Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 4th, 2012

Evaluation of a Stylometry System on Various Length Portions of Books

Ida Schulstad, Mark Boga, Cranston Jordan, Kara Pally,
John Monaco, Richard DeStefano, John Stewart, and Charles Tappert
Pace University, Seidenberg School of CSIS, White Plains, NY USA
{*ida.schulstad,mark.boga,cranston.s.jordan,kl07502n,john.v.monaco,richard.j.destefano*}*@pace.edu*

Abstract

The Pace University Stylometry Biometric System was developed as an extension of a previously developed keystroke biometric system. Earlier experiments on short texts taken from online test-takers gave poor results on the Pace system. This study tests the system on text input ranging from 250 to 10,000 words to obtain the accuracy as a function of text length. 10 samples each from 30 authors were collected for a total of 300 texts, all published between 1880 and 1930. The results show that performance as measured by the Equal Error Rate generally improves with increased text length, and that the best results are achieved with fewer subjects and "strong" training.

1. Introduction

The main goal of this study is to test the Pace University Stylometry Biometric System (PSBS) on text samples of various lengths. A previous study of the PSBS on online tests showed poor results, but improved with longer text samples [2]. This study therefore tests the PSBS on samples of 11 different lengths - 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 4000, 5000, and 10,000 words, based on the hypothesis that accuracy will increase as a function of sample length.

The PSBS was developed as an extension of a Pace Keystroke Biometric System (PKBS) previously described by Tappert et al and Zack et al [14, 15]. The motivation for introducing stylometry came from the use of the keystroke biometric system for authenticating online test takers in response to the requirement of the 2008 federal Higher Education Opportunity Act that institutions make efforts to control the identity of students accessing the system and taking exams in online courses [11]. However, while the correct student may be typing, the answers could be provided by somebody else. Although stylometry was thought to be a useful addition to the system to ensure that students were not coached, the addition of PSBS proved unsuccessful with short-answer online tests [11]. The results of the current study show the accuracy of the PSBS on longer text samples and may suggest the appropriate context for future use of the PSBS as a plagiarism detection tool. Although stylometric comparisons are likely to fail in scenarios where texts are strongly paraphrased to the point that they more closely resemble the personal writing style of the plagiarist, it has been shown by Stein et al [12] that stylometric analysis to detect plagiarism works reliably for document lengths of several thousand or tens of thousands of words.

2. Review of Stylometry Features

Stylometry is defined as "the statistical analysis of variations in literary style" [13]. The underlying assumption of stylometry is that writers have unique habits of vocabulary, sentence length, and other features, and that these habits are unconscious and difficult to mask [2]. These unique writing features are measured to create an author profile against which other texts can be compared. Grieve [5] summarizes some important events in the history of stylometry, which was first used to attribute an author to a disputed or anonymous text, one of the first examples being Edmond Malone's 1787 study of *Henry VII*. In this study, he used meter and rhyme to show that the text did not match Shakespeare's usual style. Shakespeare has since been a popular subject of stylometric studies using a variety of characteristic features. Another famous study is the 1964 study of *The Federalist Papers*, a series of essays published in *The Independent Journal* from 1787 to 1788, which were known to be written by James Madison, Alexander Hamilton, and John Jay, but the exact author of each individual paper was unknown. Mosteller and Wallace used the frequency of common words as a characteristic feature to attribute the disputed texts to Madison [5, 3, 16].

However, stylometry has not only been used for literary and historical purposes – it also has forensic applications. An early example is the Reverend Andrew Morton's use in the 1970s of collocations (sequences of words) to determine that a confession was written by multiple authors, i.e. that it was forged, which convinced the court to drop the charges [5].

