Obfuscating Document Stylometry to Preserve Author Anonymity

Gary Kacmarcik **Michael Gamon**
Natural Language Processing Group
Microsoft Research
Redmond, WA USA
{garykac,mgamon}@microsoft.com

Abstract

This paper explores techniques for reducing the effectiveness of standard authorship attribution techniques so that an author *A* can preserve anonymity for a particular document *D*. We discuss feature selection and adjustment and show how this information can be fed back to the author to create a new document *D'* for which the calculated attribution moves away from *A*. Since it can be labor intensive to adjust the document in this fashion, we attempt to quantify the amount of effort required to produce the anonymized document and introduce two levels of anonymization: shallow and deep. In our test set, we show that shallow anonymization can be achieved by making 14 changes per 1000 words to reduce the likelihood of identifying *A* as the author by an average of more than 83%. For deep anonymization, we adapt the unmasking work of Koppel and Schler to provide feedback that allows the author to choose the level of anonymization.

1 Introduction

Authorship identification has been a long standing topic in the field of *stylometry*, the analysis of literary style (Holmes 1998). Issues of style, genre, and authorship are an interesting sub-area of text categorization. In authorship detection it is not the topic of a text but rather the stylistic properties that are of interest. The writing style of a particular author can be identified by analyzing the form of the writing, rather than the content. The analysis of style therefore needs to abstract away from the content and focus on the content-independent form of the linguistic expressions in a text.

Advances in authorship attribution have raised concerns about whether or not authors can truly maintain their anonymity (Rao and Rohatgi 2000). While there are clearly many reasons for wanting to unmask an anonymous author, notably law enforcement and historical scholarship, there are also many legitimate reasons for an author to wish to remain anonymous, chief among them the desire to avoid retribution from an employer or government agency. Beyond the issue of personal privacy, the public good is often served by whistle-blowers who expose wrongdoing in corporations and governments. The loss of an expectation of privacy can result in a chilling effect where individuals are too afraid to draw attention to a problem, because they fear being discovered and punished for their actions.

It is for this reason that we set out to investigate the feasibility of creating a tool to support anonymizing a particular document, given the assumption that the author is willing to expend a reasonable amount of effort in the process. More generally, we sought to investigate the sensitivity of current attribution techniques to manipulation.

For our experiments, we chose a standard data set, the Federalist Papers, since the variety of published results allows us to simulate authorship attribution "attacks" on the obfuscated document. This is important since there is no clear consensus as to which features should be used for authorship attribution.

2 Document Obfuscation

Our approach to document obfuscation is to identify the features that a typical authorship attribution technique will use as markers and then adjust the frequencies of these terms to render them less effective on the target document.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 444–451,
Sydney, July 2006. ©2006 Association for Computational Linguistics

While it is obvious that one can affect the attribution result by adjusting feature values, we were concerned with:

- How easy is it to identify and present the required changes to the author?
- How resilient are the current authorship detection techniques to obfuscation?
- How much work is involved for the author in the obfuscation process?

The only related work that we are aware of is (Rao and Rohatgi 2000) who identify the problem and suggest (somewhat facetiously, they admit) using a round-trip machine translation (MT) process (e.g., English → French → English) to obscure any traces of the original author's style. They note that the current quality of MT would be problematic, but this approach might serve as a useful starting point for someone who wants to scramble the words a bit before hand-correcting egregious errors (taking care not to re-introduce their style).

2.1 The Federalist Papers

One of the standard document sets used in authorship attribution is the Federalist Papers, a collection of 85 documents initially published anonymously, but now known to have been written by 3 authors: Alexander Hamilton, John Madison and John Jay. Due to illness, Jay only wrote 5 of the papers, and most of the remaining papers are of established authorship (Hamilton = 51; Madison = 14; and 3 of joint authorship between Hamilton and Madison). The 12 remaining papers are disputed between Hamilton and Madison. In this work we limit ourselves to the 65 known single-author papers and the 12 disputed papers.

While we refer to these 12 test documents as "disputed", it is generally agreed (since the work of Mosteller and Wallace (1964)) that all of the disputed papers were authored by Madison. In our model, we accept that Madison is the author of these papers and adopt the fiction that he is interested in obscuring his role in their creation.

