of the method.

Pitfall 0(o) (Use precision and recall for evaluation) *Simple accuracy as a measure can be affected significantly by imbalance in numbers of known documents for different candidates, and unreliable methods may appear reliable. Better is to calculate both precision and recall for each candidate author. This will show if all authors are treated equally by the learned model or if results are biased in one way or another.*

The harmonic mean of precision and recall, called the "F1 measure," is often used to give a single numeric metric for performance of text classification or information retrieval systems. It is better to use both precision and recall, however, for a couple of reasons. First, depending on the scenario, either precision or recall may be more important—averaging them loses clarity as to the import of the results. Second, in many realistic situations, high F1 can be obtained by methods that provide no useful information (Lipton *et al.*, 2014).

Corpus Composition

In all applications of authorship attribution, we must start with a *questioned document* (or set of documents) Q, whose authorship is to be determined, and a set of *known documents* K, which are reliably known to have been authored by the candidate author(s). For some attribution methods, particularly when dealing with open-set attribution or verification, a set of *comparison documents* C is also used, comprising documents by non-candidate authors as impostors or to provide background calibration for determining what level of stylometric similarity indicates coauthorship in the case.

When inferring authorship based on stylometric comparison of different texts it is essential to keep in mind the multiplicity of factors that can influence the stylistic character of a text (see Figure 5). There are no stylometric features that uniquely indicate author identity, hence care must be taken to rule out alternative explanations for stylometric similarity between two texts. As an extreme example, suppose Q is a corporate contract, and the question is which of two authors, A1 and A2, drafted it. If we are given one known document from each, K1 and K2, respectively, where K1 is a contract, and K2 is a personal email, the fact that Q is more similar to K1 than to K2 says nothing about its likely authorship, as the similarity is easily explained by register and genre.

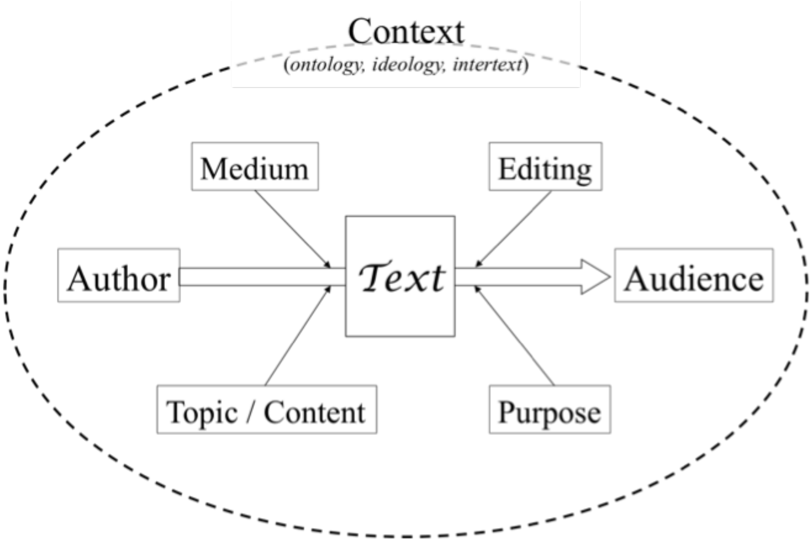

Figure 5. Summary of factors contributing to the precise form of a text (after Figure 5.1 of Argamon and Koppel, 2010). The Author seeks to express some Content about a Topic in a text via some Medium for some Purpose directed at some intended Audience. There may be Editing that affects the style and content of the text. The larger context within which the text's production is embedded also affects what text is produced, the relevant *ontology* assumed, the *ideology* encoding potential and actual social roles of the Author and Audience, and the *intertextual* relationships of the new Text with other texts that came before.

Ideally, all the documents, Q, K, and C, should be as similar as possible in all ways other than in authorship; this is the best way to ensure that inference to authorship cannot be explained by other factors. However, such a level of experimental control, exercised in laboratory research, is rarely if ever possible in the forensic context. Known documents are limited to whatever documents can be obtained for the candidate authors—there may be very few, and those that are available may be from different genres and registers from Q and from each other. It is critical to keep in mind that there are **no known** stylometric features that vary with authorship and do not vary with genre, register, topic, or other style-influencing factors (collectively, if vaguely, *text type*). Thus any differences in text type within the corpus must be accounted for, either by experimental control (which as noted is difficult to achieve in forensic cases), or by analytic procedure (see Section 'Algorithmic Attribution for discussion of how some methods can deal with differences in text types).

