Chapter 9

ANALYZING STYLOMETRIC APPROACHES TO AUTHOR OBFUSCATION

Patrick Juola and Darren Vescovi

Abstract Authorship attribution is an important and emerging security tool. However, just as criminals may wear gloves to hide their fingerprints, so too may criminal authors mask their writing styles to escape detection. Most authorship studies have focused on cooperative and/or unaware authors who do not take such precautions. This paper analyzes the methods implemented in the Java Graphical Authorship Attribution Program (JGAAP) against essays in the Brennan-Greenstadt obfuscation corpus that were written in deliberate attempts to mask style. The results demonstrate that many of the more robust and accurate methods implemented in JGAAP are effective in the presence of active deception.

Keywords: Authorship attribution, stylometry, obfuscation, deception

1. Introduction

The determination of the author of a particular piece of text has been a methodological issue for centuries. Questions of authorship are of interest to scholars, and in a much more practical sense to politicians, journalists and lawyers. In recent years, the development of improved statistical techniques [6, 11] in conjunction with the wider availability of computer-accessible corpora [4, 21] have made the automatic inference of authorship at least a theoretical possibility. Consequently, research in the area of authorship attribution has expanded tremendously.

From the legal and security perspectives, it is not enough to merely identify an unsuspecting author. Just as criminals wear gloves to hide their fingerprints, criminal authors often attempt to disguise their writing styles based on the expectation that their writings will be analyzed

by law enforcement. However, it is not clear that a method that can identify Shakespeare would correctly identify an author who is deliberately deceptive. This paper analyzes the methods implemented in the Java Graphical Authorship Attribution Program (JGAAP) against essays in the Brennan-Greenstadt obfuscation corpus [2] that were written in deliberate attempts to mask style.

2. Background

With a history stretching to 1887 [19] and 181,000 hits on Google (corresponding to a phrasal search for "authorship attribution" on June 30, 2010), it is apparent that statistical/quantitative authorship attribution or stylometrics is an active and vibrant research area. However, it is surprising that stylometrics has not been accepted by literary scholars. A discussion of this problem is beyond the scope of this paper. Interested readers are referred to [6, 11] for additional information.

In broad terms, a history of *ad hoc*, problem-focused research has emerged. A scholar interested in a particular document will develop a technique for addressing the document, with little regard to whether or not the technique generalizes to other document types, languages, etc. Similarly, new techniques are often lightly tested on toy problems – the *Federalist* papers are a common candidate – to establish that the methods "work." Since the seminal analysis by Mosteller and Wallace [20] of the distribution of function words in the *Federalist* papers, it has become almost traditional to test new methods on these essays [7, 18, 22, 24]. Rudman [23] lists no less than nineteen studies of this particular corpus and the list is by no means complete. However, it is not clear that this particular (overstudied) corpus is representative of the problem as a whole.

More recent studies [3, 5, 8, 9, 13, 14] have recognized the need for broader data and comparative analysis. Juola's Ad hoc Authorship Attribution Competition (AAAC) [9, 10] has established a moderate-scale empirical testbed for the comparative evaluation of authorship attribution methods. The standardized test corpus allows the demonstration of the ability of statistical methods to determine authorship. Moreover, it enables the methods to be further distinguished between the "successful" and the "very successful."

The AAAC corpus includes thirteen problems in a variety of lengths, styles, genres and languages, mostly gathered from the web, but also comprising some materials specifically collected for the competition. Unfortunately, the AAAC corpus is too small to be truly effective for sorting good from bad methods, which creates opportunities for further research.

2.1 Brennan-Greenstadt Corpus

The AAAC corpus primarily contains historic literary documents, but these documents were not gathered with an eye to address deliberate attempts to mask authorial style. On the other hand, the Brennan-Greenstadt corpus [2] is the first (small-scale) obfuscation corpus that was specifically created to study "adversarial attacks" where writers obfuscate their writing styles and also deliberately imitate the style of other authors. Brennan and Greenstadt collected about 5,000 words of sample writing from each of fifteen authors. The fifteen authors were then asked to write a new 500-word sample in which they hid their identity through their writing style and another sample that imitated the style of Cormac McCarthy as expressed in *The Road*.