2.2 Problem Statement

A more formal problem statement is as follows: We assume that an author A (in our case, Madison) has created a document D that needs to be anonymized. The author self-selects a set K of N authors (where $A \in K$) that some future agent (the "attacker" following the convention used in cryptography) will attempt to select between.

The goal is to use authorship attribution techniques to create a new document D' based on D but with features that identify A as the author suppressed.

3 Document Preparation

Before we can begin with the process of obfuscating the author style in D, we need to gather a training corpus and normalize all of the documents.

3.1 Training Corpus

While the training corpus for our example is trivially obtained, authors wishing to anonymize their documents would need to gather their own corpus specific for their use.

The first step is to identify the set of authors K (including A) that could have possibly written the document. This can be a set of co-workers or a set of authors who have published on the topic. Once the authors have been selected, a suitable corpus for each author needs to be gathered. This can be emails or newsgroup postings or other documents. In our experiments, we did not include D in the corpus for A, although it does not seem unreasonable to do so.

For our example of the Federalist Papers, K is known to be {Hamilton, Madison} and it is already neatly divided into separate documents of comparable length.

3.2 Document Cleanup

Traditional authorship attribution techniques rely primarily on associating idiosyncratic formatting, language usage and spelling (misspellings, typos, or region-specific spelling) with each author in the study. Rao and Rohatgi (2000) and Koppel and Schler (2003) both report that these words serve as powerful discriminators for author attribution. Thus, an important part of any obfuscation effort is to identify these idiosyncratic usage patterns and normalize them in the text.

Koppel and Schler (2003) also note that many of these patterns can be identified using the basic spelling and grammar checking tools available in most word processing applications. Correcting the issues identified by these tools is an easy first step in ensuring the document conforms to conventional norms. This is especially important for work that will not be reviewed or edited since these idiosyncrasies are more likely to go unnoticed.

However, there are distinctive usage patterns that are not simple grammar or spelling errors that also need to be identified. A well-known example of this is the usage of while/whilst by the authors of the Federalist Papers.

	Hamilton	Madison	Disputed
while	36	0	0
whilst	1	12	9

Table 1 : Occurrence counts of "while" and "whilst" in the Federalist Papers (excluding documents authored by Jay and those which were jointly authored).

In the disputed papers, "whilst" occurs in 6 of the documents (9 times total) and "while" occurs in none. To properly anonymize the disputed documents, "whilst" would need to be eliminated or normalized.

This is similar to the problem with idiosyncratic spelling in that there are two ways to apply this information. The first is to simply correct the term to conform to the norms as defined by the authors in *K*. The second approach is to incorporate characteristic forms associated with a particular author. While both approaches can serve to reduce the author's stylometric fingerprint, the latter approach carries the risk of attempted style forgery and if applied indiscriminately may also provide clues that the document has been anonymized (if strong characteristics of multiple authors can be detected).

For our experiments, we opted to leave these markers in place to see how they were handled by the system. We did, however, need to normalize the paragraph formatting, remove all capitalization and convert all footnote references to use square brackets (which are otherwise unused in the corpus).

3.3 Tokenization

To tokenize the documents, we separated sequences of letters using spaces, newlines and the following punctuation marks: .,()-:;`'?![]. No stemming or morphological analysis was performed. This process resulted in 8674 unique tokens for the 65 documents in the training set.

4 Feature Selection

The process of feature selection is one of the most crucial aspects of authorship attribution. By far the most common approach is to make use of the frequencies of common function words that are content neutral, but practitioners have also made use of other features such as letter metrics (e.g., bi-grams), word and sentence length metrics, word tags and parser rewrite rules. For this work, we opted to limit our study to word frequencies since these features are generally acknowledged to be effective for authorship attribution and are transparent, which allows the author to easily incorporate the information for document modification purposes.

We wanted to avoid depending on an initial list of candidate features since there is no guarantee that the attackers will limit themselves to any of the commonly used lists. Avoiding these lists makes this work more readily useful for non-English texts (although morphology or stemming may be required).

We desire two things from our feature selection process beyond the actual features. First, we need a ranking of the features so that the author can focus efforts on the most important features. The second requirement is that we need a threshold value so that the author knows how much the feature frequency needs to be adjusted.

