Practical Attacks Against Authorship Recognition Techniques

Michael Brennan and **Rachel Greenstadt**
Dept. of Computer Science
Drexel University
3175 JFK Blvd Room 140
Philadelphia, PA 19104
tel 215.895.2920
fax 215.895.0545
{mb553,greenie}@cs.drexel.edu

Abstract

The use of statistical AI techniques in authorship recognition (or stylometry) has contributed to literary and historical breakthroughs. These successes have led to the use of these techniques in criminal investigations and prosecutions. However, few have studied adversarial attacks and their devastating effect on the robustness of existing classification methods. This paper presents a framework for adversarial attacks including **obfuscation attacks**, where a subject attempts to hide their identity and **imitation attacks**, where a subject attempts to frame another subject by imitating their writing style. The major contribution of this research is that it demonstrates that both attacks work very well. The obfuscation attack reduces the effectiveness of the techniques to the level of random guessing and the imitation attack succeeds with 68-91% probability depending on the stylometric technique used. These results are made more significant by the fact that the experimental subjects were unfamiliar with stylometric techniques, without specialized knowledge in linguistics, and spent little time on the attacks. This paper also provides another significant contribution to the field in using human subjects to empirically validate the claim of high accuracy for current techniques (without attacks) by reproducing results for three representative stylometric methods.

Introduction

The field of stylometry (or authorship recognition) has been used to great effect by historians and literary detectives to identify the authors of the Fedaralist Papers, Civil War letters, and Shakespeare's plays (Klarreich 2003; Oakes 2004). While stylometric methods existed before computers and artificial intelligence techniques, the field is currently dominated by AI techniques such as neural networks and statistical pattern recognition. In many historical matters, authorship has been unintentially lost to time and it can be assumed that the authors did not have the knowledge or inclination to attempt to hide their linguistic style. However, this may not be the case for modern authors who wish to hide their identity. For example, stylometric techniques are currently used as evidence in courts of law in Britain, the U.S., and Australia (Morton and Michaelson 1996).

> In some criminal, civil, and security matters, language can be evidence...When you are faced with a suspicious document, whether you need to know who wrote it, or if it is a real threat or a real suicide note, or if it is too close for comfort to some other document, you need reliable, validated methods. (Institute for Linguistic Evidence 2008)

This study seeks to discover how robust current methods of stylometry are in dealing with adversarial attacks. Until now the field has focused on creating new methods that attempt to classify existing unknown works sets of authors, with little attention being given to the question of what happens when an adversary tries to intentionally circumvent the classification system that has been established. Recently attention has been brought to this problem. Dr. Patrick Juola, an expert in computer linguistics at Duquesne University, discussed the importance of research in this area in his 2008 book on authorship attribution, stating "there is obviously great potential for further work here" (Juola 2008).

This study looks at three specific approaches and their resilience against two types of adversarial attacks. The first, which will be referred to as an **obfuscation attack**, is when an author attempts to write a document in such a way that their personal writing style will not be recognized. The second, which will be referred to as an **imitation attack**, is when an author attempts to write a document such that their writing style will be recognized as that of another specific author. The three methods of stylometry investigated were chosen for their variety in both metrics and methodology. The study found that none of the methods performed better than chance in identifying the correct author in either of these attacks. Additionally, the imitation attacks were widely successful, attributing authorship of the passages to the intended victim of the attack in most instances.

This study also provides a clear validation of three prominent stylometric methods using a significantly larger data set than former studies. Furthermore, our study involves more original authors than most other stylometric studies. Our results thus provide further evidence of their effectiveness in non-adversarial scenarios. Our experimental design provides a framework for identifying adversarial attacks and analyzing the adversarial risk of stylometric methods. Finally, we show that even naive users lacking in expertise in the field of stylometry, linguistics, or even literature can successfully perform imitation and obfuscation attacks.

The findings of this research indicate that the studied techniques are insufficient to determine authorship in an adversarial context. We call upon the research community to refine our framework for testing methods of stylometry against adversarial attacks so it can play a role in determining the effectiveness of existing and future work in the field.

