The Rowling Protocol, Steven Bannon, and Rogue POTUS Staff: a Study in Computational Authorship Attribution

Patrick Juola

Juola and Associates, USA

Abstract. *A key step forward in the professionalization of forensic science is the development of standards of practice and protocols. Based on his analysis of the Rowling case, Juola (2015) proposed a systematic protocol for authorship verification. We present both a theoretical and an empirical analysis of the accuracy of this protocol. We further present a demonstration of this analysis in terms of a high-profile case of political activism. We show that this protocol produces accurate and understandable analyses of the likelihood of common authorship.*

Keywords: *Authorship attribution, standards, protocols, independence, statistical analysis.*

Resumo. *Um passo fundamental na profissionalização da ciência forense é o desenvolvimento de normas e protocolos de prática. Com base na sua análise do caso Rowling, Juola (2015) propôs um protocolo sistemático para verificação de autoria. Neste trabalho, apresentamos, quer uma análise teórica, quer uma análise empírica da precisão deste protocolo. Procedemos, ainda, a uma demonstração dessa análise em termos de um caso importante de ativismo político, mostrando que este protocolo permite produzir análises precisas e abrangentes da possibilidade de autoria comum.*

Palavras-chave: *Atribuição de autoria, normas, protocolos, independência, análise estatística.*

Introduction

The authorship of documents is a key question in many legal cases (both fictional and real), as a skim of many of Agatha Christie's mysteries will show.[1] Handwritten, or even typed, documents can be validated by physical marks of the production process.[2] Electronic documents (Chaski, 2005; Juola, 2006b, 2007) bring their own set of issues, as handwriting cannot be used to validate the documents, and one ASCII 'A' is bit-for-bit identical to any other. Stylometry, the study of individual writing style (Holmes, 1994; Grieve, 2005; Juola, 2006a; Stamatatos, 2009), can be so used. In the case of *Ceglia v. Zuckerberg, et al.* (McMenamin, 2011), for example, ownership of a significant part of Facebook depended in part on the validity of an emailed agreement between the two parties. By looking at the writing style, including aspects such as word choice, catch phrases, punctuation and spelling, McMenamin was able to find the *linguistic* marks of the producer/author.

In this paper, we describe a specific type of authorship attribution problem, that of authorship verification, with some examples. We then describe a formal protocol based on the analytic techniques used to identify J.K. Rowling, the author of the *Harry Potter* novels, as the author of Robert Galbraith's *A Cuckoo's Calling* as well. We show how this protocol can be used to address a general and common class of problems and present a software system (Envelope) that implements this protocol in a simple and easy-to-use way. We present both a theoretical and an empirical analysis of the accuracy of this protocol. Finally, we present a demonstration of this analysis in terms of a high-profile case of political activism — that of "Rogue POTUS Staff," a political activist who ostensibly posts inside information about the Trump White House.

Background

Authorship Analysis

Language is among the most individualized activities people engage in. For this reason, much can be learned about a person by looking at his or her writings. An easy example is distinguishing between different regional groups. A Commonwealth English speaker/writer can easily be spotted by her use of "lorry" instead of "truck," spelling "labor" with a 'u,' and less obviously by grammatical constructions such as "could do" or "in hospital." These insights can be extended to questions of authorship without regard to handwriting. The basic theory of traditional stylistics is fairly simple. As McMenamin (2011) describes it,

> At any given moment, a writer picks and chooses just those elements of language that will best communicate what he/she wants to say. The writer's "choice" of available alternate forms is often determined by external conditions and then becomes the unconscious result of habitually using one form instead of another. Individuality in writing style results from a given writer's own unique set of habitual linguistic choices.

Coulthard's (2013) description is also apt:

> The underlying linguistic theory is that all speakers/writers of a given language have their own personal form of that language, technically labeled an idiolect. A speaker/writer's idiolect will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in the written and spoken communications they produce. For example, in the case of vocabulary, every speaker/writer has a very large learned and stored set of words built up over many years. Such sets may differ slightly or considerably from the word sets that all other speakers/writers have similarly built up, in terms both of stored individual items in their passive vocabulary and, more importantly, in terms of their preferences for selecting and then combining these individual items in the production of texts.

These choices express themselves in a number of ways. In an expert witness report, McMenamin (2011) analyzed eleven different and distinct "features" of the writing in sets of both known (undisputed) and disputed emails. One feature, for example, hinged on the spelling of the word "cannot", and in particular whether it was written as one word ("cannot") or as two ("can not"). Another feature was the use of the single word "sorry" as a sentence opener (as opposed, for example, to "I'm sorry"). Coulthard (2013) similarly discussed (among other features) the use of the specific phrase "disgruntled employees." (Why "disgruntled" and not one of its myriad synonyms?) In both cases, significant differences in these features can be held as evidence of differences in authorship.

The legal implications of this type of evidence should be apparent. Chaski (2005) provides a dramatic example in a suspicious death. When a person's body is found next to a typed suicide note – in this case, it was typed into a computer, but it could just as easily have been typed on an actual typewriter – the specific machine used to produce the note is not in question. If the machine is shared (for example, used by several roommates in a house), fingerprint analysis may not reveal much except that the fingerprints of several people can be found on it. By analyzing the writing, Chaski was able to establish that the decedent was probably not the author of the suicide note, turning the apparent suicide into a murder, and enabling the police to eventually catch the perpetrator. But even without the drama, any case involving "anonymous" writing (such as poison-pen letters or emails) would be aided by the ability to find the actual author.