Brennan and Greenstadt applied three fairly standard stylometric methods to determine the authorship of the obfuscated essays and the imitative essays. Their results for the obfuscated essays were essentially at chance, while the results for the imitative essays were strongly below chance, suggesting that attempts to disguise or imitate style are likely to be successful against stylometric methods. Brennan and Greenstadt concluded that "obfuscation attacks weaken all three methods to the point that they are no better than randomly guessing the correct author of a document." Brennan and Greenstadt also stated that "[t]he imitation attacks were widely successful in having their authorship attributed to the intended victim of the attack. [...] Frameworks for testing methods of authorship attribution on existing texts have been around for a long time, and now it is clear that there is a need to use a similar framework for testing these very same methods in their resilience against obfuscation, imitation, and other methods of attack." A larger-scale analysis by Juola and Vescovi [17] has confirmed this finding with 160 different stylometric algorithms, none of which were able to crack the problem.

2.2 JGAAP

The Java Graphical Authorship Attribution Program (JGAAP) [15, 16], which was developed at Duquesne University and is freely available at `www.jgaap.com`, incorporates tens of thousands of stylometric methods [12]. JGAAP uses a three-phase modular structure, which is summarized below. Interested readers are referred to [10, 11] for additional information.

- **Canonicization:** No two physical realizations of events are exactly identical. Similar linguistic notions are considered to be identical to restrict the event space to a finite set. This may involve,

for example, unifying case, normalizing whitespace, de-editing to remove page numbers, or correcting spelling and typographic errors.

- **Event Set Determination:** The input stream is partitioned into individual "events," which could be words, parts of speech, characters, word lengths, etc. Uninformative events are eliminated from the event stream.

- **Statistical Inference:** The remaining events are subjected to a variety of inferential statistics ranging from simple analysis of event distributions to complex pattern-based analysis. The statistical inferences determine the results (and confidence) in the final report.

Brennan and Greenstadt were able to obtain permission to publish only twelve of the fifteen essay sets. However, we were able to re-analyze these essays against a much larger set of more than 1,000 attribution methods.

3. Materials and Methods

Twelve of the fifteen essay sets in the Brennan-Greenstadt corpus were re-analyzed using JGAAP 4.1. The following methods are available or are implemented directly in JGAAP 4.1:

- **Canonicizer (Unify Case):** All characters are converted to lower case.

- **Canonicizer (Strip Punctuation):** All non-alphanumeric and non-whitespace characters are removed.

- **Canonicizer (Normalize Whitespace):** All strings of consecutive whitespace characters are replaced by a single "space" character.

- **Event Set (Words):** Analysis is performed on all words (maximal non-whitespace substrings).

- **Event Set (2-3 Letter Words):** Analysis is performed on all words (maximal non-whitespace substrings) of two or three letters (e.g., "to" and "the").

- **Event Set (3-4 Letter Words):** Analysis is performed on all words (maximal non-whitespace substrings) of three or four letters (e.g., "the" and "have").

- **Event Set (Word Bigrams):** Analysis is performed on all word pairs.
- **Event Set (Word Trigrams):** Analysis is performed on all word triples.
- **Event Set (Word Stems):** Document words are stemmed using the Porter stemmer [25] and analysis is performed on the resulting stems.
- **Event Set (Parts of Speech):** The document is tagged with the part of speech of each word and analysis is performed on the parts of speech.
- **Event Set (Word Lengths):** Analysis is performed on the number of characters in each word.
- **Event Set (Syllables per Word):** Analysis is performed on the number of syllables (defined as separate vowel clusters) in each word.
- **Event Set (Characters):** Analysis is performed on the sequence of ASCII characters that make up the document.
- **Event Set (Character Bigrams):** Analysis is performed on all character bigrams (e.g., "the word" becomes "th," "he," "e ," " w" and so on).
- **Event Set (Character Trigrams):** Analysis is performed on all character trigrams (e.g., "the word" becomes "the," "he ," "e w," " wo" and so on).
- **Event Set (Binned Frequencies):** Analysis is performed on the frequencies of each word as measured by the English Lexicon Project [1].
- **Event Set (Binned Reaction Times):** Analysis is performed on the average lexical decision time of each word as measured by the English Lexicon Project [1].
- **Event Set (Mosteller-Wallace Function Words):** Analysis is performed on all instances of word tokens in the Mosteller-Wallace analysis set derived from the *Federalist* papers [20]. In other research (in preparation), we have shown that this method tends not to perform well because the function words appear to be overtuned to this particular document set.