More recent studies have used stylometry to determine the authorship of e-mails and online messages to counteract cybercrime [3, 16]. In addition to identifying an author, stylometry can also be used to

detect multiple authors in a text (plagiarism) or to assign an author to a sociolinguistic category such as gender. For example, Corney [3] uses various studies on gender differences in linguistic choices to compile a series of features designed to assign an author to a gender category, including the frequency of *sorry* words, as female writers had been found to be more likely to apologize. A slightly different use of stylometry is described in Lex et al [8], which looks at the use of stylometric and lexical features to determine the genre of a blog (news or non-news), and to determine how the author feels about the events (sentiment analysis, see also Abbasi et al [1]).

There is little agreement regarding which features are most useful for identifying purposes, and a summary in 1998 concluded that more than 1000 different writing features had been used in authorship analysis [16]. Such features may include lexical, syntactic, structural, content-specific, and idiosyncratic features.

Lexical features include vocabulary size and character-based features, for example the number of vowel characters compared to the total number of characters. Syntactic features are related to sentence structure, and include for example the use of function words. Structural features are related to how a text is structured, and include features such as paragraph length and use of indentation. Content-specific features include the use of content-specific words. Idiosyncratic features include spelling and grammatical errors [2, 16].

A 2005 study [5] (see also Grieve [6]) evaluates some of the commonly used features for author attribution, including several ways of measuring word-length, sentence length, vocabulary richness, graphemes (defined as one of the letters of the English alphabet), word frequencies, punctuation, word positions (e.g. the relative frequencies of words occurring in particular positions in the sentences), collocations (a sequence of two or more words), and n-grams (defined as a sequence of two or more characters). The study concludes that the most successful features are measurements of function word frequencies (for example prepositions and articles), punctuation marks, and two-and three-grams, where the word and punctuation profile outperforms the n-grams. The author hypothesizes that this is because function words and punctuation marks reveal how sentences are constructed, whereas n-grams are more influenced by the content of the text. For example, the use of the function word "or" says something about an author's preference for conjoined sentences. Counting the two-gram (or digram) o-r will pick up the use of the word "or", but its frequency will also be influenced by words such as "neighbor" and "labor", which are more related to content than personal style. On the other hand, n-grams may pick up on such characteristics as a preference for past tense through the frequency of the e-d two-gram.

Other studies have used similar features, for example Corney [3], who, in addition to two-grams, function word frequency, and punctuation, also includes several character-based features and word-length measurements. Corney finds that the two-gram and function word features are the most successful, although the best result is obtained when combining all features sets, a strategy also supported by Grieve [5, 6] and Zheng et al [16].

3. Pace Stylometry Biometric System

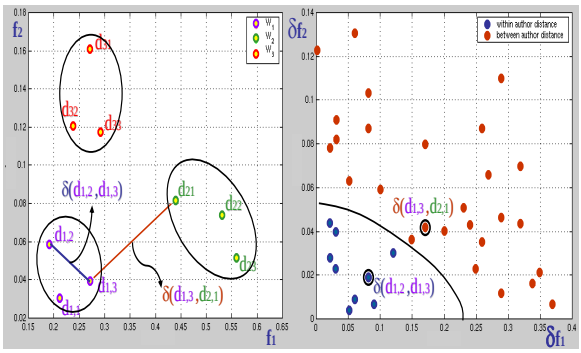
The Pace Stylometry Biometric System (PSBS) consists of a feature extractor, an authentication classifier, and a Receiver Operating Characteristic (ROC) curve generator. The input to the feature extractor is an XML file of text data from various authors.

3.1 Feature Extractor

This study used 228 features (49 character-based, 13 word-based, and 166 syntactic), listed in Appendix A. These features are based on features used in Zheng et al [16] and character-based features used in the keystroke biometric system [2]. The character-based measurements include several different one-gram measurements, such as the relative frequency of "o" to the other vowel characters, and several different two-gram measurements, such as the relative frequency of "e-a" to the number of total vowel-vowel two-grams. The word-based features measure the word-length, the vocabulary size, the frequency of words occurring once (Hapax Legomena), and the frequency of words occurring twice. It is important to note that the method used for measuring vocabulary size, the ratio of different words to the total number of words, is sensitive to sample size, and would not be appropriate for comparing samples of different lengths. The syntax-based features include punctuation marks, special characters, function words, other word frequency measurements (e.g. common adjectives/number of words), and sentence length.