Background & Related Work

The classic example case in the field of Stylometry is that of the Federalist papers. 85 papers were published anonymously in the late 18th century to persuade the people of New York to ratify the American Constitution. The authorship of 12 of these papers was heavily contested (Oakes 2004). To discover who wrote the unknown papers, researchers have analyzed the writing style of the known authors and compared it to that of the papers of unknown authorship. The features used to determine writing styles have been quite varied. Original attempts looked at the length of words, whereas later attempts looked at pairs of words, vocabulary usage, sentence structure, function words, and so on. Most studies show the author was James Madison.

It is important to note that stylometry is not handwriting analysis; it only looks at the linguistic style of the text.

There are some interesting hurdles in the field of Stylometry that, while far from unique, certainly are not common. One such aspect is the fact that the writing style of individuals tends to change over time, and compensating for that is a difficult task. Over time, the entire body of test data for a specific author slowly becomes dated and less reliable.

There is still no consensus on what method of stylometry is the most effective or even which features are the most important. The only true consensus is that the method chosen should reflect the text that is going to be classified. For example, counting the number of semicolons might be appropriate for classifying the author of code but not for classifying essays written by children in elementary school.

A number of resources are available that give an overview of stylometry methods that exist (Malyutov 2006; Uzuner and Katz 2005b), and focus on the state of the field as it relates to computer science and computer linguistics (Juola 2008). Artificial Intelligence has been embraced in the field of stylometry, leading to more robust classifiers using machine learning and other AI techniques (Tweedie, Singh, and Holmes 1996; Holmes and Forsyth 1995; Uzuner and Katz 2005a). There has also been some work on circumventing attempts at authorship attribution (Josyula and Pankaj 2006; Kacmarcik and Gamon 2000) and looking at stylometry as a method of communication security (Calix et al. 2008), but these do not deal with malicious attempts to circumvent a specific method. Some research has looked at imitation of authors but not in any widespread study, such as (Somers and Tweedie 2003) who compared the work of Gilbert Adair's pastiche of Lewis Carroll's Alice in Wonderland, and found mixed results.

Methodology

Countless methods of authorship attribution have been developed. Since there is no consensus on which available method is the best, three specific approaches were selected for their variety in both methodology and the features they used for comparison. Additionally, a strict process was set up to collect the samples used in this study for both the training text and the example attacks.

Study Format and Setup

The results of this study are based upon the participation of 15 individual authors. This is significantly larger than most studies in the field which generally deal with 2-4 authors (Tweedie, Singh, and Holmes 1996; Clark and Hannon 2007; Oakes 2004; Celikel and Dalkilic 2004). There were three basic elements to their participation.

First, each author had to submit approximately 5000 words of pre-existing sample writing. Each writing sample had to be from some sort of formal source, such as essays for school, reports for work, and other professional and academic correspondence. This was intended to eliminate slang and abbreviations, instead concentrating on consistent, formal writing style of everyone involved. This also helped to limit possible errors that are not a result of the malicious attack attempts but nonetheless could have an effect on the accuracy of the authorship attribution. Participants submitted anywhere from 1 to 5 sample documents which were combined and split into 500 word sample passages.

Second, each author had to perform an obfuscation attack in which they try to hide their identity through their writing style. This was accomplished by writing a new 500 word passage on a specific topic. The task given to them was to write a description of their neighborhood directed at a friend who had never been there. This task was designed to encourage thought about the writing style and not weigh down the author with having to think about a complex topic or be too creative with their content.

