Journal of Law and Policy

Volume 21
Issue 2
SYMPOSIUM:
Authorship Attribution Workshop

Article 2

2013

Stylometry and Immigration: A Case Study

Patrick Juola, Ph.D.

STYLOMETRY AND IMMIGRATION: A CASE STUDY

*Patrick Juola**

INTRODUCTION

This paper describes "authorship attribution" as the process of inferring authorial identity from writing style and presents several classic studies as examples. This paper further explores a case of attribution "in the wild," so to speak, where there are a number of additional constraints and challenges. These challenges, fortunately, are not insurmountable. The background of the case, an asylum case in immigration court; responses to the challenges of the case; and the results of the analysis are discussed.

I. BACKGROUND

A. Stylometry and Authorship Attribution

Standard practice for stylometric investigations involves a detailed comparison of stylistic features culled from a training set of documents.[1] The questioned document is then compared

---

* Juola & Associates, pjuola@juolaassociates.com. This material is based upon work supported by the National Science Foundation under Grant No. OCI-1032683. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

[1] *See, e.g.*, Patrick Juola, *Authorship Attribution*, 1 FOUND. & TRENDS INFO. RETRIEVAL 233 (2006); Moshe Koppel & Jonathan Schler, *Computational Methods in Authorship Attribution*, 60 J. AM. SOC'Y INFO. SCI. & TECH. 9 (2009); Mathew L. Lockers & Daniel M. Witten, *A Comparative Study of Machine Learning Methods for Authorship Attribution*,

against the training set, typically using some form of classification or machine learning algorithm. Finally, an appropriate decision is reached in line with the experimental results.

A classic example of this form is the Mosteller-Wallace study of the *Federalist* papers,[2] a collection of eighteenth-century political documents describing and arguing for the (newly proposed) Constitution of the United States. These documents were originally published pseudonymously under the name Publius, but are now known (via traditional historical methods) to have been written by Alexander Hamilton, James Madison, and John Jay. Historians have come to consensus about the authorship of each of the eighty-five essays in the collection.

Mosteller and Wallace investigated the authorship question through the frequencies of individual words such as prepositions.[3] Careful analysis of known works by Hamilton and Madison, for example, show that they vary in the use of the word "by." For instance, Hamilton tended to use it about seven times per thousand words, rarely more often than eleven times per thousand, and never (in the samples studied) more than thirteen times per thousand words.[4] Madison, by contrast, used the word "by" most often in the range of eleven to thirteen times per thousand words, never less than five per thousand, and as much as nineteen per thousand.[5] Similar studies show that Hamilton used the word "to" more often than Madison, that Madison almost never used the word "upon," and so forth.[6]

We can therefore infer that a thousand-word document with seventeen tokens of "by" is more likely to be from Madison's pen than Hamilton's. If this document also contains relatively few "to's" and "upon's," our inference is strengthened. The

---

25 LITERARY & LINGUISTIC COMPUTING 215 (2010); Efstathios Stamatatos, *A Survey of Modern Authorship Attribution Methods*, 60 J. AM. SOC'Y INFO. SCI. & TECH. 538 (2009).

[2] *See generally* FREDERICK MOSTELLER & DAVID L. WALLACE, INFERENCE AND DISPUTED AUTHORSHIP: THE FEDERALIST (1964).

[3] *Id.* at 29 tbl.2.3–3.

[4] *Id.* at 17 tbl.2.1–1.

[5] *Id.*

[6] *Id.*

notion of "more likely," with respect to identifying authorship, can be formalized using statistics (particularly Bayes' theorem)[7] to yield a precise odds ratio. With enough data, the odds ratio can achieve practical certainty. For example, Madison is millions of times more likely to have written Federalist Paper 51 than Hamilton.[8]

A similar example is the study by Binongo of the fifteenth *Oz* book, *The Royal Book of Oz*.[9] The original *Wonderful Wizard of Oz* was of course written by L. Frank Baum, as were the second through fourteenth books in that series. When Baum died, the publisher found another writer, Ruth Plumly Thompson, to serve as Baum's successor, working from "notes and a fragmentary draft"[10] for the fifteenth book and then writing eighteen more original *Oz* books. The question is whether a substantial "draft" of the fifteenth book ever existed, or whether the *Royal Book* was also largely Thompson's work.