3.2 Authentication Classifier

The PSBS uses the same classification system as the Pace University Keystroke Biometric System described in Stewart et al [11], Yoon et al [17], Tappert et al [14], and Zack et al [15] and summarized here. A multi-class problem is transformed into a two-class problem, resulting in the two classes within-class (intra-person) (authenticated), and between-class (inter-person) (not authenticated) (Fig. 1). This method has been shown to be especially effective for multidimensional feature-space problems [17].



(a) Feature space (b) Feature-difference space

Figure 1. Transformation from feature space (a) to feature distance space (b), adapted from [10].

Yoon et al explain the dichotomy transformation process as follows: take an example of three people $\{P_1,P_2,P_3\}$ where each person supplies three biometric samples. Figure 2 (a) plots the biometric sample data for these three people in two-dimensional feature space. This feature space is transformed into a feature-difference space by calculating vector distances between pairs of samples of the *same* person (*intra-person distances*, denoted by $x_{\oplus}$) and distances between pairs of samples of *different* people (*inter-person distances*, denoted by $x_{\varnothing}$). Let d_{ij} represent the individual feature vector of the i^{th} person's j^{th} biometric sample, then $x_{\oplus}$ and $x_{\varnothing}$ are calculated as follows [17]:

$$x_{\oplus} = |d_{ij} - d_{ik}| \text{ where } i=1 \text{ to } n, \text{ and } j,k=1 \text{ to } m, j \neq k$$
$$x_{\varnothing} = |d_{ij} - d_{kl}| \text{ where } i,k=1 \text{ to } n, i \neq k \text{ and } j,l=1 \text{ to } m \quad (1)$$

where n is the number of people, m is the number of samples per person, and the absolute value is of the elements of these vectors. Figure 2 (b) shows the transformed feature distance space for the example problem. If n people provide m biometric samples each, the numbers of intra-person and inter-person distance samples, respectively, are [17]:

$$n_{\oplus} = \frac{m \times (m-1) \times n}{2} \quad n_{\varnothing} = m \times m \times \frac{n \times (n-1)}{2} \quad (2)$$

In the authentication process, a sample requiring authentication is first converted into a feature vector. The difference between this feature vector and an earlier-obtained enrollment feature vector from this user is computed, and the resulting difference vector is classified as within-class (intra-person) for authentication or between-class (inter-person) for non-authentication. The system uses a *k*-nearest neighbor algorithm, which assigns a testing sample to the class most common among its *k* nearest neighbors among the feature-difference vectors in the training set [7]. System performance is calculated by letting the intra-person and inter-person distances simulate true users and impostors respectively [11]. The system measures the rate at which "true users" are falsely rejected and "impostors" falsely accepted, as well as the rate of correctly assigned "users".

3.3 ROC Curve Generator

The authentication results are used by the Receiver Operating Characteristic (ROC) curve generation process to provide further analysis. This process generates a graphical curve that shows the trade-off between the FAR and FRR at different threshold operating points. The method for obtaining the ROC curve using a weighted procedure of the *k* nearest neighbors has been described in Stewart et al [11] and Zack et al [15]. The basic unweighted procedure is that for each test sample, the *k* nearest-neighbor outputs are examined and the number of in-class matches counted. If the number of matches is equal to or greater than a threshold *m*, which varies from 0 to *k*, the sample is accepted as within-class; otherwise, it is rejected. For example, if *m* is 0, the sample is accepted if 0 or more of the *k* choices are within-class. For each value of *m*, the FAR and FRR are calculated. For *m*=0, all samples would be accepted, yielding an FRR of 0% and FAR of 100%. For the weighted method, the choices are weighted differently – the first choice is given a weight of *k*, the second *k*-1, and the k^{th} a weight of 1. The maximum score when all choices are within-class is k(k+1)/2. In this method, a sample is authenticated if the sum of all the weighted within-class choices is equal to a threshold *m* that varies between 0 and k(k+1)/2. For each value of *m*, the FAR and FRR are computed and plotted to obtain the ROC curve [11, 15].