Pitfall 0(p) (Control corpora for text type) *If possible, ensure that all documents to be compared are of the same, or very similar, text types (genre, register, topic). If this cannot be assured, be very clear about the similarities and differences in text type and their likely influence on stylometric comparisons.*

Pitfall 0(q) (Exercise caution when Q differs in text type from K) *When Q differs in text type from known documents in K, and when known documents by different candidate authors differ in text type, consider carefully to what extent similarity judgments might be attributed to text type, as opposed to authorship, and could be misleading.*

In addition to controlling for text type, any method that relies on statistical analysis of textual features, as do the computational attribution methods discussed in this article, must also control for text length. One is tempted to assume that the relative frequency of a given feature will be roughly the same no matter the length of the text. However, this is not the case. Common features will tend to drop in frequency as a text gets longer, due to the introduction of new vocabulary (cf. Zipf's law (1935)). See, for references, the frequencies of the words 'the' and 'you' in texts of different lengths in Figure 6—after an early rise, frequencies tend to drop until the text is long enough to give a near-constant frequency. Since forensic texts tend to be short, this variability is important to account for. Hence comparison of texts should be of segments of approximately equal length—if Q is 600 words long, comparing it to K_1 of length 600 words and K_2 of length 2500 words will not be a fair comparison, as we expect K_2 to have noticeably different frequency statistics from Q on general principle having nothing to do with authorship. Better is to use excerpts of (near-)equal length from all the documents to be compared.

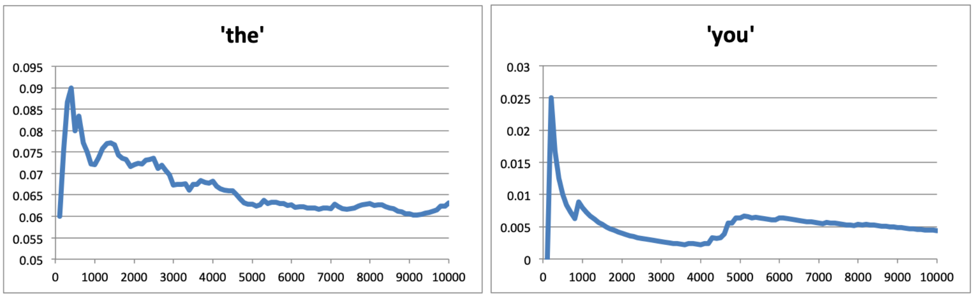

Figure 6. Frequencies of the words 'the' and 'you' in prefixes of different word lengths from the concatenated Congressional Record of the 104th–109th Congresses (Gentzkow and Shapiro, 2013).

Pitfall 0(r) (Control comparisons for text length) *Whenever possible, ensure that texts being compared are of nearly the same length, since estimates of lexical frequencies vary based on the length of the text (due to Zipf's Law). Such control is less critical when using syntactic features.*

Once we consider segmenting documents, however, we must contend with discourse structure—different sections of many kinds of document have different characters. This is true for many genres of text, including correspondence, suicide notes, contracts, essays, and more. Hence segments should respect boundaries between recognizable sections of each document, instead of comprising arbitrary segments of text of a given length. Also, assuming comparison is being done between texts of similar genre (recommended whenever possible), comparison should be between similar sections. Hence, the initial segment of one (say) letter should be compared against the initial segment of another, not its final segment. Inconsistency in this regard can lead to similarity judgments that are misleading.

Pitfall 0(s) (Control comparisons for discourse structure) *To the extent possible, ensure that text segments to be compared come from comparable portions of their respective documents. Ideally this would be based on a genre-relevant decomposition of the documents, but this can usually be approximated by using portions from the same relative positions in the respective documents (beginning, middle, end).*

The discussion above assumes that every document has a single author. While in some cases (e.g., ransom notes) this is a reasonable assumption, it is not always realistic. Editorial influence in published work can influence textual style, and some genres of text, such as contracts, are inherently multiply authored, due to collaboration and text reuse.

Pitfall 0(t) (Consider multiple authorship) *Consider directly the possibility that Q or known documents have multiple authorship or strong editorial influence. If single authorship can be assumed, make the assumption explicit and justify it.*

If single authorship cannot be strongly assumed, ideally plausibly multiply authored documents should be removed from consideration; however, this is rarely possible. Hence analysis must take the possibilities of collaborative authorship and editorial influence. How this is done depends greatly on the specifics of the case, but some general principles can be sketched.