As typically defined (e.g. Mosteller and Wallace, 1964; Binongo, 2003), authorship "attribution" involves selecting the most likely author from a small but finite set of candidate authors. In the real world, cases often involve simply determining whether or not a single specific author wrote a single specific document, where the alternative answer is that the actual author is simply "someone else." Examples of this include Juola (2013c), Brooks and Flyn (2013) and Collins (2013). Authorship "verification," as this subproblem is called, is a more difficult task because there is no obvious way to assess the properties of millions or billions of potential authors who are not part of the document set. While a specific misspelling of "toutch" (Wellman, 1936) may be idiosyncratic to one person, that does not exclude the possibility of other, unrelated people also using that spelling.

Computational text analysis

Computer-based stylometry applies the same general theory, but with a few major differences. The basic assumption that people make individual choices about language still holds, but instead of ad hoc features selected by examination of the specific documents, the analysts use more general feature sets that apply across the spectrum of problems (Binongo, 2003; Burrows, 1989; Hoover, 2004; Koppel *et al.*, 2009; Juola, 2006a; Juola *et al.*, 2013; Mikros and Perifanos, 2013). Examples of feature sets include word use, character clusters, and so forth. Using these feature sets or others (Rudman, 1998), the features present in a document are automatically identified, gathered into collections of feature representations (such as vector spaces), and then classified using ordinary classification methods (Jockers and Witten, 2010; Juola, 2006a, 2012a; Koppel *et al.*, 2009; Noecker Jr. and Juola, 2009) to establish the most likely author.

Binongo (2003) provides a clear example of this. For background: the first fourteen books of the *Oz* series were written by L. Frank Baum before his death. After his death, the publisher approached another author, Ruth Plumly Thompson, to finish the then-incomplete (and arguably nonexistent) fifteenth book, *The Royal Book of Oz*. In his study of the authorship of the *Oz* books, Binongo collected the frequencies of the fifty most frequent words in English from the books of undisputed authorship (his feature set). He applied principal component analysis (his classification method) to obtain a data visualization of the stylistic differences, then showed that the disputed 15th book clearly lay in a stylistic space corresponding to only one candidate author, Thompson. This would clearly be highly relevant evidence if the authorship (perhaps for copyright reasons) were being disputed in court.

From a legal standpoint, there are three key issues with this technology. The first, admissibility, has been addressed in detail elsewhere (Chaski, 2013; Coulthard, 2013; Juola, 2014, 2015) but is closely tied to the second issue, the scientific validity of the technology itself. Numerous surveys (Grieve, 2005; Jockers and Witten, 2010; Juola, 2006a; Koppel *et al.*, 2009; Stamatatos, 2009) and TREC-style conferences (Juola, 2004, 2012b; Juola and Stamatatos, 2013; Stamatatos *et al.*, 2014) have shown that authorship can be determined with high accuracy (typically 80% or better) using realistically-sized samples. Large-scale studies (Juola, 2012a; Vescovi, 2011) have confirmed that there are often many different "best practices" that perform well based on different features. This allows for ordinary data fusion techniques such as mixture-of-experts to boost accuracy rates to practical levels.

Types of Authorship Problem

The usefulness of the above technology has been demonstrated in actual disputes. Chase's murder case (Chaski, 2005) has already been mentioned. For example, Collins (2013) used a mixture of experts to validate a newly discovered short story by Edgar Allan Poe, and Juola (Brooks and Flyn, 2013; Brooks, 2013; Juola, 2013b) used a similar method to identify J.K. Rowling (the author of the *Harry Potter* series) as the author of the pseudonymously published detective novel *A Cuckoo's Calling*. In a legal context, Juola (2013c) was able to verify the authorship of anonymous newspaper columns in support of an asylum claim in a US immigration court. Finally, Grant (2013) was able to perform a similar analysis without the aid of computers and determine the identity of a murderer.

A detailed examination of these cases, however, reveals key differences among them. As typically defined (e.g. Mosteller and Wallace, 1964; Binongo, 2003; Grant, 2013), authorship "attribution" involves selecting the most likely author from a small but finite set of candidate authors. In the case studied by Grant, there were, realistically, only two actors of interest. This may be typical of crimes-of-person, where someone needs to be present to commit the crime, and only a small group of candidates (those, for example, who had physical access to the crime scene) need to be considered. In Grant's case, it is understood that without sophisticated technological spoofing, only a person with physical access to a cell phone can use that phone to send text messages. In this case, the task of the analyst is to assess the comparative similarity/likelihood of each possible candidate author for the documents in question. This so-called "closed class" task, then, does not need to consider "none of the above" as a serious contender, and is the simplest and easiest formulation of the problem of authorship attribution.

By contrast, cases often involve simply determining whether or not a single specific author wrote a single specific document, where the alternative answer is that the actual author is simply "someone else." This may be typical of the analysis of published documents, as the questioned manuscript might have been written from literally anywhere in the world. Similar issues arise with the analysis of electronically transmitted documents such as web pages and emails. Even an obvious idiosyncrasy may be shared with someone else thousands of kilometers away.

Work has been done in authorship verification such as the "imposter" method (Koppel and Winter, 2014), but their work is hard to use and to understand. First, their use of huge numbers (tens of thousands) of distractor authors may provide statistical power, but makes the task of data collection arduous and expensive. Second, the authors focus on one analysis repeated in a rather untransparent way, an analysis focusing on what a reviewer of this paper has correctly identified as perhaps the least understandable analysis method we ourselves use. Thirdly, their protocol relies on an undescribed ad-hoc cutoff threshold and does not lend itself well to intuitive odds judgements about what people are typically interested in, the actual likelihood that a given author wrote a specific work—the sort of intuitive presentation that is easily understandable to a judge or jury.