4. Experiment Design

The 300 text samples used in this project were retrieved from Project Gutenberg, which offers more than 36,000 free e-books with expired copyrights for download in plain text format as well as formatted for popular e-readers [10]. Founded in 1971 by Michael S. Hart, it is the oldest and largest digital library [9]. Our text samples are taken from books published between 1880 and 1930. This particular period was chosen based on the availability of books with expired copyrights. The period was restricted to fifty years to ensure that linguistic differences between authors would be more related to personal style than the time of writing. It should be noted that the samples were not restricted geographically; for example, authors were included from Great Britain, Ireland, and the United States. The samples from each author may also span a great variety of text types. For example, Oscar Wilde's samples include essay (De Profundis), novel (The Picture of

Dorian Gray), and play (The Importance of Being Earnest). This is a much greater variety of text types than what the system has previously been tested on.

We used only authors who had more than ten texts available on Project Gutenberg. All texts had to be longer than 5,000 words and originally written in English. Our experiments employ ten samples each from thirty authors for a total of 300 texts. The thirty authors wrote in various genres – fiction (8), action/adventure fiction (3), science fiction (1), British literature (6), mystery and thriller (3), classical literature (7), and horror (2). An overview of the authors and their genres is shown in table 1.

TABLE 1. Overview of Authors

Author	Genre
Bennett, Arnold. Janvier, Thomas A. Lang, Andrew. Montgomery, L.M. Wodehouse, Pelham Grenville. Johnston, Annie Fellows. Tracey, Louis. Jacobs, W.W.	Fiction
Blackwood, Algernon, Lee, Vernon	Horror
Buchan, John, Haggard, H. Rider. London, Jack.	Action/Adventure Fiction
Burroughs, Edgar Rice	Science Fiction
Chesterton, G.K. Conrad, Joseph. Kipling, Rudyard. Shaw, George Bernard. Stevenson, Robert Louis. Wilde, Oscar.	British Literature
Doyle, Arthur Conan. Green, Anna Katharine. Rohmer, Sax	Mystery & Thriller
Harte, Bret. James, Henry. Stockton, Frank Richard. Twain, Mark. Wells, H.G. Wharton, Edith. Laut, Agnes C.	Classic Literature

The plain text files were prepared for our experiment by removing all the extraneous text such as publishing information, table of contents, book title, Project Gutenberg disclaimer, etc. The final result was that each text sample had the following format:

Line 1: Author first/middle name
Line 2: Author last name
Line 3: [Blank]
Line 4: Text begins

The stylometry feature extractor accepts xml files as input. To help facilitate the creation of these sample sizes (11 sizes for each text, 10 texts for each author), we created a text-splitting program that accepts a plain text file as input and creates 11 XML files of different sample sizes as output. A sample run of the program is shown in figure 2.

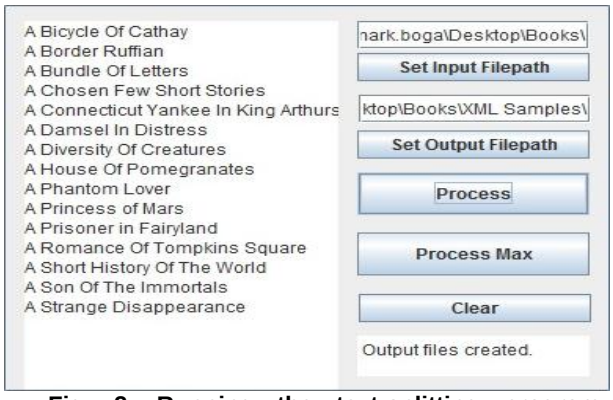


Fig. 2. Running the text-splitting program.