If coauthorship of the questioned document Q is suspected, one approach is to attempt attribution separately for different sections of the document, which should detect if different individuals were primarily responsible for different sections. Except in cases where natural divisions are available (e.g., for plays, which divide into acts and scenes), overlapping sections should be chosen since we do not know in advance which portions may have been written by different people. The same idea can be applied to allowing for co-authorship of known documents, for attribution methods that treat known documents separately, so that each document section is compared to Q in isolation. In this case, it may be that Q (or a section thereof) matches only some sections of a known document, giving reason to believe that the known document may have multiple authors, and that Q may be attributable to whoever wrote those sections.

The likelihood of multiple authorship can also be directly tested by comparing the style of a document's sections to each other (Glover and Hirst, 1996; Graham *et al.*, 2005; Rybicki *et al.*, 2014; Stamatatos *et al.*, 2016). If different sections appear to show different authorial styles, they should be treated as separate units of analysis. If a known document seems to be multiply-authored, a conservative approach would simply remove it from consideration, provided that there are sufficient other known documents to proceed with the analysis.

Pitfall 0(u) (Segment documents to test and control for multiple authorship) *If multiple authorship cannot be ruled out, consider segmenting Q (and known documents) to be separately attributed. Stylistic comparison of segments of the same document can also be used to estimate the likelihood of the document being multiply authored.*

Segmenting documents will not, however, help us with the possibility of editorial influence, where authorial style is directly overlaid with other stylistic characteristics. In many cases, of course, the likelihood of editorial changes is virtually zero, as for ransom or suicide notes, but in cases involving published or institutional documents, this possibility is much more likely. Such influence may be from an editor's individual style, or from the imposition of a 'house style' on the document. Note that the implications for attribution of editorial influence are different when considering the questioned document or the known documents.

If Q's style may have been significantly affected by editorial changes, it will lessen the likelihood that any given candidate author is a strong match, since Q will bear a mixture of stylistic characteristics. Thus, if, nonetheless, just one candidate author is a strong match, the value of the evidence will be at least as large as had there been no editorial interference. However, it will be impossible to distinguish between attributed authorship and editorship—if candidate author A is a good match for Q, we cannot know if A was the author without external evidence that A was not the editor.

On the other hand, if editorial influence is suspected among the known documents, spurious similarities may be found with Q, invalidating analytic conclusions. One way to control for this, when multiple documents from each candidate author are available, is to compare known documents to each other. Each known document K is taken in turn as a questioned document, and attributed based on the remaining known documents. If editorial influence is minimal, we expect each known document to most likely be attributed correctly. If most are, but a small number are not, this may indicate editorial interference with those, and reason to exclude those known documents from consideration.

Pitfall 0(v) (Test for likelihood of editorial interference) *If significant editing of known documents cannot be ruled out, test for stylistic consistency among documents of each candidate author, and remove those that do not fit in with the rest.*

The above discussion assumes editorial influence varies for different documents. If the same editorial influence obtains for (say) all known documents by a single candidate, those documents may be stylistically consistent without clearly reflecting the style of the author—they may instead reflect the editor's style or a mixture of the two, without revealing an inconsistency. In such a case, we cannot reliably distinguish attribution to the candidate or to the editor.

Feature Extraction

We now consider the different sorts of textual features that are typically used in computational stylometric analyses, for authorship attribution as well as for others. Choice of such features must balance three considerations: their linguistic significance, their effectiveness at measuring true stylometric similarity, and the ease with which they can be identified computationally. Some potentially useful and linguistically meaningful features may not be easily (or at all) identified accurately by existing computational techniques. For example, metaphor use may be a useful feature for authorship analysis, but current automated metaphor identification methods are not accurate enough to rely upon.

Statistical complexity

The earliest work in stylometrics sought statistical measures invariant across documents by a single author but vary between authors. A great variety of such measures have been proposed, such as average word or sentence length (Fucks, 1952; Brinegar, 1963; Yule, 1939) and more complex statistics using type/token ratios and numbers of *hapax legomena* and the like, such as Yule's (1939) K, Sichel's (1975) S or Honore's (1979) R. However, no such measures have proven to be reliable for authorship attribution (Burrows, 1992; Grieve, 2007).

Pitfall 0(w) (Complexity measures are not reliable alone) *Overall measures of textual or linguistic complexity are not generally reliable for authorship attribution. Hence they should not be used except together with other features, if they increase a method's reliability. This must be demonstrated by empirical testing.*

Lexical choice

Lexical choice is a key dimension of variation between individual authors, who exhibit statistical preferences for different words that can be used in particular contexts. There are different kinds of feature sets built on this notion, as discussed below.