- **Inference (Histogram Distance):** Events are treated as "bags of events" (without regard to ordering). Histograms are created for each document pair, pairwise distances are calculated using the standard Euclidean (root-mean-square) metric, and authorship is assigned to the single nearest document of known authorship (one-nearest neighbor).
- **Inference (Manhattan Distance):** Same as above, except that distances are calculated using the Manhattan or L_1 Minkowski distance.
- **Inference (Cosine Distance):** Same as above, except that distances are calculated using the normalized cosine or dot product distance.
- **Inference (Kolmogorov-Smirnov Distance):** Same as above, except that distances are calculated using the nominal Kolmogorov-Smirnov distance.
- **Inference (Cross-Entropy Distance):** Same as above, except that distances are calculated using cross-entropy divergence.
- **Inference (Kullback-Leibler Distance):** Same as above, except that distances are calculated using Kullback-Leibler divergence.
- **Inference (Linear Discriminant Analysis):** Authorship is assigned using linear discriminant analysis based on the vector space defined by the event histograms described above.
- **Inference (Gaussian Support Vector Machine):** Authorship is assigned using Gaussian support vector machines based on the vector space defined by the event histograms described above.
- **Inference (Naive Bayes):** Authorship is assigned using naive Bayesian inference.

In all cases, the analysis was restricted to the fifty most common events of a given type (i.e., the fifty most common character trigrams instead of the million or so trigrams, most of which hardly appear in English). Each experiment used any or all of the 8 ($= 2^3$) possible canonicizer sets (including "null"), one event set and one inference method. This resulted in a total of 1,080 experiments.

Each experiment analyzed all twelve obfuscated essays against all possible training documents, with a resulting chance-level performance of approximately 8.33% (one in twelve) correct. Each experiment also

analyzed all twelve imitative essays (including samples from Cormac McCarthy) to determine whether or not they are correctly attributed (among all thirteen candidates), and whether or not they are incorrectly attributed to McCarthy.

4. Experimental Results

The results of any given experiment could vary widely. In some cases, specific combinations of canonicizer, event set and inference got no attributions correct; in other cases, combinations got as many as 42% (five out of twelve) of the obfuscated attributions correct. Similarly, some of the methods were able to correctly identify the author of as many as 25% of the imitative essays, with no attributions to the imitated author. On the other hand, some combinations made no correct attributions or attributed more than 80% of the imitated essays to McCarthy.

Table 1 summarizes the overall results by the method used. The table shows the results of all the experiments using a particular method, e.g., all 540 experiments in which unifying case was performed (or not performed), or all 120 experiments in which the inference engine used the cosine distance.

5. Discussion

From a formal statistical point of view, no method was able to perform "significantly" above chance at the standard 0.05 level, but several interesting patterns emerged. In this strict sense, we are reporting a negative result – we were unable to find a "silver bullet" that reliably solves the hostile author problem. Nevertheless, we consider these findings to be interesting and provocative.

We first observe that there indeed is a confirmed problem. Although the analyses performed (on the average) above chance, they did not do so by robust margins, and there is enough variance in individual performance that no "significant" improvement can be claimed. We also observe that no amount of pre-processing (canonicization) significantly improves the accuracy for any problem; similarly, the type of inference employed appears to make little difference.

On the other hand, the type of event and feature set analyzed do appear to make a difference. Character-based events (bigrams and trigrams in the analysis) appear to be more robust to obfuscative attacks than word-based events; preliminary analysis of the Juola-Vescovi statistics [17] bears this out at a marginal value of $p < 0.18$. This makes intuitive sense. To the extent that authorial "style" is a function of specific vocabulary items, it is easy for an author to mask his or her style by picking

Table 1. Experimental results.