Similarly to the Mosteller-Wallace study, Binongo chose to study lexical items, analyzing the relative frequency of the fifty most common words in the combined *Oz* series, a set containing words like "the," "and," "with," "into," and so forth.[11] Using a dimensionality reduction technique called Principal Component Analysis ("PCA"), he combined the variation among these fifty words down to two dimensions and plotted each work on a two-dimensional graph.[12] The results were clear and compelling; there were distinct clouds representing Baum's and Thompson's respective work, with a notable separation between them (in Binongo's words, a "stylistic gulf").[13] The *Royal Book* fell squarely on Thompson's side of the fence, "reveal[ing] that the

---

[7] Here and elsewhere, we omit the detailed mathematical description for clarity and brevity.

[8] *Id*. at 211 tbl.5.5–2, 263.

[9] José Nilo G. Binongo, *Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution*, 16 CHANCE, no. 2, 2003 at 9.

[10] RAYLYN MOORE, WONDERFUL WIZARD, MARVELOUS LAND 89 (1974).

[11] Binongo, *supra* note 9, at 11–12.

[12] *Id.* at 12.

[13] *Id.* at 15.

writing style in the 15th Book of *Oz* is more compatible with Thompson's than Baum's."[14]

There are notable differences between these studies. Mosteller and Wallace studied a variety of possible features before settling on a hand-picked set of thirty words (including some rather rare words such as "direction") chosen for their discriminative abilities in this specific study.[15] Binongo, on the other hand, simply used the fifty most common words in the corpus.[16] In this volume, Stamatatos argues for the use not of words but of character sequences;[17] we have argued elsewhere for the use both of character sequences and word sequences.[18] Mosteller and Wallace used a form of Bayesian statistical analysis,[19] Binongo used PCA,[20] Stamatatos uses a third technique called "support vector machines,"[21] and we have argued elsewhere for similarity-based nearest neighbor methods.[22]

More striking than the differences, however, are the similarities in both the Mosteller-Wallace and Binongo studies:

- the set of candidate authors was limited to only a small and clearly defined group of people;
- all candidate authors had an extensive body of unquestioned work to compare;

---

[14] *Id.* at 16.

[15] MOSTELLER & WALLACE, *supra* note 2, at 67–68.

[16] Binongo, *supra* note 9, at 11–12.

[17] *See generally* Efsathios Stamatatos, *On the Robustness of Authorship Attribution Based on Character* N-*Gram Features*, 21 J.L. & POL'Y 421 (2013).

[18] *See generally* Patrick Juola & Darren Vescovi, *Analyzing Stylometric Approaches to Author Obfuscation*, *in* ADVANCES IN DIGITAL FORENSICS VII, at 115, 115–25 (Gilbert Peterson & Sujeet Shenoi eds., 2011).

[19] *See generally* MOSTELLER & WALLACE, *supra* note 2.

[20] Binongo, *supra* note 9, at 12–17.

[21] Stamatatos, *supra* note 17, at 431.

[22] John Noecker, Jr. & Patrick Juola, *Cosine Distance Nearest-Neighbor Classification for Authorship Attribution*, PROC. DIGITAL HUMAN., 2009, at 208.

- this body of work was huge (in the *Oz* study, more than a dozen novels each), large enough to provide statistical confidence; and
- the body of work was similar to the disputed document in style, topic, and genre, and thus provided a representative sample.[23] This is key because many of the factors that separate individuals also vary systematically between types of writing. Passive writing is very common in technical prose, for example, but uncommon in conversation or narrative.[24]

One might suspect that the choice of topics and works to study was in part driven by these considerations. Unfortunately, many cases of practical interest (especially in the court system) do not have these attributes, as will be seen in Part II.