A Proposed Protocol

Juola (2014, 2015) presented a formal protocol for authorship verification and showed how it could be applied to several separate authorship disputes.

Key elements of this proposed protocol are:

- Suitable data for analysis, including an ad hoc set of distractor authors believed not to be connected to the case;
- A set of independent analysis methods that have been found to perform well on similar tasks;
- A predefined data fusion framework amenable to formal statistical analysis, so that the likelihood of error can be assessed mathematically;
- A predefined interpretation of the statistical results in human-understandable terms.

As an illustration, we here describe its application in the immigration case reported in Juola (2013c). The background is relatively straightforward; an immigrant, whose name and other identifying details have been changed for personal safety, was applying for asylum in the United States. Bilbo Baggins, as we have renamed him, was originally a citizen of Mordor, a successful journalist under his own name, but also an anonymous online critic of the Mordor government. He feared persecution for his political activities, but, of course, the political activities had not been performed openly under his own name. Could the author of these anonymous articles be linked with the articles published under Mr. Baggins' own name?

I was able to collect a set of 160 news articles by five different named authors, none of whom was Baggins. This, in turn, provided me with five separate "baseline document corpora" against which to compare the anonymous writings. Using the JGAAP software platform (Juola *et al.*, 2006, 2009),[3] stylistic "distances" (Noecker Jr. and Juola, 2009) were calculated between the anonymous documents and each of the candidate authors as well as Baggins' undisputed writings. These distances had been shown in prior work to be able to select (with relatively high accuracy) the correct author out of a set of candidate authors based on the principle that the smallest distance represents the most similar and therefore most likely author.

Should Baggins be the actual author of the anonymous articles, then, one would expect Baggins to be the closest author by distance measurement. In the event that, by chance, I had selected the actual author of the anonymous articles as one of the distractors, we would expect Baggins not to be the closest, but that person instead. Should the actual author be a seventh person, not in the set (which is more likely than accidentally finding the actual author as a distractor), one would have no reason *a priori* to believe that Baggins is particularly likely to write using a similar style, so there is roughly one chance in six that he would be the most similar author.

Table 1. Potential outcomes of Baggins article analysis

Case 1		Case 2	
Position	Author	Position	Author
1	*Baggins*	1	Distractor 1
2	Distractor 1	2	Distractor 2
3	Distractor 2	3	Distractor 3
4	Distractor 3	4	*Baggins*
5	Distractor 4	5	Distractor 4
6	Distractor 5	6	Distractor 5
Probably Baggins		Probably not Baggins	

One can therefore describe the potential outcomes of this analysis in table 1. Case 1 describes a situation where Baggins is chosen as the closest and most likely author; in the event that Baggins is, in fact, the actual author, we would consider this the most probable case. Case 2 describes a situation where Baggins is not observed to be the closest author, which we would consider to be the most probable case in the event either that the true author had inadvertently been among the distractors (a highly unlikely coincidence) or that the actual author was not in our data set of known authors.

Thus, with high probability, we expect case 2 if Baggins is not the actual author, and case 1 only if either Baggins is the true author, or the unlikely event that the true author is someone who writes with a similar style to Baggins. If Baggins is not in the data set, then we would expect, by chance, case 2 to arise roughly $\frac{5}{6}$ of the time. Thus, if one treats "none-of-the-above" as the null hypothesis, we would have an effective p-value (for rejecting the null hypothesis) of 0.167 in case 1.

If greater confidence is desired, one can, however, improve upon these results using ensemble methods. Juola (2013c) wrote:

> 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 (Juola, 2008). 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 [the Baggins analysis], 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.

Repeating this test (as Juola did) with a second, independent analysis (and getting the same result) would give an effective p-value of roughly 0.0278, enough to reject the null hypothesis on a standard one-tailed cutoff of 0.05. Similarly, repeating this test a third time (as Juola did not), and again getting the same result would get an effective p-value of 0.00463. In rejecting the null hypothesis, he would thus have demonstrated evidence tending to show that it is highly unlikely that anyone other than Baggins wrote the disputed documents, and hence that Baggins is the true author of the questioned documents. This was, in fact, the outcome of the case, and Bilbo Baggins was permitted to remain in the United States.

Of course, there is no reason to restrict oneself to only two tests, and similarly no reason to restrict oneself to exactly five distractor authors. Similarly, a simple "first/not-first" cutoff may be impractical, but this test lends itself well to statistical tests such as Fisher's exact test (Fisher, 1971) applied to computed scores such as the rank sum of Baggins' positions. (The case above, for example, would be equivalent – under the null hypothesis – of rolling two dice to determine Baggins' score; the reader can confirm for himself that there is one chance in 36 of getting a rank sum of 2, and three chances in 36, less than one in ten, of getting a rank sum of 3 or smaller.)

As discussed in the following section, we have extended this proposed protocol to permit more accurate probability assessments by using more tests and a larger number of distractor authors. We have implemented this protocol in a software-as-a-service (SaaS) platform, named Envelope (Juola, 2016) to provide low-cost, high-accuracy resolution of authorship disputes.