Third, each author had to perform an imitation attack in which they try to imitate another author's style. For this task the participants were given a 2500 word sample from *The Road* by Cormac McCarthy to model their passage after. This selection was made for a variety of reasons. Most importantly Cormac McCarthy has a distinct writing style:

> On the far side of the river valley the road passed through a stark black burn. Charred and limbless trunks of trees stretching away on every side. Ash moving over the road and the sagging hands of blind wire strung from the blackened lightpoles whining thinly in the wind.[1]

Since the participants are not linguists this would allow them to make a fair attempt at emulating another style. Furthermore the passage is an engaging read which once again encourages the reader to think more about style than forcing them to trudge through boring content. The assumption was that engaged participants yield more representative passages and thus more accurate results. The writing task given to the participant was to narrate their day from the point at which they get out of bed, and to do so using a third-person perspective. This is also similar to the events in the sample text. For testing purposes an additional 2500 words were

[1]Full excerpt at `http://www.bookbrowse.com/excerpts/index.cfm?book_number=1964`

taken from The Road and used as training text for Cormac McCarthy along with the original sample. It should also be noted that the excerpt distributed to readers is freely available as a promotional passage from the book.

Asking the participants to attempt the obfuscation attack before the imitation attack was intentional. We were concerned that if participants chose to do the imitation attack first then all of the obfuscation attacks would simply read as a second Cormac McCarthy imitation attack.

Method 1: Statistical Method using the Signature Stylometric System

Method 1 analyzes the text, then uses a basic statistical method for comparison. The features used for the analysis are word lengths, letter usage, and punctuation. The method of authorship attribution compares a sample text with each training corpus (one per author, consisting of all sample texts from that author) and sums the Chi-square result of the two texts. The higher the value, the less likely it is that the author of the sample piece and the author of the reference text are the same individual. The minimum of the sum of the Chi-Square values for each comparison pair is then selected as the author of the sample text.

The reliability of this method was confirmed by randomly selecting a sample text from each author in each set of authors and classifying it amongst that set. Signature is an established platform for author recognition but the current version unfortunately lacks the ability to automate testing. Because of this, all testing had to be performed manually and led to less robust testing since fewer random samples could be tested for each group than in the other methods. We feel, however, that these results combined with the off-the-shelf nature of the software correctly reflect the strength of the Signature system. On average this method correctly identified the original author 95% of the time.

Method 2: Neural Network Approach

Method 2 is adapted from the approach outlined in Neural Network Applications in Stylometry (Tweedie, Singh, and Holmes 1996). This approach is one of the first and most widely used approaches of combining Stylometry with the field of Artificial Intelligence. The method was developed and tested using the classic Federalist Papers. The features previously used in this system are function words that were hand-picked after comparing each of the known Federalist Papers to discover which words would be the most effective for classification. The function word feature set was not used in this study due to the wide range of topics in the training data. In the original method the use of the Federalist Papers allowed for many assumptions to be made about the content, but here—as in the real world—this luxury does not exist. Instead a series of nine features were used which could be automatically generated from each sample text: number of different words, lexical density, Gunning-Fog readability index, character count without whitespace, average syllables per word, sentence count, average sentence length, and an alternative readability measure. These metrics were automatically generated using Textalyser, a freely available text analysis tool[2]. The tool also ignored the numbers in the text, applied a stop list on common words, and ignored words less than 3 characters in length.

The effectiveness of Method 2 was confirmed through repeated random sub-sampling validation. We randomly splitting the sample texts from the authors in each group, using 90% for training and 10% for testing. A total of 16 random trials were conducted for each group giving an overall average accuracy of 78.5%.

Method 3: Synonym-Based Classifier

The final method is based on the vocabulary of the author. This method uses three different models created and evaluated by (Clark and Hannon 2007). The authors found the most effective model to be the second, giving them 94% to 98% accuracy depending on the number of authors, but results in this study showed a hybrid of models 1 and 2 to be the best approach for the training text, giving us 89% to 99% accuracy for the same numbers of authors. This accuracy was measured through leave-one-out-cross-validation.

The general approach of this method is to examine how each author chooses synonyms. The theory behind the method is that when a word has a large number of synonyms to choose from, the choice the author makes is significant in understanding his or her writing style.

A feature vector is created for each word w in a text, having two elements: the number of synonyms s that the word has according to Princeton's WordNet lexical database (Miller 1995), and the shared frequency n of the word w between the sample text and the training text of a known author. The match value for a sample text u from an unknown author and a reference text k from a known author is then the sum of $n * s$ for all shared words between the two texts. Authorship is attributed to a text based on the known author with the highest match value to the sample text.