The user begins by typing in the names of the books that they want split. A ".txt" extension is automatically appended to the file name entered, so it is only necessary to enter the name of the .txt file, and that book will be processed as long as it has a .txt file extension. Although the example above uses only 15 .txt files for this simple test, one could use as many as desired. Next, the user has the ability to specify input and output file directories. By setting an input file path, a user can specify where the program will find the books listed. The complete folder file path must be entered, ending in a "\" symbol, and the button "Set Input Filepath" must be pressed. Similarly, the user can specify where the output XML files will be placed by setting the output file path using the same procedure. If those fields are unaltered and the file paths are not set, the program will look for the input files and output all files created in the same directory that the program's .jar file resides in.

For each file name entered, the program processes "fileName.txt" and will split it up into samples of varying sizes. When a book is processed, it tokenizes the strings of the text and splits the text up into words that are separated by blank spaces. (If two words are joined by symbols, such as "but--wait" without a space, this program will count them as one word.) A word count is tallied that keeps track of how many words the sample text contains. Upon reaching the text limit, which corresponds to the 11 sample sizes (250, 500, 750, 1000, 1500, 2000, 2500, 3000, 4000, 5000, and 10,000 words) that we want output files created for, a file of type .xml is created for that book. The first file created would be "fileName_250.xml" all the way to "fileName_10000.xml," assuming the initial input file contains at least that many words. If an input text file only has 100 words, no output file will be created. This ensures that every sample with "_250" appended to its name has exactly 250 words; every sample with "_500" appended to its name has exactly 500 words, etc. In order for a book to produce all 11 samples, it must contain at least 10,000 words. However, the program may also produce sample sizes of the maximum book

length by using the "Process Max" button. The .xml output files follow the pattern:

```
<?xml version="1.0" encoding="UTF-8"
standalone="no"?>
<biometrics-data>
<firstname>Author's first name</firstname>
<lastname>Author's last name</lastname>
<textinput>
The sample text will go here, size varies by filename.
</textinput>
</biometrics-data>
```

5. The Experiments

The 300 text samples were changed to XML files of eleven different sizes (250, 500, 750, 1000, 1500, 2000, 2500, 3000, 4000, 5000, and 10000 words) in order to obtain system performance as a function of text length. 8 of the 10K samples had less than 10000 words due to the size of the original text file. The XML files were input to the stylometry feature extractor, which produced a stylometry features vector file in the form of a text file. The text file includes the name of the file, the date and time of feature extraction, and the number of samples. The remaining entries consist of the name of the author and file, the number of feature attributes, and a sequence of feature measurements in the range 0-1.

The stylometry feature extractor has been previously described by Canales et al [2] and Stewart et al [11]. For the 30-author main experiment, the output file from the feature extractor was split into training and testing files with five books from each author in each file, using "strong" training where the system is trained on the test subjects. Figure 3 illustrates the tradeoff between FAR and FRR on the Receiver Operating Characteristic (ROC) curves for selected sample sizes. The ROC curves were derived using 9 nearest neighbors to provide weighted scores in the range 0–45 [15].

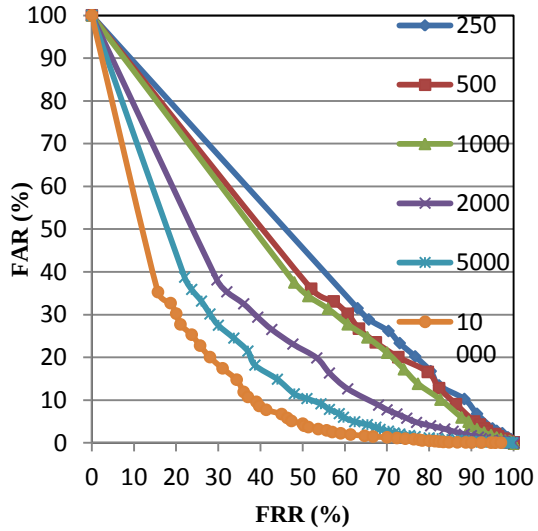


Fig. 3. ROC Curves: 250, 500, 1K, 2K, 5K, 10K words.