Envelope, a SaaS Platform for Authorship Verification

Design and implementation

Envelope, in its current version, focuses on a specific (and relatively common) type of disputed document, electronic mail (Chaski, 2005; Coulthard, 2013; McMenamin, 2011) written in English. The system is presented with client-supplied copies of the disputed email(s) as well as samples known to have been written by the purported author. These documents are compared against a set of distractor authors (currently a set of ten gender-balanced authors extracted from the Enron corpus (Klimt and Yang, 2004)) and rank-ordered for similarity along five human-understandable features that have been shown to work well in large-scale testing (Juola, 2012a; Vescovi, 2011). The five measured dimensions are as follows:

- Authorial Vocabulary (Vocabulary overlap): Words are, of course, what a work is fundamentally all about. A crime novel is usually about a dead body and how people deal with the problem it poses, while a romance novel is about a small group of people and their feelings for each other. Even emails differ in word choices as discussed above (Coulthard, 2013; McMenamin, 2011; Juola, 2013a). Authorial vocabulary is also one of the best ways to tell individual writers apart, by looking at the choices they make, not only in the concepts they try to express, but the specific words they use to create their own individual expression. The degree of shared vocabulary is thus a key authorial indicator. This was calculated using a modified Jaccard distance that does not take into account frequency distribution, and hence is sensitive only to the question of whether the author does or does not use a particular word token.

- Expressive Complexity (Word length): One key attribute of authors is, on the one hand, their complexity, and on the other, their readability. A precise author who uses the exact specific word to every event – "that's not a car, that's a Cadillac; that's not a cat, but a tabby" – will more or less be forced to use rarer words. These rarer words, by their very nature, are typically longer (Zipf, 1949). A large and complex vocabulary will naturally be reflected in longer words, producing a very distinctive style of writing. By tracking the distribution of word lengths (n.b.: the percentage of words with various lengths, not just the average word length, which is known not to perform well), we can assess the expressive complexity of a given author.
- Character n-grams: In addition to comparing words directly, scholarship has shown (Cavnar and Trenkle, 1994; Mikros and Perifanos, 2013; Noecker Jr. and Juola, 2009; Stamatatos, 2013) that comparison of the frequency spectra of character clusters (for example, four adjacent letters, whether as part of a word like "eXAMPle" or across two words as in "iN␣THe") is a useful way to assess document similarity. This allows matching of similar but not identical words, such as different forms of the same stem or words with similar affixes, and even preferred combinations of words. We used normalized cosine distance (Noecker Jr. and Juola, 2009) to compare the frequencies of various character n-grams.
- Function words: One of the most telling and oft-studied aspects of an individual writer is their use of function words (Binongo, 2003; Burrows, 1989; Hoover, 2004), the simple, short, common, and almost meaningless words that form a substantial fraction of English writing. These words thus provide a good indication of the tone of the writing and the specific types of relationship expressed throughout the manuscript. We evaluated function words by restricting our attention to the fifty most frequent words using normalized cosine distance as above.
- Punctuation: Although not necessarily linguistically interesting, and often the choice of editor instead of author, punctuation offers an insight into social conventions that have little effect on the meaning of the text. Because they have little effect, they are often freely variable between authors. For example, an author's decision to use an Oxford comma, their choice of marking extraneous material (for example, with commas, parentheses, or brackets), the way they split sentences with semicolons, periods, or comma splices, and even whether punctuation is put inside or outside quotation marks, do not change the meaning. In unedited documents (such as email), they therefore provide a strongly topic-independent cue to authorship that is not directly related to the other dimensions. (See Grant, 2013; McMenamin, 2011 for some non-computational examples.)

Numerical analysis of the Envelope protocol

Along each document, the eleven possible authors (ten implausible distractor authors plus one plausible suspect) are ranked from #1 (most similar/likely) to #11. The rank sum of the purported author across all dimensions is calculated and used to fuse the different analyses. For example, if the purported author scored as the most similar author on all five dimensions (the most compelling possible result), the rank sum would be five. The system then uses Fisher's exact test (Fisher, 1971) to determine a likelihood that the specific experimental result could have been obtained by chance.

In more detail, we consider the null hypothesis that the disputed document was not written by the purported author, and that there is, in fact, no relationship between them. Under this assumption, the purported author would rank anywhere from #1 to #11 (with equal probability), averaging at the sixth slot. Consistently appearing closer than the sixth slot, then, is evidence of systematic similarity between the two authors across a variety of independent stylometric variables. An unrelated person is unlikely to show this kind of systematic similarity, and hence if the calculated rank sum is small enough, we can reject the null hypothesis at any specific alpha cutoff desired. The system as currently developed uses standard cutoffs: if the p-value is 0.05 or less, we consider this to be "strong indications of common authorship," while trend-level values (p-value of 0.10 or less) are "indications of common authorship." "Weak indications" occur at p-values of 0.20 or less. Inconclusive or outright contraindications are handled appropriately. Indications of different authorship are handled at the other tail of the distribution; for example "strong indications of different authorship" are defined as a p value of 0.95 or greater. (From a theoretical perspective, of course, we would expect two unrelated authors to produce a p-value, on average, of 0.50; we thus acknowledge that the category names are biased somewhat against dissimilar authorship.)

Investigating independence

One major issue is the unwarranted independence assumptions implicit in the fusion framework. Two analyses are "independent" if knowing the outcome of one analysis gives you no information that would let you predict the other analysis. A classic example of this would be two well-shuffled decks of cards; drawing an ace from one deck tells you little about what you would get drawing from the other. By contrast, two draws from the same deck are not independent; if you draw an ace for your first card, there are fewer aces left to be drawn, and the odds of drawing an ace are lowered slightly . Similarly, the odds of drawing a queen are raised slightly. "Card counters" use this lack of independence to estimate odds in a professional gambling context.