B. JGAAP

In light of the differences among possible analyses, an obvious question is "which method works best?" To address this question, the Evaluating Variations in Language Laboratory at Duquesne University has developed a modular system for the development and comparative testing of authorship attribution methods.[25] This system, Java Graphical Authorship Attribution Program ("JGAAP"), provides a large number of interchangeable analysis modules to handle different aspects of the analysis pipeline such as document preprocessing, feature selection, and analysis/visualization. Taking combinatorics into account, the number of different ways to analyze a set of documents ranges in the millions and can be expanded by the inventive user with a moderate knowledge of computer programming.

---

[23] MOSTELLER & WALLACE, *supra* note 2, at 2–3; Binongo, *supra* note 9, at 9–10.

[24] DOUGLAS BIBER, VARIATION ACROSS SPEECH AND LANGUAGE 50 (1988).

[25] Juola, *supra* note 1; Patrick Juola et al., *JGAAP 4.0—A Revised Authorship Attribution Tool*, PROC. DIGITAL HUMAN., 2009, at 357.

II. A CASE STUDY

To illustrate the issues and complications that can arise in "the real world," we present the following as a case study in the application of authorship attribution in actual forensic practice. All identifying details have been changed to protect the privacy (and possible physical well-being) of the individuals involved.

A. Statement of the Case

Bilbo Baggins, a native of Mordor, was facing immigration procedures that might have led to his removal from the United States. He claimed in immigration court that deportation was inappropriate and sought asylum because he was a noted and published activist against the Mordor government and he feared negative consequences if forcibly repatriated. As evidence for this claim, he offered a number of articles he had written for an Elvish-language newspaper, as well as a set of newer (antigovernment) articles he claimed to have written but that had been published anonymously while outside Mordor. Juola & Associates was asked by Baggins' counsel to analyze these articles. The basic theory of the case was that if Baggins had, in fact, written the newer articles (the older articles were unquestioned, as they had been published under his name), and if that fact could be demonstrated, that would establish that his fears were well founded.

Superficially, this appears to be an ordinary questioned-documents case, but there are a few twists. We started by rejecting "traditional" document forensics, handwriting analysis and such, as there are no original documents to study. All documents had been submitted to newspapers and subjected to editorial review and publication; the older documents were in the form of photocopies of printed clippings, while the new documents were born-digital web pages that had no originals. All that was available was the content of the documents, suggesting a need for authorship analysis as defined above.

At the same time, there was no clearly defined set of candidate authors; either Baggins wrote the questioned documents or "someone else" did, and all we know about this

"someone else" is that they had access to the Internet. Additionally, the set of documents available was rather small: a dozen newspaper articles each in the known and questioned sets. The documents were also in Elvish, an understudied language with little computational support available.

The last point is probably the least important, as JGAAP provides a relatively language-agnostic method of analysis. Certainly, the idea of "fifty most common words" is computationally tractable in any language with a clear notion of a word (such as a language like English, German, Russian, or Spanish where spaces separate words). Furthermore, previous research has shown that there is a high cross-linguistic correlation in performance of authorship attribution methods or, in other words, that in the absence of compelling counterinformation, methods that are known to perform well in English are likely to perform well in other unstudied languages.[26] But structuring the problem as a verification instead of classification problem forced us to use a somewhat nonstandard approach. In a typical classification problem, there are a number of possible answers, one "correct" answer and a number of "distractor" answers. (In an authorship context, Marlowe and Kyd could be distractors for a play we believe to be written by Shakespeare; in the context of criminal investigation, all of the suspects except for the actual guilty party are *de facto* distractors.) By contrast, in a verification problem, we have only one "suspect" but need to evaluate whether the evidence is sufficient to tie him to the acts in question.