To rank and threshold the features, we used decision trees (DTs) and made use of the readily available WinMine toolkit (Chickering 2002). DTs produced by WinMine for continuously valued features such as frequencies are useful since each node in the tree provides the required threshold value. For term-ranking, we created a Decision Tree Root (DTR) ranking metric to order the terms based on how discriminating they are. DTR Rank is computed by creating a series of DTs where we remove the root feature, i.e. the most discriminating feature, before creating the next DT. In this fashion we create a ranking based on the order in which the DT algorithm determined that the term was most discriminatory. The DTR ranking algorithm is as follows:

1) Start with a set of features
2) Build DT and record root feature
3) Remove root feature from list of features
4) Repeat from step 2

It is worth noting that the entire DT need not be calculated since only the root is of interest. The off-the-shelf DT toolkit could be replaced with a custom implementation[1] that returned only the root (also known as a *decision stump*). Since

[1] Many DT learners are information-gain based, but the WinMine toolkit uses a Bayesian scoring criterion described in Chickering et al. (1997) with normal-Wishart parameter priors used for continuously valued features.

Token	*DTR*	*Threshold*	*Occurrence #49*
upon	1	> 0.003111	0 → 6
whilst	2	< 0.000516	1 → 0
on	3	< 0.004312	16 → 7
powers	4	< 0.002012	2 → 2
there	5	> 0.002911	2 → 5
few	6	< 0.000699	1 → 2
kind	7	> 0.001001	0 → 2
consequently	8	< 0.000513	1 → 0
wished	9	> 0.000434	1 → 0
although	10	< 0.000470	0 → 0

Table 2 : Top 10 DTR Rank ordered terms with threshold and corresponding occurrence count (original document → obfuscated version) for one of the disputed documents (#49).

our work is exploratory, we did not pursue optimizations along these lines.

For our first set of experiments, we applied DTR ranking starting with all of the features (8674 tokens from the training set) and repeated until the DT was unable to create a tree that performed better than the baseline of p(Hamilton) = 78.46%. In this fashion, we obtained an ordered list of 2477 terms, the top 10 of which are shown in Table 2, along with the threshold and bias. The threshold value is read directly from the DT root node and the bias (which indicates whether we desire the feature value to be above or below the threshold) is determined by selecting the branch of the DT which has the highest ratio of non-*A* to *A* documents.

Initially, this list looks promising, especially since known discriminating words like "upon" and "whilst" are the top two ranked terms. However, when we applied the changes to our baseline attribution model (described in detail in the Evaluation section), we discovered that while it performed well on some test documents, others were left relatively unscathed. This is shown in Figure 1 which graphs the confidence in assigning the authorship to Madison for each disputed document as each feature is adjusted. We expect the confidence to start high on the left side and move downward as more features are adjusted. After adjusting all of the identified features, half of the documents were still assigned to Madison (i.e., confidence > 0.50).

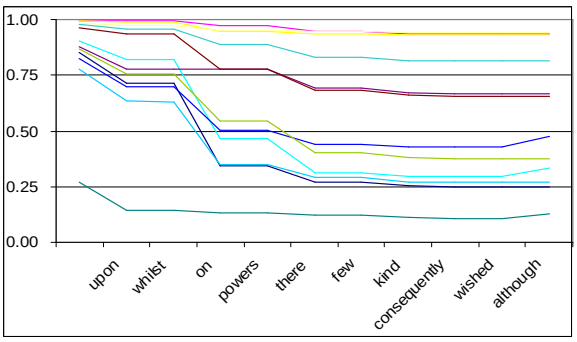

Figure 1 : Confidence in assigning disputed papers to Madison graphed as each feature is adjusted. Each line corresponds to one of the 12 disputed documents. Features are ordered by DTR Rank and the attribution model is SVM30. Values above 0.5 are assigned to Madison and those below 0.5 are assigned to Hamilton.

Choosing just the high-frequency terms was also problematic since most of them were not considered to be discriminating by DTR ranking (see Table 3). The lack of DTR rank not only means that these are poor discriminators, but it also means that we do not have a threshold value to drive the feature adjustment process.