In a forensic linguistics context, if we determine that the questioned document is closer in terms of word length distribution to unrelated distractor author A than it is to author B, does this imply that the questioned document will also be more similar in terms of punctuation to A than to B?

If method 1 is right 90% percent of the time, and method 2 is right 90% of the time, that does not mean that both methods will be wrong only one time in 100. That calculation (perhaps obviously) only holds if method 1 and method 2 are independent. However, method 1 and method 2 might *never* both be wrong, or, more worrisomely, if method 2 is a simple replication of method 1, method 2 might be wrong every time method 1 is wrong, so both are wrong a full 10% of the time. To properly validate this system will require analysis of potential inter-analysis dependencies and updating the fusion appropriately.

We can see some preliminary data from Juola's Baggins case (Juola, 2013c). As discussed above, the data were analyzed twice using two methods (that differed primarily in feature weighting). If these two methods are, indeed, independent, the fact that a particular distractor author is first in one condition would not provide any information about that author's position in the other. More formally, we would expect zero correlation between the rank-orders of the two conditions.

Table 2. Actual outcomes of Baggins article analysis

Condition 1	Condition 2
Baggins	Baggins
Distractor 1	Distractor 2
Distractor 2	Distractor 4
Distractor 4	Distractor 5
Distractor 5	Distractor 1
Distractor 3	Distractor 3

A quick inspection shows that these two orderings do not appear independent; for example, distractor 3 is last in both conditions. This is precisely as unlikely as Baggins being first in both analyses. Expressed more formally, the calculated rank-order correlation is 0.4, yielding a p-value of roughly 0.50. Given the small sample size, this is not strong enough to reject the hypothesis of no correlation, but it is not sufficient either to make one feel comfortable claiming that independence is likely, or even plausible.

In this case, the assumptions behind Fisher's exact test do not hold and the numerical calculations described above are not necessarily accurate. Further work is obviously required to assess the relative independence of any proposed methods for an Envelope-like system. The joint accuracy can be determined using more sophisticated fusion methods, but these methods are typically not easily understandable and not something one would wish to bring into court to present to a judge or jury. Alternatively, one can perform experiments to determine empirically the accuracy of such a system under controlled conditions and present the results of those experiments as an estimate of the overall accuracy of the analysis. This method, discussed in the following sections, provides us with an easily understandable assessment of the accuracy and therefore the weight to be given to any particular piece of evidence.

Discounting for a moment the potential issue of independence assumptions, the system is designed to be capable of delivering a sophisticated stylometric analysis quickly, cheaply, and without human intervention (thereby minimizing analyst bias effects).

Accuracy and Validation

System accuracy

To enhance validity, the system as implemented performs a data validation process. Both the known and disputed documents need to be of sufficient length (currently defined as $=<$ 200 words), and cannot include header information (which can be picked up, for example, by looking for *From:* lines). Furthermore, the documents must be in English (Cavnar and Trenkle, 1994) (which we currently approximate by confirming the existence of "the" in the files). Violations of these conditions are documented in the generated report but do not prevent analysis; more sophisticated (and expensive) human-based analysis may be necessary in these circumstances. For example, stylometric analysis technology is known to transfer well between languages (Hasanaj, 2013; Hasanaj *et al.*, 2014; Juola, 2009), but a new distractor corpus would be necessary.

Preliminary testing: English-language email

The accuracy of this system has been tested on a variety of other email samples drawn from 20 additional authors in the Enron (Klimt and Yang, 2004) corpus. Out of 375 trials, 179 produced "strong" indications of authorship, and all 179 (100%) were correct. Similarly, "Weak" indications of authorship were correct in 21 of 23 cases (91%). Only 2 cases showed just "indications", and 1 of those (50%) was correct, while the remaining 43 inconclusive cases could not be validated, but showed significant numbers of both same (6) and different (37) author pairs. Thus, as expected, this method does not return an answer in all cases, but when an answer is returned, the accuracy is very high.

Additional testing: English-language blogs

For a more extensive evaluation, we turned to the Blog Authorship Corpus (Schler *et al.*, 2006)[4]. This corpus provides the collected posts of nearly 20,000 bloggers, containing nearly 700,000 posts and 140 million words. We extracted approximately 8,000 blogs containing 300 or more sentences. We first extracted a fixed set (as per *Envelope* design of ten designated distractor authors and collected the first 100 sentences as distractor samples. For same-author tests, we randomly selected 4,000 blogs. From these blogs, we collected the first hundred sentences as a sample KD, and the last hundred sentences as a sample QD (thus providing maximal opportunity for topic and stylistic drift).

For different-author tests, we selected 4,000 additional blogs and paired them randomly in a daisy-chain structure. From these blogs, we used the first hundred sentences as a known document and the last hundred sentences from a different blog as a questioned document. No blogger appeared in both the same-author and different author tests. Aside from the distractor authors, no passage was analyzed more than once across the entire experiment suite.

This procedure yielded 4000 independent tests of same-author accuracy and of different-author accuracy. Table 3 shows the results, using the previously defined Envelope categories.