Method	Number Expts.	Proportion Correct [Obfuscative] Mean(SD)	Proportion Correct [Imitative] Mean(SD)	Proportion Attributed to McCarthy Mean(SD)
Chance	—	0.083	0.077	0.077
ALL	1,080	0.099(0.075)	0.040(0.060)	0.478(0.295)
Unify case	540	0.098(0.070)	0.038(0.058)	0.478(0.286)
No unify case	540	0.100(0.080)	0.041(0.063)	0.478(0.305)
Strip punctuation	540	0.101(0.072)	0.045(0.066)	0.476(0.295)
No strip punctuation	540	0.098(0.078)	0.034(0.054)	0.479(0.296)
Norm. white space	540	0.099(0.075)	0.037(0.058)	0.486(0.294)
Non-norm. white space	540	0.100(0.076)	0.042(0.062)	0.470(0.296)
Character event sets	216	0.161(0.96)	0.034(0.051)	0.524(0.289)
Numeric event sets	216	0.080(0.045)	0.050(0.049)	0.403(0.278)
Word event sets	648	0.085(0.064)	0.038(0.066)	0.487(0.299)
Words	72	0.079(0.056)	0.014(0.037)	0.574(0.240)
2-3 letter words	72	0.039(0.046)	0.025(0.065)	0.559(0.193)
3-4 letter words	72	0.083(0.063)	0.014(0.031)	0.521(0.199)
Word bigrams	72	0.063(0.072)	0.095(0.097)	0.292(0.345)
Word trigrams	72	0.097(0.058)	0.074(0.062)	0.141(0.311)
Word stems	72	0.081(0.054)	0.014(0.037)	0.593(0.230)
Parts of speech	72	0.120(0.073)	0.052(0.076)	0.591(0.252)
Word lengths	72	0.088(0.044)	0.0(0.0)	0.620(0.254)
Syllables per word	72	0.083(0.0)	0.065(0.053)	0.454(0.244)
Characters	72	0.110(0.074)	0.043(0.048)	0.517(0.261)
Character bigrams	72	0.161(0.060)	0.012(0.029)	0.654(0.257)
Character trigrams	72	0.213(0.117)	0.047(0.062)	0.402(0.293)
Binned frequencies	72	0.076(0.061)	0.046(0.050)	0.488(0.263)
Binned reaction times	72	0.081(0.050)	0.039(0.042)	0.266(0.276)
MW function words	72	0.116(0.063)	0.053(0.066)	0.493(0.260)
Histogram distance	120	0.095(0.089)	0.040(0.055)	0.465(0.209)
Manhattan distance	120	0.108(0.094)	0.040(0.058)	0.464(0.195)
Cosine distance	120	0.101(0.091)	0.028(0.049)	0.472(0.198)
KS distance	120	0.108(0.094)	0.040(0.058)	0.464(0.195)
Cross-entropy	120	0.097(0.067)	0.035(0.051)	0.453(0.239)
KL distance	120	0.097(0.067)	0.035(0.051)	0.453(0.239)
LD analysis	120	0.117(0.075)	0.072(0.091)	0.415(0.226)
Gaussian SVM	120	0.083(0.0)	0.0(0.0)	1.0(0.0)
Naive Bayes	120	0.090(0.041)	0.067(0.063)	0.113(0.225)

different words, but it is difficult to change large-scale emergent statistics such as character frequency. Consider, for example, how reasonable an editorial request to "use American spelling" for a journal article appears, especially in comparison with a request like "use no more than 10% e's."

Oddly enough, this pattern is only borne out for obfuscative attacks; character-based event sets actually perform slightly (and non-significantly) worse than word-based ones in their ability to attribute

imitative essays (both in their overall ability and in their ability to eliminate the imitated target).

Examining the individual top (and bottom) performers offers some interesting suggestions. For example, of the top 28 performers on the obfuscative attacks, no less than 26 use character trigrams as an event set (the other two use parts of speech tags). The seven best performers all use character trigrams and the nearest-neighbor, but with four different distances, and for the most part (4 out of 7) use no canonicization. All the bottom performers on this task give no correct answers and do so for a variety of methods, essentially representing the floor effect.

Similar domination is seen in the imitative event sets. The best performance (33% correct attribution with no misattribution to McCarthy) is achieved by four different versions of word bigrams using linear discriminant analysis (LDA) as the analysis method, but LDA, in particular, dominates the top performing cases, with fifteen of the top fifteen sets.

6. Conclusions

The results of this paper provide partial support and partial refutation of the research of Brennan and Greenstadt. Active deception is a problem for the current state of stylometric art. Tests of about a thousand of the more than 20,000 methods available in the stylometric tool suite that was employed indicate that some of the individual combinations appear to perform at levels much beyond chance on the deceptive corpus. At the same time, no "silver bullets" were discovered that could help pierce the deception.