Function words

One of the oldest and most generally reliable feature sets used in stylometric authorship attribution is *function words*, used at least since Mosteller and Wallace's landmark study of the *Federalist Papers* (1964). Function word use (a) does not vary substantially with topic (but does with genre and register) and (b) constitutes a good proxy for a wide variety of syntactic and discourse-level phenomena. Furthermore, it is largely not under conscious control, and so should reduce the risk of being fooled by deception (Chung and Pennebaker, 2007).

Function word lists used in English are typically up to a few hundred words long and include pronouns, prepositions, auxiliary and modal verbs, conjunctions, and determiners, as well as numbers and interjections, even though they are not function words, since they tend to vary with authorship and are mostly topic-independent. The function words available for use in different languages will vary of course, and for synthetic languages will likely be incomplete and need to be supplemented by morphological analysis. Results of different studies using somewhat different lists of function words have been similar, indicating that the precise choice of function words is not crucial. Discriminators built from function word frequencies often perform at levels competitive with those constructed from more complex features.

Pitfall 0(x) (Use morphological analysis on synthetic languages) *Function word lists in synthetic languages will likely miss many important features of the idiolect, so morphological analysis is needed to extract a more complete set of features.*

When using function words for authorship attribution, attention must be paid to the fact that genre and register variation in the corpus will also affect function word frequencies. For example, pronouns (particularly first and second person) are much more frequent in narrative text than in informative text. Depending on the analysis methodology, some classes of function words may need to be removed from consideration.

Pitfall 0(y) (Filter function words based on genre and register) *Frequencies of many function words will vary greatly between different genres and registers of text, and so appropriate methods or controls need to be applied if the corpus must comprise diverse text types. This may involve removing some function words from consideration. All such controls must be validated empirically on the data.*

Content words

Other aspects of lexical choice variation are not captured by function word use. For example, one candidate author may prefer to use words like 'start' and 'large', where another may prefer 'begin' and 'big' (Mosteller and Wallace, 1964; Koppel *et al.*, 2006, 2009). This sort of pattern can be analyzed by modeling the relative frequencies of content words. Typically very rare words and those with near-uniform distribution over the corpus of interest can be omitted (Forman, 2003), so that a set of several to ten thousand words may be used. Content words, however, require even tighter experimental care and control, since their frequencies will vary with topic, as well as with text type. This may lead to both false attributions and to missing valid attributions, depending on how such irrelevant dimensions of variation may influence attribution.

Pitfall 0(z) (Using content words requires tighter corpus control) *Content words may indicate topic more strongly than authorship, so tests using them need tight controls* *on topic of corpus documents, or methods that can be shown to be stable in the face of topic differences. Examining the features that the analysis identifies as key to the attribution should be done to check if such interference is present.*

Word embeddings

Using words as features for stylometric comparison, whether function words or content words, finds similarity by comparing occurrences of the exact same word. However, some words are more similar than other. Consider a comparison between the sentences "The President spoke about tariffs" and "The administration issued a statement about import taxes." The only words shared between them are "the" and "about," however, they are very similar. Significant semantic closeness is seen in the pairs (President, administration), (spoke, statement), and (tariffs, taxes), but is not taken into account by word-based methods. A popular way to generalize word comparison is to use a *word embedding*, which represents each word by a multidimensional numeric vector such that words that occur in similar contexts will have similar vectors. One of the most popular methods, Word2vec (Mikolov *et al.*, 2013), uses a neural network model to derive such embeddings, largely capturing semantic and syntactic connections between words such that similar words have nearby vectors. They show, for example, that $\text{vec}[king]$+($\text{vec}[woman]$-$\text{vec}[man]$) $\approx$ $\text{vec}[queen]$. Recent development of *contextual* word embeddings (Devlin *et al.*, 2018; Peters *et al.*, 2018) give more precise word vectors for particular word occurrences, that are sensitive to context. These embeddings thus encode different word senses and parts-of-speech, giving a more fine-grained representation.

The hope of using such vectors for stylometric comparison, is to get more general and more precise measures of semantic similarity in lexical choice. Indeed, some recent research has shown word embeddings to give useful features for authorship analysis in research studies (Sari and Stevenson, 2016; Posadas-Durán *et al.*, 2017). Results seem fairly insensitive to what corpus was used to compute the embedding, provided it was large enough—standard embeddings trained on very large corpora are now easily available for such use. The main caveat when using word embeddings is that, just like content words, their occurrence is dependent on document topic, genre, and register, and so these factors need to be tightly controlled in any authorship analysis using them.