Table 3. Results of blog analysis under same- and different-author conditions

Result	Same-author	Different-author	Odds Ratio
Strong same	2,948	748	3.941
Same	246	359	0.686
Weak same	195	396	0.492
Inconclusive	409	1,390	0.294
Weak different	54	234	0.231
Different	47	230	0.204
Strong different	91	663	0.137

The final column of table 3 shows the odds ratio – the number of same-author attributions in that category divided by the number of different-author attributions in that category.

Application: The Case of "Rogue POTUS Staff"

We present one such case study, that of "Rogue POTUS ("President Of The United States.") Staff" (henceforth RPS), an anonymous political commentator, ostensibly from within the Trump White House staff.

Case background

The Twitter microblogging platform provides an easy way to publish short texts (140 characters, recently raised to 280) to a wide audience. It has become one of the largest social media platforms, with more than 330 million monthly active users (https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/) and more than a billion visits per month. Among those users is "Rogue POTUS Staff" (@RoguePOTUSSTAFF), self-described[5] as "The unofficial resistance team inside the White House. We pull back the curtain to expose the real workings inside this disastrous, frightening Administration."

Since even before its inauguration, the Trump administration has been plagued by controversy; staff turnover in the White House has been "off the charts" [6]. While historians will dig at the inner workings of the Trump White House for decades, one clear factor will be the tension between two groups of Trump supporters that are largely at odds, the traditional Republican establishment (such as former DIA director Flynn) and the alt-right ideologues (such as Bannon). Bannon's firing on August 18, 2017 was a result of just such a power struggle.

Approximately a month after Bannon's dismissal, the British online newspaper METRO published an article [7] suggesting that RPS was, in fact, a Twitter account run and written by Bannon himself. The primary evidence cited in this article was a suspicious correspondence in timing; RPS' last tweet was on August 16, two days before Bannon's firing. (After this time, no more tweets were issued from this account until October 29, more than a month after the article in question, and the October tweet contains no actual information, simply a threat: "Did you let silence become a false friend of security Mr. President? Tick tock, tick tock.") This is in marked contrast to previous activity — for example, on July 21, RPS posted twenty separate messages. Did something happen to RPS in mid-August?

Others have disputed this account, claiming that RPS is simply a hoax.[8] This, then, can be viewed as a classic instance of anonymous political discourse such as the publication by "Publius" of the Federalist Papers. The question of whether RPS and Bannon are the same author is thus a typical authorship verification problem, very similar to the question of whether "Publius" is the same person as Alexander Hamilton (Mosteller and Wallace, 1964) or whether Galbraith and Rowling are the same, and to be handled in a very similar way.

Materials and Methods

The Envelope method described in the previous sections was applied to answer this question.

As is typical, it was first necessary to collect undisputed samples of both RPS' writing style and Bannon's. Unfortunately, collecting writing sample data on Twitter is problematic. Much of Twitter is simply "re-tweets" of other people's writings (presumably in other people's style), and "almost 50% of (Twitter) traffic is generated and propagated by a rapidly growing bot population" (Juola *et al.*, 2018) (which again would not reflect the ostensible author's style). Bannon's official Twitter feed, in particular, consists almost exclusively of procedurally generated tweets. However, Bannon has written a lot of articles for the Breitbart media platform which are published under his own name. Similarly, RPS has a web presence (http://potusstaff.com/) containing, among other things, editorial articles similar to those on Breitbart. We collected two RPS articles, and articles by Bannon as well as nine other Breitbart authors as distractors.

All analyses were performed using the Envelope engine. As described above, five sub-analyses were performed and folded into the overall system. To recap, we analyzed the authorial vocabulary (vocabulary overlap), expressive complexity (word lengths), character 4-grams, function words (the 50 most frequent words) and punctuation.

Our first comparison was of one RPS document to another, to confirm that RPS was, in fact, a single-author project (their use of "we" notwithstanding). One RPS document was used as the "questioned" or "unknown" document, and compared to eleven other documents (ten non-Bannon distractors and the second RPS document). If RPS were, in fact, a unitary author, then we would expect a ranking near 1 reflecting the stylistic uniformity. If RPS were not a unitary author, then there is little reason to suppose that the two RPS documents would be similar, and the expected rank might be anywhere from 1 to 11, averaging a 6. Fisher's exact test can measure the likelihood of a chance similarity with precision. Similarly, we compared Bannon's Breitbart articles to each of the two RPS articles separately, resulting in two additional results. All results are presented in the following subsection.

Results

Comparing the two RPS articles against each other produced a measured (theoretical) p-value of 0.008. Table 3 shows that when blog posts of comparable p-value ($p < 0.05$) are analyzed, the results are 4:1 that they are by the same author. We therefore conclude that these two articles (and by extension, the RPS editorials on http://potusstaff.com) are by the same, single author.

Conversely, the two RPS articles compared to the known Bannon article produced p-values of 0.6343 and 0.9729, respectively. In other words, not only were the articles not particularly similar, they were in fact more dissimilar than they were similar (more than half of the distractor authors were more similar). The odds ratio from table 3 are roughly 3:1 and 7:1 (respectively) against the articles being by the same person.

This demonstrates that there are substantial and robust stylistic differences between Bannon's writing and that of the (unknown) RTS author, while the style of the RTS author is uniform enough to allow us to believe him/her to be a single person. This strongly suggests that Metro was wrong and that "Rogue POTUS Staff" was not, in fact, Stephen Bannon. Our results further suggest that we can have high confidence in this finding.