Token	*DTR*	*Frequency*	*Token*	*DTR*	*Frequency*
the	-	0.094227	it	-	0.013404
,	595	0.068937	is	-	0.011873
of	-	0.063379	which	-	0.010933
to	39	0.038404	as	-	0.008811
.	-	0.027977	by	58	0.008614
and	185	0.025408	;	57	0.007773
in	119	0.023838	this	575	0.007701
a	515	0.021446	would	477	0.007149
be	-	0.020139	have	-	0.006873
that	-	0.014823	or	-	0.006459

Table 3 : Top 20 terms sorted by frequency.

We next combined the DTR and the term frequency approaches by computing DTR one the set of features whose frequency exceeds a specified threshold for any one of the authors. Selecting a frequency of 0.001 produces a list of 35 terms, the first 14 of which are shown in Table 4.

Token	*Frequency*	*Threshold*	*Δ 49*
upon	0.002503	> 0.003111	+6
on	0.004429	< 0.004312	-9
powers	0.001485	< 0.002012	0
there	0.002707	< 0.002911	+3
to	0.038404	> 0.039071	+7
men	0.001176	> 0.001531	+1
;	0.007773	< 0.007644	0
by	0.008614	< 0.008110	-2
less	0.001176	< 0.001384	-1
in	0.023838	> 0.023574	+6
at	0.002990	> 0.003083	0
those	0.002615	> 0.002742	+4
and	0.025408	< 0.025207	-1
any	0.002930	> 0.003005	+2

Table 4 : Top 14 DTR(0.001) ranked items. The last column is the number of changes required to achieve the threshold frequency for document #49.

Results for this list were much more promising and are shown in Figure 2. The confidence of attributing authorship to Madison is reduced by an average of 84.42% (σ = 12.51%) and all of the documents are now correctly misclassified as being written by Hamilton.

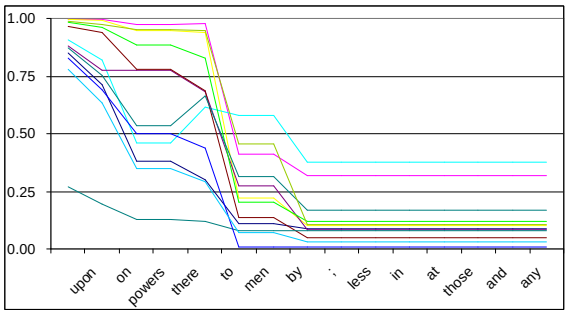

Figure 2 : Confidence in assigning disputed papers to Madison graphed as each feature is adjusted. Feature order is DTR(0.001) and the attribution model is SVM30.

5 Evaluation

Evaluating the effectiveness of any authorship obfuscation approach is made difficult by the fact that it is crucially dependent on the authorship detection method that is being utilized. An advantage of using the Federalist Papers as the test data set is that there are numerous papers documenting various methods that researchers have used to identify the authors of the disputed papers.

However, because of differences in the exact data set[2] and machine learning algorithm used, it is not reasonable to create an exact and complete implementation of each system. For our experiments, we used only the standard Federalist Papers documents and tested each feature set using linear-kernel SVMs, which have been shown to be effective in text categorization (Joachims 1998). To train our SVMs we used a sequential minimal optimization (SMO) implementation described in (Platt 1999).

The SVM feature sets that we used for the evaluation are summarized in Table 5.

For the early experiments described in the previous section we used SVM30, which incorporates the final set of 30 terms that Mosteller & Wallace used for their study. As noted earlier, they made use of a different data set than we did, so we did expect to see some differences in the results. The baseline model (plotted as the left-most column of points in Figure 1 and Figure 2) assigned all of the disputed papers to Madison except one[3].

SVM70	(Mosteller & Wallace 1964)	70 common function words.[4]
SVM30	(Mosteller & Wallace 1964)	Final 30 terms.[5]
SVM11	(Tweedie, Singh & Holmes 1996)	on, upon, there, any, an, every, his, from, may, can, do
SVM08	(Holmes & Forsyth 1995)	upon, both, on, there, whilst, kind, by, consequently
SVM03	(Bosch & Smith 1998)	upon, our, are

Table 5 : Summary of feature words used in other Federalist Papers studies.

5.1 Feature Modification

Rather than applying the suggested modifications to the original documents and regenerating the document feature vectors from scratch each time, we simplified the evaluation process by adjusting the feature vector directly and ignoring the impact of the edits on the overall document probabilities. The combination of insertions and deletions results in the total number of words in the document being increased by an average of 19.58 words (σ = 7.79), which is less than 0.5% of the document size. We considered this value to be small enough that we could safely ignore its impact.