B. Materials and Methods

Baggins himself supplied us with ten copies of newspaper articles published under his name approximately ten years before the date of the case; these articles comprised a set of known documents. These documents (photocopies of clippings) were hand-transcribed by Elvish-speaking typists into a machine-

---

[26] Patrick Juola, *Cross-Linguistic Transference of Authorship Attribution, or Why English-Only Prototypes Are Acceptable*, PROC. DIGITAL HUMAN., 2009, at 162.

readable corpus. In addition, he supplied us with eleven web page images from a recent news site, published anonymously, as the set of questioned documents.[27]

The JGAAP software package provided the necessary technology for this text analysis. All relevant files were preprocessed to convert them into plain text (Unicode) format. All case distinctions were neutralized, and all whitespace (interword spacing, line breaks, paragraphing, etc.) was normalized to avoid any spurious findings of dissimilarity caused by simple formatting and editing issues. (Again, JGAAP has a button for this kind of preprocessing, and in fact no manual processing was required at all for this analysis.) All documents were converted into word trigrams (phrases of three adjacent words, as in the English phrase "in the English"), a unit of processing known to give good results in authorship queries.[28]

To establish with reasonable certainty that Baggins had or had not written the document, it was necessary for us to create our own distractor set, which we did by gathering a collection of Elvish-language newspaper articles on political issues from another online newspaper. This corpus consisted of 160 news articles by five different named authors, none of whom were Baggins. This provided us with five separate comparison "baseline document corpora" each containing at least thirty articles known to be authored by a distractor author.

The word trigram distributions of the ten documents in the known document set were averaged to produce a central or typical example of Baggins' writings. Each individual document in the questioned corpus as well as the five baseline corpora was individually compared against this "typical" Baggins style to determine a stylistic distance—a numerical measure of stylistic similarity. Two identical documents would be at distance zero, and, in general, the smaller the distance (the "closer" the document pair), the more likely two documents were to share

---

[27] Of these eleven documents, one was in English and unsuitable for study, so the actual questioned documents comprised ten web pages from which text was extracted. No typists were needed to extract text from these pages as they were in standard HTML; JGAAP will in fact do that automatically.

[28] *See* Juola, *supra* note 1, at 265–66.

authorship. These distances were averaged to produce a per-author average distance from the known documents.

1. Preliminary Results

The preliminary results can be summarized in Table 1.

Table 1: Preliminary results using cosine distance

Subcorpus	Distance to KD (Known Document Set)
BD-1 (Baseline Document Set 1)	0.9437975
BD-2	0.9517967
BD-3	0.9576155
BD-4	0.9530338
BD-5	0.9534134
QD (Questioned Document Set)	*0.8840330*

These results provided preliminary evidence in favor of Baggins's claim; his style is notably closer to that of the questioned documents than it is to other, similar writers. But can we turn this preliminary observation into quantifiable probability judgments? And if so, how compelling are these probabilities? Unfortunately, standard parametric tests (such as *t*-tests) did not help. Interdocument variation (not shown here) dominated the small differences between groups, and the difference in distance was not significant, in a technical sense.

However, there is still an argument to be made here using a non-parametric framework. Assuming that the questioned documents were written by a seventh author outside the set, we have no *a priori* reason to assume that this seventh author would be particularly similar or dissimilar to Baggins. Thus, the probability of this seventh author being the closest to Baggins (as we found in this study) is one in six, approximately 16.7%. Nonparametrically, we can reject this idea (that the documents were written by a seventh author) at the *p*-value of 0.167. This confirms our intuitions that the results support his claim and provide (weak) numerical support, but enough, perhaps, to overcome a "balance of probabilities" burden of proof in a civil case.