Pitfall 0() (Word embeddings encode topic dependence) *Word embeddings enable better determination of lexical similarity by generalizing beyond identity of word tokens. However, they share the properties of topic- and text type-dependence of content words, and analysis must be controlled accordingly.*

Syntax

Another category of style markers is the relative frequencies of different choices of syntactic structure, either measured directly, or by proxy via looking at occurrences of parts of speech. Different authors have different preferences for type and complexity of different constructs, and both absolute and relative frequencies of syntactic constructs have shown to be useful for authorship attribution (Baayen *et al.*, 1996; Stamatatos *et al.*, 2001; Gamon, 2004; Hirst and Feiguina, 2007). In all such cases, feature frequency is likely to be influenced by text type, and so experimental control is necessary (or text-type invariance needs to be demonstrated).

Pitfall 0() (Syntax also requires text type control) *Despite its facial and empirical topic independence, syntactic choice is* not *invariant to text type; different genres and registers have difference characteristic relative frequencies for various syntactic constructs. Hence full control for text type is necessary when using syntactic features as well.*

Extracting syntactic structure from text in English and most other European languages can be done accurately using current natural language processing tools, for texts in reasonably standard prestige dialect. These tools will have more difficulty on less formal text that includes orthographic and grammatical errors or variations, as well as on most languages outside the European mainstream.

Pitfall 0() (Understand accuracy of syntactic analysis tools) *Automated syntactic analysis tools vary in the accuracy of their output depending on the language (they are best for English and major European languages) and text type. They are particularly poor on informal texts. Their accuracy should be evaluated on texts of the same kind as the analysis corpus before use.*

A simple type of syntax-based feature is using relative frequencies of different parts-of-speech and of short part-of-speech sequences, e.g., "the fraction of common nouns that are immediately preceded by an adjective". A number of research studies have shown that such features can be useful in authorship attribution (Argamon *et al.*, 1998; Kukushkina *et al.*, 2001; Corney *et al.*, 2001; Koppel *et al.*, 2002; Koppel and Schler, 2003; Zhao *et al.*, 2006; Zheng *et al.*, 2006).

More complex automated parsing tools can be used to identify full syntactic structures, and compute the frequencies of noun and verb phrases or of relative clauses, for example. These have also been shown to work for authorship attribution in the research literature (Baayen *et al.*, 1996; Stamatatos *et al.*, 2001, 2000; Gamon, 2004; Halteren, 2007; Chaski, 2005; Uzuner *et al.*, 2005; Hirst and Feiguina, 2007).

Specific examples of such features are:

- N-grams of parts-of-speech: "determiner–adjective–adjective" or "common noun–common noun" (Argamon-Engelson *et al.*, 1998),
- Syntactic phrase categories: XYZ (Stamatatos *et al.*, 2001)
- Syntactic category bigrams: "coordinating conjunction followed by clause" or "name starting with proper noun" (Hirst and Feiguina, 2007), and
- Marked syntactic structures: "non-head-final noun phrase" (in English) (Chaski, 2005).

In most attribution studies, syntactic features are used together with lexical features, as syntactic features alone are not usually fine-grained enough to attain high accuracy.

Pitfall 0() (Evaluate if syntax is reliable for the specific case) *Consider well whether the syntactic features to be used are likely to be reliable for the kinds and numbers of texts in the corpus, and empirically test them. Use lexical features (e.g., function words) as well, if needed.*

Character n-grams

The relative frequencies of character n-grams (sequences of several characters), such as "ing", "auth", "opos", or the like, has been proposed as a feature set for attribution, subsuming lexical choice features (function and content words) and morphology (by capturing many affixes). Such features have the big advantage of being largely language-independent (for non-ideographic writing systems); a number of research studies have shown their efficacy for attribution in various languages and contexts (Kjell, 1994; Clement and Sharp, 2003; Houvardas and Stamatatos, 2006; Ledger and Merriam, 1994; Grieve, 2007; Kešelj *et al.*, 2003; Peng *et al.*, 2004). Since they are sensitive to topic as well as text type, all of the concerns regarding function and content words apply as well to character n-grams.

Presenting Results

No text analytic method can conclusively prove who the author of a questioned text is—a good result is one which shows where the weight of the evidence lies, with respect to the authorship question at hand, and gives some measure of the strength of that evidence. An attribution result is one of two types: it may *rule-in* a particular candidate A as a likely author of Q, or it may *rule-out* a candidate B, tagging B as an unlikely author of Q. In both cases, one must be careful to determine and explain who the alternative authors are that A (or B) is being compared against (see the discussion in Section 'Task Formulation' above comparing open- and closed-set classification and verification tasks).