Modifying the feature vector directly also allows us to consider each feature in isolation, without concern for how they might interact with each other (e.g. converting whilst→while or rewriting an entire sentence). It also allows us to avoid the problem of introducing rewrites into the document with our distinctive stylometric signature instead of a hypothetical Madison rewrite.

5.2 Experiments

We built SVMs for each feature set listed in Table 5 and applied the obfuscation technique described above by adjusting the values in the feature vector by increments of the single-word probability for each document. The results that we obtained were the same as observed with our test model – all of the models were coerced to prefer Hamilton for each of the disputed documents.

[2] Mosteller & Wallace and some others augmented the Federalist Papers with additional document samples (5 Hamilton and 36 Madison), but this has not been done universally by all researchers.

[3] Document #55. However, this is not inconsistent with Mosteller &Wallace's results: "Madison is extremely likely […] to have written all the disputed Federalists […] with the possible exception of No. 55. For No. 55 our evidence is relatively weak […]." (Mosteller & Wallace 1964) p.263.

[4] *ibid* p.38.

[5] *ibid* p.66.

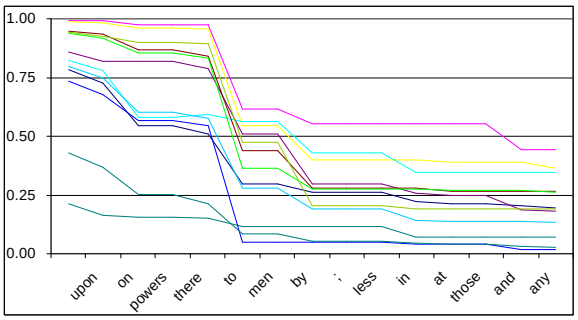

Figure 3 : Confidence in assigning disputed papers to Madison graphed as each feature is adjusted. Feature order is DTR(0.001) and the attribution model is SVM70.

Figure 3 shows the graph for SVM70, the model that was most resilient to our obfuscation techniques. The results for all models are summarized in Table 6. The overall reduction achieved across all models is 86.86%.

	% Reduction	σ
SVM70	74.66%	12.97%
SVM30	84.42%	12.51%
SVM11	82.65%	10.99%
SVM08	93.54%	4.44%
SVM03	99.01%	0.74%

Table 6 : Percent reduction in the confidence of assigning the disputed papers to Madison for each of the tested feature sets.

Of particular note in the results are those for SVM03, which proved to be the most fragile model because of its low dimension. If we consider this case an outlier and remove it from study, our overall reduction becomes 83.82%.

5.3 Feature Changes

As stated earlier, an important aspect of any obfuscation approach is the number of changes required to effect the mis-attribution. Table 7 summarizes the absolute number of changes (both insertions and deletions) and also expresses this value related to the original document size. The average number of changes required per 1000 words in the document is 14.2. While it is difficult to evaluate how much effort would be required to make each of these individual changes, this value seems to be within the range that a motivated person could reasonably undertake.

More detailed summaries of the number of feature changes required for single document (#49) are given in Table 2 and Table 4.

By calculating the overall number of changes required, we implicitly consider insertions and deletions to be equally weighted. However, while deletion sites in the document are easy to identify,

Document	Changes	Doc Size	Changes/1000
49	42	3849	10.9
50	46	2364	19.5
51	67	4039	16.6
52	52	3913	13.3
53	62	4592	13.5
54	53	4246	12.5
55	52	4310	12.1
56	59	3316	17.8
57	60	4610	13.0
58	54	4398	12.3
62	78	5048	15.5
63	91	6429	14.2

Table 7 : Changes required per document

proposing insertion sites can be more problematic. We do not address this difference in this paper, although it is clear that more investigation is required in this area.

6 Deep Obfuscation

The techniques described above result in what we term *shallow* obfuscation since they focus on a small number of features and are only useful as a defense against standard attribution attacks. More advanced attribution techniques, such as that described in (Koppel and Schler 2004) look deeper into the author's stylometric profile and can identify documents that have been obfuscated in this manner.