Model 2 expands upon Model 1 in two ways; it first adds a stop word list based on the 319 most common words in the English language according to the Glasgow University Information Retrieval Group[3]. It also takes into account the overall frequency of a word in all of the available text and uses it to help determine the match value. We found the former change to be beneficial while the latter was not. The resulting accuracy of Model 1 plus the stop word list was 91.67% on average. The accuracy of the method on the training set was accomplished by comparing each individual sample passage against the rest of the entire set of test passages for all authors, minus the one being tested. This test was repeated exhaustively for each sample passage in each set of authors.

Since the training text for each author was split up into 500 word passages there were two methods of classifying authors. A sample text could be attributed to the author of any single passage with the highest match value, or it could be attributed to the author with the highest match value over

[2] www.textalyser.net

[3] www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils

all samples from that author. It was decided to use an average of the two methods.

Deployed Applications Used

A number of tools were used in implementing the three stylometry methods.

Signature Stylometric System. Signature is freeware developed by Dr. Peter Millican of Oxford University[4]. It uses a basic statistical approach for comparing texts. It is not inherently a classifier, but the similarity measurements within it were adapted for this study to be used for classification.

Textalyser. A free, web-based text analysis tool.

Weka. Weka is a collection of open source machine learning algorithms intended to be used for data mining[5]. It has an excellent neural networks model that was used to generate the neural networks for Method 2.

WordNet. Princeton has long developed and maintained this large lexical database of English[6]. It has many functions but the part used for this study is the ability to traverse synsets, or the sets of synonyms for adjectives, nouns, verbs, and adverbs.

Evaluation

The bulk of the work in the field of Stylometry looks at comparing small sets of authors (Tweedie, Singh, and Holmes 1996; Clark and Hannon 2007; Oakes 2004; Celikel and Dalkilic 2004). The relatively large number of subjects in this study enabled us to compare multiple groups of authors at various "standard" sizes. Four randomly chosen sets of 2, 3, 4, and 5 authors each were tested across all three methods of authorship recognition. The accuracy of the methods when tested against each attack was calculated by comparing the attack passages to the entire set of training texts for each author in each set. The methodology for evaluating the initial accuracy of each method has been explained in their respective sections. The accuracy measurements in all of the graphs are an average of the four sets in each group.

Figure 1 shows the strength of the various stylometric methods that have been implemented in this study. The top three lines represent the three methods used in this study and the bottom line represents the probability of selecting the correct author by random chance. It is important to note the significance of the accuracy measurements when compared to random chance as the adversarial attacks weaken all methods to the point of being equivalent or worse than random selection. With accuracies ranging from 91.5% to 100%, Method 1 had the best performance, but it is important to stress that this is not a study in comparing these stylometric methods with each other but rather against the prospect of resisting adversarial attacks. Methods 2 and 3 still performed very well, with Method 2 ranging from 73% to 90% and Method 3 ranging from 88.7% to 98.7%.

[4] www.philocomp.net/humanities/signature

[5] www.cs.waikato.ac.nz/ml/weka

[6] wordnet.princeton.edu

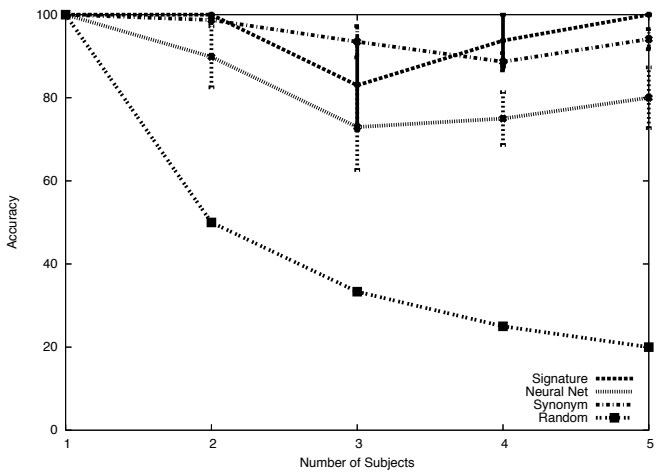

Figure 1: Accuracy of the three methods without active attack. X-axis shows the number of subjects, y-axis shows the average percentage of documents correctly classified. Error bars show the standard error for each experiment.