Pitfall 0() (No analysis can prove authorship) *Never claim that an analysis "demonstrates" authorship. The best that can be said is where the strength of the evidence points, compared to particular alternatives.*

Strength of evidence

If A is being ruled in as a likely author (or coauthor) of Q, the strength of the evidence will be that A's known documents K_A are particularly similar to Q, relative to known documents by other potential candidates and/or background authors representing the rest of the world. The metric for similarity needs to be calibrated, and that calibration shown, to show how similar is similar enough to determine likely authorship, and what the error rates, both false positive and false negative, are likely to be.

Pitfall 0() (Exhibit calibration on known documents) *Attribution measures for relevant documents with known authorship should be shown for calibration, to enable the jury to evaluate themselves the significance of your attribution results.*

When presenting quantitative results, particularly estimates of reliability of the analysis, it is important to do so in a way that avoids fallacies in interpretation. For example, suppose an analysis is performed to find the author of a questioned text Q from (say) a thousand candidates, and one candidate, X, matches Q such that the estimated probability of the match happening by chance is just one in a thousand. If that probability is presented as-is to a jury, their direct (and fallacious) conclusion may be that there is a 99.9% chance that X is the author of Q. This, however, is an instance of the prosecutor's fallacy (Thompson and Schumann, 1987). The actual probability of *some* candidate among the thousand reaching this level of match with Q by chance is $1-(1-\frac{1}{1000})^{1000} \approx 63.2\%$, thus, on its own, this is weak evidence for X's authorship indeed.

A less misleading presentation of evidential power of the attribution to X would be to present it in terms of Bayesian updating of the probability of the attribution given the new evidence (Berger, 2013), by giving the *Bayesian update factor* to the prior probability for X's authorship given the analysis:

$$\frac{P(\text{author=X}|\text{analysis})}{P(\text{author=X})}$$

This formulation directly shows how evidence adduced by the analysis should be combined with other available evidence to form a conclusion, and can be intuitively explained as an update to prior beliefs about the candidate.

Precise probability estimates are not always available, and such estimates often themselves rely on probabilistic assumptions. This can be most clearly expressed by giving a confidence interval, saying, for example, that the Bayes update factor is most likely between 1.5 and 6, so it is at least 50% more likely that X is the author given the analysis, and perhaps as much as six times more likely.

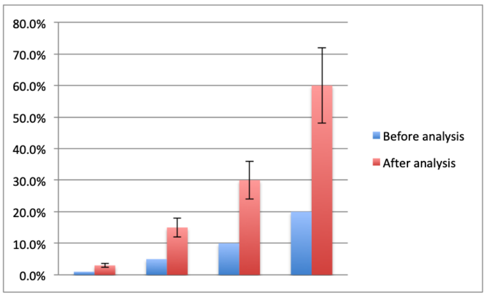

Figure 7. Example bar-graph showing belief updating for a Bayes update factor of $3 \pm 20\%$.

Visualizations can also be helpful. One way is to concretize these notions of probability—a bar graph (say) such as in Figure 7, showing how to update one's level of belief in authorship given the analysis. Another is to graphically show the similarity of Q to the known documents by different candidates (and comparison documents when relevant), as in Figures 3 and 4 above, for example. If this is done, care must be taken to address the potential pitfalls described in Section ‘Visual attribution, and to explain how this was done, of course.

Opening the black box

In addition, whatever attribution method is used should not be treated as a black box that simply takes documents as input and outputs attributions (with confidence scores). The box needs to be opened up to show what features are doing the attribution work, that is, which features Q shares more with K_A than with documents not by A. This helps to establish the trustworthiness of the method, as well as give more detail to the evidential claim of authorship.

The same principle applies when ruling out an author B. In such a case, the claim is supported by the similarity of B's known documents K_B to Q being notably less than would be expected if B was an author of Q. Here, opening up the box means showing features that are shared between different texts authored by B, but that are not shared with Q.

Pitfall 0() (Show the features that support the analysis) *Do not treat an analysis method as a black box, but show what textual features it bases its result on. This is necessary to establish the strength and the basis of the evidence for authorship being adduced.*

Examining the features used by the algorithms to classify authorship is also essential as a check on the entire text-processing pipeline. It is surprisingly easy, when dealing with diverse input formats, for text preprocessing to let through tokens that are not part of the actual text such as “page 3” or the name of an author in a page header, or the like. If such errors affect attribution, telltale features will show up, letting the analyst know to debug the text processing subsystem.