Discussion

Precision and Recall

As was seen in the large-scale tests, nearly three-fourths of the actual same-author cases were identified as "strong indicators of common authorship," a recall rate of nearly 75%. Less than 2.5% were categorized as "strong indicators of different authorship." Thus documents actually by the same author are highly likely to be identified as such. Similarly, documents classified as "strong indicators of common authorship" included 2,948 correct attributions out of 3,696 so classified, a precision of 80%.

At the same time, documents by different authors are substantially less skewed, with the paradoxical result that "weak indications" or even merely "indications" of similar authorship are actually less likely to arise from same-author analyses than from different-author analysis.

The odds-ratio clearly shows that, as the strength of indication of similar authorship goes down, the probability of similar authorship also decreases. The directionality of this relationship is perfect, in keeping with previous research (e.g. DeCarlo, 2013).

At the same time, it is also clear that, in contrast to theoretical predictions, the distribution of Fisher scores and by extension p-values in the different-author case is not uniform. For example, only 5% of the scores are expected to return p-values of 0.05 or less, while 18.7% actually did. This indicates, as discussed in the following section, a need for greater independence among the individual tests.

Genre effects

A second concern relates to the relationship between the calibration studies and RPS analysis; the calibration studies were done with blogs, not editorial articles. This genre will probably not directly affect our conclusion about RPS' identity, but may affect our confidence in unknown ways. One factor that should not be a concern are issues of representativeness in the distractor set, as recent research has shown that this does not affect accuracy very much (DeCarlo, 2013).

Finally, our analysis hinges crucially on the non-mathematical assumption, first, that the articles published under Bannon's by-line in Breitbart are actually by him and not by a ghost-writer, and similarly that the articles on RPS' web site are by the same RPS writer who writes the tweets. While we have shown some stylistic similarities in the RPS articles, there is no practical way to actually validate physical authorship. Of course, analysts have similar issues with many other authors; few if any modern scholars have seen a physical manuscript of Einstein's 1905 relativity paper, so there is no way to disprove the idea that it was written by someone else. In the absence of evidence to the contrary, we make the assumption that the claims of authorship mean what they say.

Conclusions

Despite these concerns, we feel the Envelope system delivers a high-quality analysis quickly and at low cost. Being fully automatic, the analysis is reproducible and is not influenced by analyst bias in any specific case. The probability of error has been confirmed empirically to be low (as expressed in table 3, and is lowest precisely when the analysis yields the strongest results. It is easy to extend the current system to additional languages, additional document types, or even additional classification tasks such as author profiling (Argamon *et al.*, 2009).

Interpreting an Envelope report is fairly straightforward; in the event of a "same author" finding, it means that, at the minimum, the actual author of the questioned document shared five human-understandable characteristics of writing style with the person who wrote the known document of interest. If the author of the will was not the decedent, it was, at a minimum, someone who used the same characteristic vocabulary, syntax, her characteristic style of punctuation, and even the function words in the same way. The computer can characterize the likelihood of this kind of match occurring with a person off the street using the statistics described above. As the old joke has it, "if it looks like a duck, walks like a duck, and uses punctuation like a duck. . . ."

There are a number of fairly obvious extensions and possible improvements. Extension to new genres and/or languages (Hasanaj, 2013; Hasanaj *et al.*, 2014; Juola, 2009) can be as simple as the creation of a new set of distractor documents. It may be possible to improve the accuracy by the incorporation of other analysis methods and feature sets (for example, the distribution of part-of-speech tags), although high-level processing such as POS tagging may limit its use in other languages. We continue our preliminary testing and will be expanding our offerings in terms of genre.

So, who did write the will, or at least (in the modern Christie remake) the Web posting? Who wrote the email ostensibly dividing up ownership of the startup, or revealing confidential business information? Computational analysis, as typified by Envelope, may not be able to provide definitive answers, but the evidence it creates can provide valuable information to help guide investigations or suggest preliminary conclusions. This system provides low-cost advice without the time and cost of human analysis, while retaining high accuracy.

Notes

[1] For those not familiar with Christie's work, an appropriate start might be *Peril at End House.*

[2] The reader is invited to look at Dorothy L. Sayer's *Strong Poison.*

[3] The JGAAP program is freely available as an open-source program; we have used it as well for the present study.

[4] See also `http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm`

[5] See https://twitter.com/roguepotusstaff?lang=en.

[6] https://www.newyorker.com/news/news-desk/a-year-into-the-trump-era-white-house-staff-turnover-is-off-the-charts

[7] http://metro.co.uk/2017/09/29/was-rogue-white-house-twitter-account-actually-steve-bannon-6965449/

[8] Cf. https://theoutline.com/post/2396/trump-resistance-phonies

References

Argamon, S., Koppel, M., Pennebaker, J. W. and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119–123.

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.

Brooks, R. (2013). Whodunnit? JK Rowling's secret life as wizard crime writer revealed. *Sunday Times*, 14 July.

Brooks, R. and Flyn, C. (2013). JK Rowling: The cuckoo in crime novel nest. *Sunday Times*, 14 July.

Burrows, J. F. (1989). 'an ocean where each kind…' : Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5), 309–21.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *1994 Symposium on Document Analysis and Information Retrieval*, 161–176.

Chaski, C. (2013). Best practices and admissibility of forensics author identification. *Journal of Law and Policy*, XXI(2), 333–376.

Chaski, C. E. (2005). Who's at the keyboard: Authorship attribution in digital evidence invesigations. *International Journal of Digital Evidence*, 4(1), n/a. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

Collins, P. (2013). Poe's debut, hidden in plain sight. *The New Yorker*, October.