Not all of the methods follow a smooth degradation curve as the number of authors in each set increases. In fact, all methods jump up in accuracy between sets of 4 and 5 authors. We believe this is a function of the sets of authors used and that further testing these methods will result in a smooth degradation curve. For example, when particularly "distinct" authors were included in more 5-author sets than 4-author sets, the accuracy of those sets increased.

Figure 2 shows how each method performed against obfuscation attacks under different numbers of subjects, represented by the three bar graphs above each author count on the x-axis. The line represents the probability of selecting the correct author through random chance. In only three cases, a method did better than chance in selecting the correct author of an obfuscation attack with Method 1 achieving 62.5% and Method 3 achieving 56.3% success in sets of two authors and Method 2 achieving 25% accuracy in sets of 5 authors. There were also a few instances where the methods performed worse than chance but the overall degradation of the accuracy closely follows the probability of selecting the correct author at random. This demonstrates the weakness of these methods in resisting such attacks.

Imitation attacks were effective in circumventing all three methods, as can be seen in Figure 3 which illustrates the accuracy measurements in the same way as Figure 2. Only once did any method have higher than 10% success rate in detecting the correct author and the average accuracy was below 5% for all three. It is also interesting to note that all three methods did much worse in the face of an imitation attack than an obfuscation attack in detecting the correct author. This could be attributed to the participants in the study having more direction in creating the imitation attack which makes it easier to write in a different style.

Figure 4 shows that all three methods largely attributed authorship of the imitation attacks to the intended victim of the attack. Only once did any method classify less than 60%

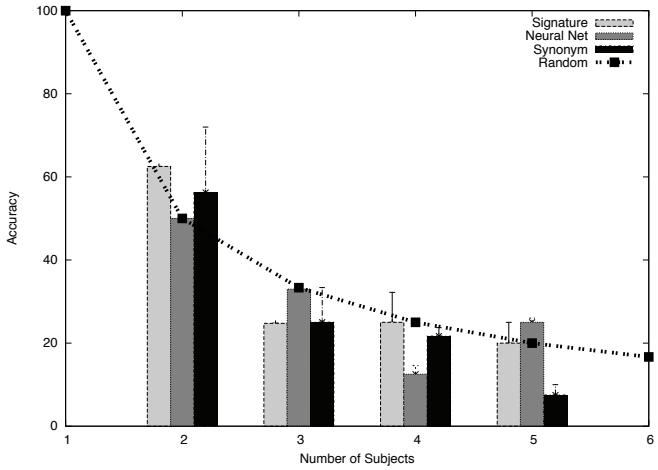

Figure 2: Accuracy in detecting obfuscation attacks. The x-axis shows the number of subjects, the y-axis shows the average percentage of obfuscation attacks correctly classified. The error bars show the standard error for each experiment.

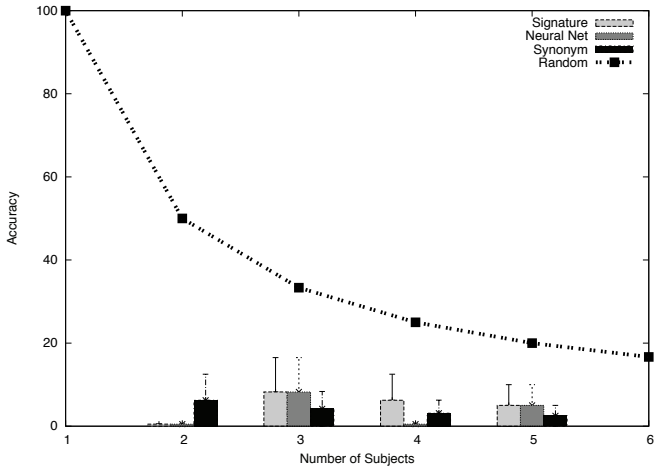

Figure 3: Accuracy in detecting imitation attacks. The x-axis shows the number of subjects, the y-axis shows the average percentage of imitation attacks correctly classified. The error bars show the standard error for each experiment.

of the imitation attacks as being authored by Cormac McCarthy. Method 3 was particularly susceptible to this attack, classifying 87.5% to 93.75% of the imitation attacks as authored by McCarthy. This result also did not appear to degrade with the addition of more authors.