It should be noted that in the currently popular ‘deep learning’ techniques, as well as some others, it is not possible to directly determine what features are used to determine authorship. Indeed, explaining why a particular result was reached by such a model is, in general, an important unsolved research problem (Biran and Cotton, 2017; Samek *et al.*, 2017).

Concluding Thoughts

Computational authorship analysis methods can often allow reliable attribution even in cases where purely manual linguistic analysis is difficult or impossible, by statistical analysis of a very large number of subtle stylistic markers. However, establishing the reliability of a particular method for a particular case can be tricky, as it depends critically on many specifics of the case—one cannot simply rely on previous experience or experiments with the method. The list of potential pitfalls in this article should serve as guidelines for ensuring good methodology in developing computational authorship analyses, but the reader should always keep in mind that no such list can ever be complete. Expertise, experience, and careful human judgment must always be used and never supplanted by blind adherence to any predetermined methodology.

Summary List of Potential Pitfalls

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT press.

Argamon, S. and Koppel, M. (2010). The rest of the story: Finding meaning in stylistic variation. In *The Structure of Style*. Springer, Berlin, Heidelberg, 79–112.

Argamon, S., Koppel, M. and Avneri, G. (1998). Routing documents according to style. In *First International workshop on innovative information systems*, 85–92.

Argamon-Engelson, S., Koppel, M. and Avneri, G. (1998). Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, 1–4.

Arun, R., Suresh, V. and Madhavan, C. V. (2009). Stopword graphs and authorship attribution in text corpora. In *2009 IEEE International Conference on Semantic Computing*, 192–196: IEEE.

Baayen, H., Van Halteren, H. and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3), 121–132.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Berry, M. W. and Castellanos, M. (2004). Survey of text mining. *Computing Reviews*, 45(9), 548.

Binongo, J. N. G. (2003). Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2), 9–17.

Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, 8.

Brinegar, C. S. (1963). Mark twain and the quintus curtius snodgrass letters: A statistical test of authorship. *Journal of the American statistical Association*, 58(301), 85–96.

Burrows, J. (2004). Textual analysis.". *A companion to digital humanities*, 323–347.

Burrows, J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2), 91–109.

Cancedda, N., Gaussier, E., Goutte, C. and Renders, J.-M. (2003). Word-sequence kernels. *Journal of machine learning research*, 3(Feb), 1059–1082.

Chaski, C. E. (2005). Who's at the keyboard? authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1), 1–13.

Chung, C. and Pennebaker, J. W. (2007). The psychological functions of function words. *Social communication*, 1, 343–359.

Clement, R. and Sharp, D. (2003). Ngram and bayesian classification of documents for topic and authorship. *Literary and linguistic computing*, 18(4), 423–447.

Corney, M. W., Anderson, A. M., Mohay, G. M. and de Vel, O. (2001). Identifying the authors of suspect email. *Communications of the ACM*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.

Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities*, 32(1), 50–64.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar), 1289–1305.

Fucks, W. (1952). On mathematical analysis of style. *Biometrika*, 39(1/2), 122–129.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, 611: Association for Computational Linguistics.

Gentzkow, M. and Shapiro, J. (2013). *Congressional Record for 104th-109th Congresses: Text and Phrase Counts.* Rapport interne ICPSR33501-v2, University of Michigan, Ann Arbor, MI.

Glover, A. and Hirst, G. (1996). Detecting stylistic inconsistencies in collaborative writing. In *The new writing environment.* Springer, 147–168.

Graham, N., Hirst, G. and Marthi, B. (2005). Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4), 397–415.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3), 251–270.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3), 107–145.

Halteren, H. V. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1), 1.

Halvani, O., Winter, C. and Graner, L. (2018). Unary and binary classification approaches and their implications for authorship verification. *arXiv preprint arXiv:1901.00399.*

Han, J., Pei, J. and Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4), 405–417.

Holmes, D. I. and Forsyth, R. S. (1995). The federalist revisited: New directions in authorship attribution. *Literary and Linguistic computing*, 10(2), 111–127.

Honore, T. (1979). 'imperial'rescripts ad 193–305: Authorship and authenticity. *The Journal of Roman Studies*, 69, 51–64.

Hoover, D. L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, 18(4), 341–360.

Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, 77–86: Springer.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3), 264–323.

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*, volume 3, 255–264.