2. Ensemble Methods and Mixture of Experts

We can, however, (potentially) improve upon these results using ensemble methods.[29] The basic idea is the one behind getting a second opinion: if two (or more) independent experts agree in their analysis, our confidence in that result is increased.[30] This can be formalized using probability theory: if the chance of an expert being right is x, the chance of her being wrong is therefore $(1 - x)$. The chance of two such experts independently being wrong is $(1 - x)(1 - x)$ or $(1 - x)^2$, and in general, the chance of k experts all being wrong is $(1 - x)^k$. For example, if experts in general are right 90% of the time, the chance of one expert being wrong is 0.1 or 10%. The chance of two both being wrong is 0.01 or 1%, and for three experts, 0.001 or 0.1%. In this case, the chance of our analysis being wrong, from above, is 16.7%. If a similar analysis yields the same result, the chance of them both being wrong is a mere 0.167 times 0.167, one chance in thirty-six, or about 2.78%.

We therefore performed these distance comparisons twice, using two different distance formulae and hence two different analyses. The first analysis was performed using normalized dot product or cosine distance,[31] in which the frequency of each individual word trigram is taken into account. The second was done with Jaccard or intersection distance[32] between the sets of word trigrams, which does not take into account frequency but simply measures whether or not a particular author used a particular three-word phrase at any point in the samples.

---

[29] *See generally* Patrick Juola, Authorship Attribution: What Mixture-of-Experts Says We Don't Yet Know, Conference Presentation at AACL 2008 Am. Ass'n for Corpus Linguistics (Mar. 13, 2008), *available at* http://corpus.byu.edu/aacl2008/ppt/115.ppt (discussing various authorship attribution studies).

[30] *See id*.

[31] Noecker & Juola, *supra* note 22.

[32] Tanguy Urvoy et al., *Tracking Web Spam with Hidden Style Similarity*, PROCEEDINGS OF AIRWEB'06 (Aug. 10, 2006), *available at* http://airweb.cse.lehigh.edu/2006/urvoy.pdf.

As hoped, the results of the second experiment (Table 2) confirmed the first:

Table 2: Results using Jaccard/intersection distance

Subcorpus	Distance
BD-1	0.806731
BD-2	0.739381
BD-3	0.852844
BD-4	0.747444
BD-5	0.777530
QD	*0.735449*

An alert reader will see the card that has just been palmed. Our argument for ensemble methods hinges on an assumption of independence, an assumption that is almost certainly untrue. A document in another language or *a fortiori* another alphabet/writing system will share almost no words or phrases, and hence be strongly different. But within a set of documents of more limited scope—in this case, sharing language, genre, and even general topic—we can argue that a certain amount of independence can be expected. From a purely empirical standpoint, the fact that the baseline distractor authors are ordered differently in the two experiments (e.g., #2 is the closest in Jaccard distance, followed by #4; #1 is first in cosine distance) suggests that these analyses are to a large degree independent. From a theoretical standpoint, Jaccard distance is sensitive only to the distribution of rare features (word trigrams that one author does not use at all), while cosine distance is more sensitive to more common features (as they have greater frequency variance). But in light of the fact that we have no formal measure of the degree of independence, we can, strictly speaking, only say that the chance of this result occurring is no more than 16.7% and could be as small as 2.78%.

C. Why Stop Here?

JGAAP provides many more than two possible methods. However, we provided no further analysis for this particular case. In theory, we could have used ten methods, and if they all

showed the same result, the odds of a false positive would have been approximately 0.000000165% or one in just over sixty million. However, we would also have run a risk of significantly weakening the case if the analyses did not turn out the way Baggins hoped. The additional costs and risks were, in the opinion of Baggins's counsel, not worth the marginal increase in confidence. This, of course, is a tactical and legal decision based in part on the type of case and the strength of the other evidence available.

CONCLUSION

Authorship analysis in the field can pose substantially different challenges than in the lab. The Baggins case presented several unusual aspects in stylistic investigations; the standard stylometric analysis paradigm selects among others rather than giving a simple yes/no answer. Using nonparametric rank order statistics and an ad-hoc set of distractor authors, we could still get an answer and validate it statistically.

Oh, and Bilbo Baggins himself? The judge permitted him to remain in the United States.