Koppel and Schler introduce an approach they term "unmasking" which involves training a series of SVM classifiers where the most strongly weighted features are removed after each iteration. Their hypothesis is that two texts from different authors will result in a steady and relatively slow decline of classification accuracy as features are being removed. In contrast, two texts from the same author will produce a relatively fast decline in accuracy. According to the authors, a slow decline indicates deep and fundamental stylistic differences in style - beyond the "obvious" differences in the usage of a few frequent words. A fast decline indicates that there is an underlying similarity once the impact of a few superficial distinguishing markers has been removed.

We repeated their experiments using 3-fold cross-validation to compare Hamilton and Madison with each other and the original (D) and obfuscated (D') documents. The small number of documents required that we train the SVM using the 50 most frequent words. Using a larger pool of feature words resulted in unstable models, especially when comparing Madison (14 documents) with D and D' (12 documents). The results of this comparison are shown in Figure 4.

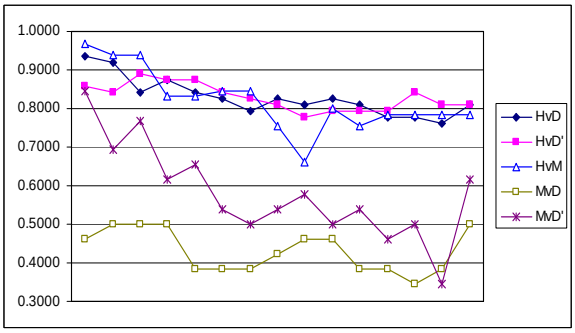

Figure 4 : Unmasking the obfuscated document. The y-axis plots the accuracy of a classifier trained to distinguish between two authors; the x-axis plots each iteration of the unmasking process. The top three lines compare Hamilton (H) versus Madison (M), the original document (D) and the obfuscated document (D'). The bottom line is M vs. D and the middle line is M vs. D'.

In this graph, the comparison of Hamilton and the modified document (MvD') exhibits the characteristic curve described by Koppel and Schler, which indicates that the original author can still be detected. However, the curve has been raised above the curve for the original document which suggests that our approach does help insulate against attacks that identify deep stylometric features.

Modifying additional features continues this trend and raises the curve further. Figure 5 summarizes this difference by plotting the difference between the accuracy of the HvD' and MvD' curves for documents at different levels of feature modification. An ideal curve in this graph would be one that hugged the x-axis since this would indicate that it was as difficult to train a classifier to distinguish between M and D' as it is to distinguish between H and D'. In this graph, the "0" curve corresponds to the original document, and the "14" curve to the modified document shown in Figure 4. The "35" curve uses all of the DTR(0.001) features.

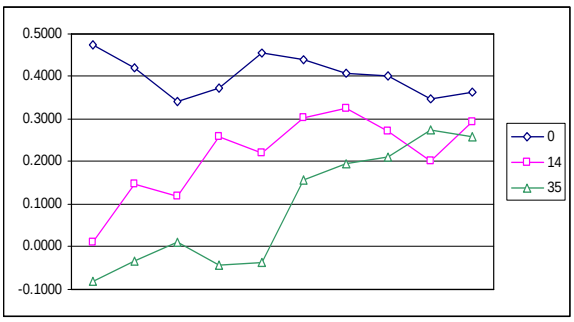

Figure 5 : Overall impact of feature modification for different levels of obfuscation. The y-axis plots the accuracy delta between the HvD' and MvD' curves; the x-axis plots each iteration of the unmasking process. The legend indicates the number of features modified for each curve.

This graph demonstrates that using DTR ranking to drive feature adjustment can produce documents that are increasingly harder to detect as being written by the author. While it is unsurprising that a deep level of obfuscation is not achieved when only a minimal number of features are modified, this graph can be used to measure progress so that the author can determine enough features have been modified to achieve the desired level of anonymization. Equally unsurprising is that this increased anonymization comes at an additional cost, summarized in Table 8.

Num Features	Changes/1000
7	9.9
14	14.2
21	18.3
28	22.5
35	25.1

Table 8 : Relationship between number of features modified and corresponding changes required per 1000 words.

While in this work we limited ourselves to the 35 DTR(0.001) features, further document modification can be driven by lowering the DTR probability threshold to identify additional terms in an orderly fashion.