Kestemont, M., Luyckx, K., Daelemans, W. and Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, 93(3), 340–356.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, K. and Daelemans, W. (2016). Authorship verification with the ruzicka metric. In *Proceedings of Digital Humanities*, 246–249.

Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9(2), 119–124.

Koppel, M., Argamon, S. and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401–412.

Koppel, M. and Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, 72–80.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), 9–26.

Koppel, M., Schler, J. and Argamon, S. (2012a). Authorship attribution: What's easy and what's hard. *JL & Pol'y*, 21, 317.

Koppel, M., Schler, J., Argamon, S. and Messeri, E. (2006). Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 660: ACM.

Koppel, M., Schler, J., Argamon, S. and Winter, Y. (2012b). The "fundamental problem" of authorship attribution. *English Studies*, 93(3), 284–291.

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

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1), 178–187.

Kukushkina, O. V., Polikarpov, A. A. and Khmelev, D. V. (2001). Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2), 172–184.

Ledger, G. and Merriam, T. (1994). Shakespeare, fletcher, and the two noble kinsmen. *Literary and Linguistic Computing*, 9(3), 235–248.

Lipton, Z. C., Elkan, C. and Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 225–239: Springer.

Lodhi, H., Shawe-Taylor, J., Cristianini, N. and Watkins, C. J. (2001). Text classification using string kernels. In *Advances in neural information processing systems*, 563–569.

López-Escobedo, F., Solorzano-Soto, J. and Sierra Martínez, G. (2016). Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution. *Journal of Quantitative Linguistics*, 23(2), 154–176.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Moh'd A Mesleh, A. (2007). Chi square feature extraction based SVMs Arabic language text categorization system. *Journal of Computer Science*, 3(6), 430–435.

Mosteller, F. and Wallace, D. (1964). *Inference and disputed authorship: The Federalist.* Addison-Wesley.

Mostrous, A. (2013). JK Rowling unmasked as author of bestselling crime novel. *The Times (UK).*

Peng, F., Schuurmans, D. and Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4), 317–345.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, 2227–2237.

Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D. and Chanona-Hernández, L. (2017). Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, 21(3), 627–639.

Potha, N. and Stamatatos, E. (2017). An improved impostors method for authorship verification. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 138–144: Springer.

Quinlan, J. R. (2014). *C4. 5: Programs For Machine Learning*. Morgan Kaufman.

Rybicki, J., Hoover, D. and Kestemont, M. (2014). Collaborative authorship: Conrad, ford and rolling delta. *Literary and Linguistic Computing*, 29(3), 422–431.

Salton, G. and Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM (JACM)*, 15(1), 8–36.

Samek, W., Wiegand, T. and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sari, Y. and Stevenson, M. (2016). Exploring word embeddings and character n-grams for author clustering. In *CLEF (Working Notes)*, 984–991.

Seidman, S. (2013). Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*: Citeseer.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a), 542–547.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4), 471–495.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2), 193–214.

Stamatatos, E., Tschnuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B. and Potthast, M. (2016). Clustering by authorship within and across documents. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, 691–715.

Stover, J. A., Winter, Y., Koppel, M. and Kestemont, M. (2016). Computational authorship verification method attributes a new work to a major 2nd century african author. *Journal of the Association for Information Science and Technology*, 67(1), 239–242.

Thompson, W. C. and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials. *Law and Human Behavior*, 11(3), 167–187.

Uzuner, Ö., Katz, B. and Nahnsen, T. (2005). Using syntactic information to identify plagiarism. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, 37–44: Association for Computational Linguistics.

Vilariño, D., Pinto, D., Gómez, H., León, S. and Castillo, E. (2013). Lexical-syntactic and graph-based features for authorship verification. In *PAN workshop at CLEF*.

Xing, Z., Pei, J. and Keogh, E. (2010). A brief survey on sequence classification. *ACM Sigkdd Explorations Newsletter*, 12(1), 40–48.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165–193.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4), 363–390.

Zaïane, O. R., Foss, A., Lee, C.-H. and Wang, W. (2002). On data clustering analysis: Scalability, constraints, and validation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 28–39: Springer.

Zhao, Y., Zobel, J. and Vines, P. (2006). Using relative entropy for authorship attribution. In *Asia Information Retrieval Symposium*, 92–105: Springer.

Zheng, R., Li, J., Chen, H. and 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 and technology*, 57(3), 378–393.

Zimmer, B. (2013). The science that uncovered J.K. Rowling's literary hocus-pocus. *The Wall Street Journal*.

Zipf, G. (1935). *The Psychology of Language*. Houghton-Mifflin.