Still, we remain hopeful. Clearly, much more work remains to be done in investigating other methods of attribution. More importantly, there is the distinct possibility that some principles could improve our search. For example, character-based methods could, perhaps, outperform word-based ones, at least for simple attempts to disguise style without focusing on specific imitation.

Acknowledgements

This research was partially supported by the National Science Foundation under Grant Numbers OCI-0721667 and OCI-1032683.

References

[1] D. Balota, M. Yap, M. Cortese, K. Hutchison, B. Kessler, B. Loftis, J. Neely, D. Nelson, G. Simpson and R. Treiman, The English Lexicon Project, *Behavior Research Methods*, vol. 39, pp. 445–459, 2007.

[2] M. Brennan and R. Greenstadt, Practical attacks against authorship recognition techniques, *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, pp. 60–65, 2009.

[3] C. Chaski, Empirical evaluations of language-based author identification techniques, *International Journal of Speech, Language and the Law*, vol. 8(1), pp. 1–65, 2001.

[4] G. Crane, What do you do with a million books? *D-Lib Magazine*, vol. 12(3), 2006.

[5] R. Forsyth, Towards a text benchmark suite, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 1997.

[6] D. Holmes, Authorship attribution, *Computers and the Humanities*, vol. 28(2), pp. 87–106, 1994.

[7] D. Holmes and R. Forsyth, The *Federalist* revisited: New directions in authorship attribution, *Literary and Linguistic Computing*, vol. 10(2), pp. 111–127, 1995.

[8] D. Hoover, Delta prime? *Literary and Linguistic Computing*, vol. 19(4), pp. 477–495, 2004.

[9] P. Juola, Ad hoc Authorship Attribution Competition, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[10] P. Juola, Authorship attribution for electronic documents, in *Advances in Digital Forensics II*, M. Olivier and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 119–130, 2006.

[11] P. Juola, Authorship attribution, *Foundations and Trends in Information Retrieval*, vol. 1(3), pp. 233–334, 2008.

[12] P. Juola, 20,000 ways not to do authorship attribution – and a few that work, presented at the *Biennial Conference of the International Association of Forensic Linguists*, 2009.

[13] P. Juola, Cross-linguistic transference of authorship attribution, or why English-only prototypes are acceptable, presented at the *Digital Humanities Conference*, 2009.

[14] P. Juola and H. Baayen, A controlled-corpus experiment in authorship attribution by cross-entropy, *Literary and Linguistic Computing*, vol. 20, pp. 59–67, 2005.

[15] P. Juola, J. Noecker, M. Ryan and S. Speer, JGAAP 4.0 – A revised authorship attribution tool, presented at the *Digital Humanities Conference*, 2009.

[16] P. Juola, J. Sofko and P. Brennan, A prototype for authorship attribution studies, *Literary and Linguistic Computing*, vol. 21(2), pp. 169–178, 2006.

[17] P. Juola and D. Vescovi, Empirical evaluation of authorship obfuscation using JGAAP, *Proceedings of the Third ACM Workshop on Artificial Intelligence and Security*, pp. 14–18, 2010.

[18] C. Martindale and D. McKenzie, On the utility of content analysis in authorship attribution: The *Federalist* papers, *Computers and the Humanities*, vol. 29(4), pp. 259–270, 1995.

[19] T. Mendenhall, The characteristic curves of composition, *Science*, vol. IX, pp. 237–249, 1887.

[20] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Massachusetts, 1964.

[21] J. Nerbonne, The data deluge: development and delights, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[22] M. Rockeach, R. Homant and L. Penner, A value analysis of the disputed *Federalist* papers, *Journal of Personality and Social Psychology*, vol. 16, pp. 245–250, 1970.

[23] J. Rudman, The non-traditional case for the authorship of the twelve disputed *Federalist* papers: A monument built on sand, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2005.

[24] F. Tweedie, S. Singh and D. Holmes, Neural network applications in stylometry: The *Federalist* papers, *Computers and the Humanities*, vol. 30(1), pp. 1–10, 1996.

[25] P. Willett, The Porter stemming algorithm: Then and now, *Program: Electronic Library and Information Systems*, vol. 40(3), pp. 219–223, 2006.