Coulthard, M. (2013). On admissible linguistic evidence. *Journal of Law and Policy*, XXI(2), 441–466.

DeCarlo, E. (2013). Inferring authorship through Myers-Briggs Type Inventory. In *Proceedings of DHCS 2013*, Chicago.

Fisher, R. A. (1971). *The Design of Experiments.* New York: Macmillan, 9th ed.

Grant, T. (2013). Txt 4n6: Describing and measuring consistency and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy*, XXI(2), 467–494.

Grieve, J. W. (2005). Quantitative authorship attribution: A history and an evaluation of techniques. Master's thesis, Simon Fraser University. URI: http://hdl.handle.net/1892/2055, accessed 5.31.2007.

Hasanaj, B. (2013). Authorship attribution methods in Albanian. In *Duquesne University Graduate Student Research Symposium.*

Hasanaj, B., Purnell, E. and Juola, P. (2014). Cross-linguistic transference of authorship attribution. In *Proceedings of the International Quantitative Linguistic Conference (QUALICO).*

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2), 87–106.

Hoover, D. L. (2004). Delta prime? *Literary and Linguistic Computing*, 19(4), 477–495.

Jockers, M. L. and Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *LLC*, 25(2), 215–23.

Juola, P. (2004). Ad-hoc authorship attribution competition. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden.

Juola, P. (2006a). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Juola, P. (2006b). Authorship attribution for electronic documents. In M. Olivier and S. Shenoi, Eds., *Advances in Digital Forensics II*, volume 222 of *International Federal for Information Processing*. Boston: Springer, 119–130.

Juola, P. (2007). Future trends in authorship attribution. In P. Craiger and S. Shenoi, Eds., *Advances in Digital Forensics III*, International Federal for Information Processing. Boston: Springer, 119–132.

Juola, P. (2008). Authorship attribution : What mixture-of-experts says we don't yet know. In *Proceedings of American Association for Corpus Linguistics 2008*, Provo, UT USA.

Juola, P. (2009). Cross-linguistic transference of authorship attribution, or why english-only prototypes are acceptable. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Juola, P. (2012a). Large-scale experiments in authorship attribution. *English Studies*, 93(3), 275–283.

Juola, P. (2012b). An overview of the traditional authorship attribution subtask. In *Proceedings of PAN/CLEF 2012*, Rome, Italy.

Juola, P. (2013a). A critical examination of the Ceglia/Zuckerberg email authorship study. In *Proceedings of the 11th Biennial Conference on Forensic Linguistics/Language and Law of the International Association of Forensic Linguists (IAFL 2013)*, Mexico City, MX.

Juola, P. (2013b). How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. *Scientific American*, August.

Juola, P. (2013c). Stylometry and immigration: A case study. *Journal of Law and Policy*, XXI(2), 287–298.

Juola, P. (2014). The Rowling case: A proposed standard protocol for authorship attribution. In *Proceedings of Digital Humanities 2014*, Lausanne, Switzerland.

Juola, P. (2015). The Rowling case: A proposed standard protocol for authorship attribution. *DSH (Digital Scholarship in the Humanities).*

Juola, P. (2016). Did Aunt Prunella really write that will? a simple and understandable computational assessment of authorial likelihood. In *Proc. A Workshop on Legal Text, Document, and Corpus Analytics (LTDCA 2016)*, 37–41.

Juola, P., Mikros, G. K. and Vinsick, S. (2018). Correlations and potential cross-linguistic indicators of writing style. *Journal of Quantitative Linguistics*, 26(2), 146–171.

Juola, P., Noecker, Jr. J., Ryan, M. and Speer, S. (2009). Jgaap 4.0 — a revised authorship attribution tool. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Juola, P., Noecker Jr, J. I., Stolerman, A., Ryan, M. V., Brennan, P. and Greenstadt, R. (2013). Keyboard behavior-based authentication for security. *IT Professional*, 15, 8–11.

Juola, P., Sofko, J. and Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21(2), 169–178. Advance Access published on April 12, 2006; doi: doi:10.1093/llc/fql019.

Juola, P. and Stamatatos, E. (2013). Overview of the authorship identification task. In *Proceedings of PAN/CLEF 2013*, Valencia, Spain.

Klimt, B. and Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004*, 217–226.

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. 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.

McMenamin, G. (2011). Declaration of Gerald McMenamin. Available online at `http://www.scribd.com/doc/67951469/Expert-Report-Gerald-McMenamin`.

Mikros, G. K. and Perifanos, K. (2013). Authorship attribution in greek tweets using multilevel author's n-gram profiles. In *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013, Stanford, California.* Palo Alto, California: AAAI Press, 17–23.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship : The Federalist*, volume 58. Addison-Wesley.

Noecker Jr., J. and Juola, P. (2009). Cosine distance nearest-neighbor classification for authorship attribution. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, 351–365.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.*

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–56.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, XXI(2), 420–440.

Stamatatos, E., Stein, B., Daelemans, W., Juola, P., Barrón-Cedeño, A., Verhoeven, B. and Sanchez-Perez, M. A. (2014). Overview of the authorship identification task at PAN 2014. In *Proceedings of PAN/CLEF 2014*, Sheffield, UK.

Vescovi, D. M. (2011). Best practices in authorship attribution of English essays. Master's thesis, Duquesne University.

Wellman, F. L. (1936). *The Art of Cross-Examination.* New York: MacMillan, 4th ed.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* New York: Hafner Publishing Company. Reprinted 1965.