7 Conclusion

In this paper, we have shown that the standard approaches to authorship attribution can be confounded by directing the author to selectively edit the test document. We have proposed a technique to automatically identify distinctive features and their frequency thresholds. By using a list of features that are both frequent and highly ranked according to this automatic technique, the amount of effort required to achieve reasonable authorship obfuscation seems to be well within the realm of a motivated author. While we make no claim that this is an easy task, and we make the assumption that the author has undertaken basic preventative measures (like spellchecking and grammar checking), it does not seem to be an onerous task for a motivated individual.

It not surprising that we can change the outcome by adjusting the values of features used in authorship detection. Our contribution, however, is that many of the important features can be determined by simultaneously considering term-frequency and DTR rank, and that this process results in a set of features and threshold values that are transparent and easy to control.

Given this result, it is not unreasonable to expect that a tool could be created to provide feedback to an author who desires to publish a document anonymously. A sophisticated paraphrase tool could theoretically use the function word change information to suggest rewrites that worked toward the desired term frequency in the document.

For our experiments, we used a simplified model of the document rewrite process by evaluating the impact of each term modification in isolation. However, modifying the document to increase or decrease the frequency of a term will necessarily impact the frequencies of other terms and thus affect the document's stylometric signature. Further experimentation is clearly needed in this area needs to address the impact of this interdependency.

One limitation to this approach is that it applies primarily to authors that have a reasonably-sized corpus readily available (or easily created). However, for situations where a large corpus is not available, automated authorship attribution techniques are likely to be less effective (and thus obfuscation is less necessary) since the number of possible features can easily exceed the number of available documents. An interesting experiment would be to explore how this approach applies to different types of corpora like email messages.

We also recognize that these techniques could be used to attempt to imitate another author's style. We do not address this issue other than to say that our thresholding approach is intended to push feature values just barely across the threshold away from *A* rather than to mimic any one particular author.

Finally, in these results, there is a message for those involved in authorship attribution: simple SVMs and low-dimensional models (like SVM03) may appear to work well, but are far less resilient to obfuscation attempts than Koppel and Schler's unmasking approach. Creating classifiers with the minimum number of features produces a model that is brittle and more susceptible to even simplistic obfuscation attempts.

8 Acknowledgements

Thanks are in order to the reviewers of earlier drafts of this document, notably Chris Brockett and our anonymous reviewers. In addition, Max Chickering provided useful information regarding his implementation of DTs in the WinMine toolkit.

References

R. A. Bosch and J. A. Smith. 1998. Separating Hyperplanes and the Authorship of the Federalist Papers. *American Mathematical Monthly*, Vol. 105 #7 pp. 601-608.

D. M. Chickering, D. Heckerman and C. Meek. 1997. A Bayesian Approach to Learning Bayesian Networks with Local Structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (UAI97 Providence, RI), pp. 80-89.

D. M. Chickering. 2002. The WinMine Toolkit. Technical Report MSR-TR-2002-103.

D. I. Holmes and R. S. Forsyth. 1995. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing* 10(2), pp.111-127.

D. I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13(3), pp.111-117.

T. Joachims. 1998. Text Categorization with Support Vector Machines: Learning with many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning*, pp.137-142.

M. Koppel and J. Schler. 2003. Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* (Acapulco, Mexico). pp.69-72.

M. Koppel and J. Schler, 2004. Authorship Verification as a One-Class Classification Problem. In *Proceedings of the Twenty-First International Conference on Machine Learning* (ICML 04 Banff, Alberta, Canada), pp.489-495.

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

J. Platt. 1999. Fast Training of SVMs Using Sequential Minimal Optimization. In B. Schölkopf, C. Burges and A. Smola (eds.) *Advances in Kernel Methods: Support Vector Learning*. MIT Press (Cambridge, MA, USA), pp.185-208.

J. R. Rao and P. Rohatgi. 2000. Can Pseudonymity Really Guarantee Privacy?, In *Proceedings of the 9th USENIX Security Symposium* (Denver, Colorado, USA), pp.85-96.

F. J. Tweedie, S. Singh and D. I. Holmes. 1996. Neural Network Applications in Stylometry: The Federalist Papers. In *Computers and the Humanities* 30(1), pp.1-10.