We also ran a test to see how Methods 2 and 3 held up in the case eight unique authors. This is a very large sample when compared to most studies in stylometry. While this testing is not as exhaustive as the trials with 2-5 authors, it indicates a continuation of the trends that have been observed. Method 3 had a training set accuracy of 91.7% and Method 2 had a training set accuracy of 64.1%, both in line with the degradation curves shown in Figure 1. Method 2 did classify 25% of the obfuscation attacks correctly, but none of the imitation attacks. Method 3 classified only 6.25% of each of the attacks correctly, and imitation attacks were successful more than 75% of the time in both cases.

Discussion & Future Work

There are a number of design choices made in this study that warrant discussion and have implications for future work.

Training Set Content and Size. The amount of training text for each author is not exceptionally large, but the total number of authors in the study is significantly bigger than most studies of stylometry methods. Having such a large number of authors to work with allowed us to do more extensive testing by choosing random groups of authors and averaging the accuracies across them. This also allowed us to see interesting patterns in the study such as which authors did a better job than others in creating successful attacks. It was clear that certain authors did a poor job in creating their attacks and that was displayed in the average results of the groups they were in. Certain authors also had a style that seemed to be particularly susceptible to obfuscation attacks. Authorship of the attacks were often attributed to these authors when they were a member of a test set. An interesting avenue of research would be to determine if it is possible to create a "generic" writing style, though this may be better suited for the field of linguistics than computer science.

The domain of possible content in this study was fairly open. Participants were allowed to present samples from a variety of subjects so long as they were formal in nature. It could be beneficial to study the effects of adversarial attacks in stricter domains where there is less room for maneuvering and thus less options for how an author could hide his or her identity. Stylometric methods used in restricted domains may prove less susceptible to our attacks.

Participant Skill Level. In general the participants in this study were unskilled in the fields of linguistics, stylometry and literature. Despite this lack of expertise they were able to consistently circumvent the authorship attribution methods that were put in place. This strengthens our findings as it would be reasonable to expect authors with expertise in such areas could do a better job at attacking the system.

In informal discussions with participants after completing the study we found that many of them tried to obfuscate their style by "dumbing down" their writing—using shorter sentences and less descriptive words. In imitating the writing style of Cormac McCarthy the participants described attempting to use descriptive and grim language. Formally incorporating the ways in which participants attempted to change their writing style into the study could provide insight into how to better defend against these attacks.

Other Stylometic Methods. While the methods chosen for this study were picked for their effectiveness and coverage of the design space, other methods may be more resistant to adversarial attacks. A possible avenue of research is to examine which metrics and methods might offer resistance, including looking at the possibility of using ensembles of classifiers to create more effective possible methods.

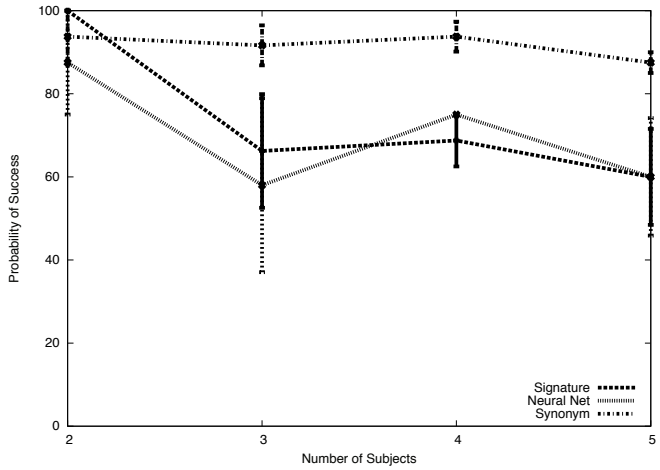

Figure 4: Success rate of imitation attacks. The x-axis shows the number of subjects, the y-axis shows the average percentage of imitation attacks that were classified as being authored by the intended victim of the attack. The error bars show the standard error for each experiment.

Conclusion

The most important conclusion to draw from this study is that we must test stylometric methods for their resistance to adversarial attacks in situations where attacks are likely. This paper looks at only a small fraction of the authorship attribution techniques that exist, but these techniques use a wide range of metrics and methodology that is representative of the field of stylometry. Yet all three methods showed significant failings when faced with adversarial attacks.

The obfuscation attacks weaken all three methods to the point that they are no better than randomly guessing the correct author of a document. The imitation attacks were widely successful in having their authorship attributed to the intended victim of the attack. In addition the attacks were generated by participants in very short periods of time with no expert knowledge in linguistics or stylometry.

This study also strengthens the original claims of high accuracies by validating the methods on a large set of new data produced for a variety of purposes. When these methods are used in situations where adversarial attacks is not considered to be a threat, they perform quite well.

Our framework for testing against adversarial attacks can be used as a basis for testing the resistance of both existing and future methods of authorship attribution against obfuscation and imitation attacks.

This study shows stylometry failing in the face of basic attacks by individuals relying solely on intuition (they have no formal training or background in authorship attribution). While the use of stylometry as a security feature is not common, it is used as forensic evidence (Morton and Michaelson 1996). Our results showing that stylometry methods are weak against trivial attacks should have an impact on how such evidence is weighed.

Does this mean the end of stylometry in sensitive areas? Certainly not. Stylometry is a massive field with a lot of research and a history of positive contributions to science and the humanities. The results of this study should bring further attention to an element of the field that has been overlooked. 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.

References

Calix, K.; Connors, M.; Levy, D.; Manzar, H.; MCabe, G.; and Westcott, S. 2008. Stylometry for e-mail author identification and authentication. *Proceedings of CSIS Research Day, Pace University*.

Celikel, E., and Dalkilic, M. E. 2004. Investigating the effects of recency and size of training text on author recognition problem. *ISCIS* 21–30.

Clark, J., and Hannon, C. 2007. A classifier system for author recognition using synonym-based features. *Lecture Notes in Computer Science* 4827:839–849.

Holmes, D., and Forsyth, R. 1995. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing* 10:111–127.

Institute for Linguistic Evidence, I. 2008. Mission & philosophy. `www.linguisticevidence.org`.

Josyula, R., and Pankaj, R. 2006. Can pseudonymity really guarantee privacy. *COLING/ACL* 444–451.

Juola, P. 2008. Authorship attribution. *Foundations and Trends in information Retrieval* 1(3):233–334.

Kacmarcik, G., and Gamon, M. 2000. Obfuscating document stylometry to preserve author anonymity. *Proceedings of the 9th USENIX Security Symposium* 9:7–7.

Klarreich, E. 2003. Bookish math. *Science News* 164(25).

Malyutov, M. 2006. Information transfer and combinatories. *Lecture Notes in Computer Science* 4123(3):233–334.

Miller, G. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38:39–41.

Morton, A., and Michaelson, S. 1996. The qsum plot. *Internal Report CSR-3-90*.

Oakes, M. 2004. Ant colony optimisation for stylometry: The federalist papers. *Proceedings of the 5th International Conference on Recent Advances in Soft Computing* 86–91.

Somers, H., and Tweedie, F. 2003. Authorship attribution and pastiche. *Computers and the Humanities* 37:407–429.

Tweedie, F.; Singh, S.; and Holmes, D. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities* 30(1):1–10.

Uzuner, U., and Katz, B. 2005a. Capturing expression using linguistic information. In *AAAI*.

Uzuner, U., and Katz, B. 2005b. A comparative study of language models for book and author recognition. In *IJCNLP*, 969.