The ROC curves show an Equal Error Rate (EER), the point on the ROC curve where FAR = FRR [4], of approximately 25% for the 10K, 30% for the 5K, and 34% for the 2K-word samples. Performance gradually increases (lower EER) with increasing text length, although the performance increase was not entirely monotonic for samples sizes under 1K words.

Additional experiments – experiments A, B, and C – were performed on the 2K, 5K, and 10K-word samples to obtain performance on 15 authors (half the number used in the main experiment), using three different training methods. Figure 4 shows the ROC curves for experiment A, which used strong training on 15 of the authors. The system was trained on five books from each author and tested on the remaining five books from each author, similar to the main experiment. As expected, performance improved with fewer subjects, with EERs of approximately 20% for 10K, 24% for 5K, and 30% for 2K-word samples.

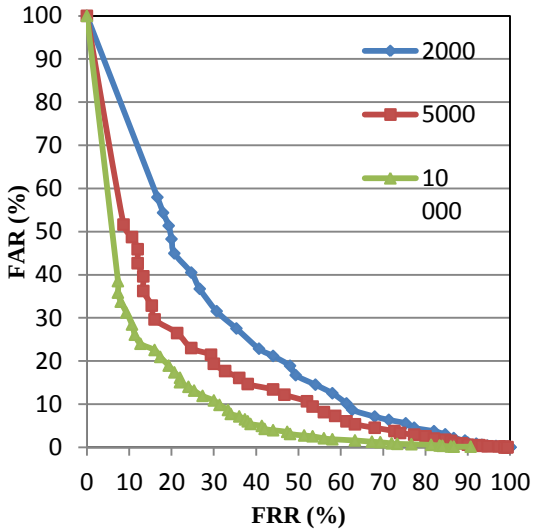


Fig. 4. Exp. A ROC Curves: 2K, 5K, 10K words.

Figure 5 shows the ROC curves for experiment B, which is similar to experiment A, except all ten samples from each of the remaining 15 authors were added to the training file. The additional training did not positively affect the results. Compared with experiment A, the EER was unchanged for 10K words, slightly worse for 5K words at approximately 25%, and clearly worse for the 2K-word samples at approximately 36%.

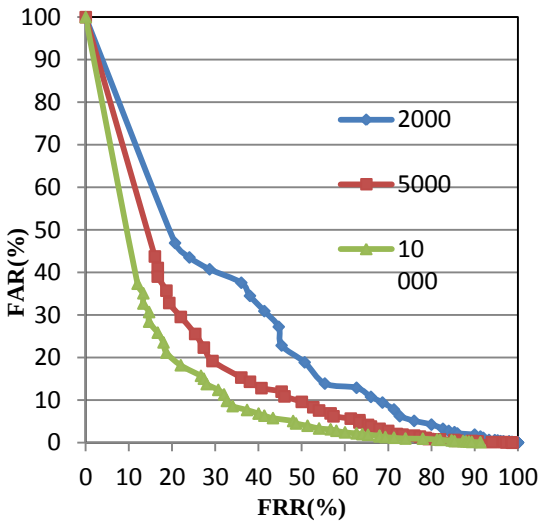


Fig. 5. Exp. B ROC Curves: 2K, 5K, 10K words

Experiment C trained the system on 15 authors (10 samples each) and tested on the remaining 15 authors (10 samples each), the weak training method. The performance results are shown in Figure 6. As expected, weak training yielded the weakest performance results of the three 15-author experiments, with EERs of approximately 30%, 34%, and 37%, respectively.

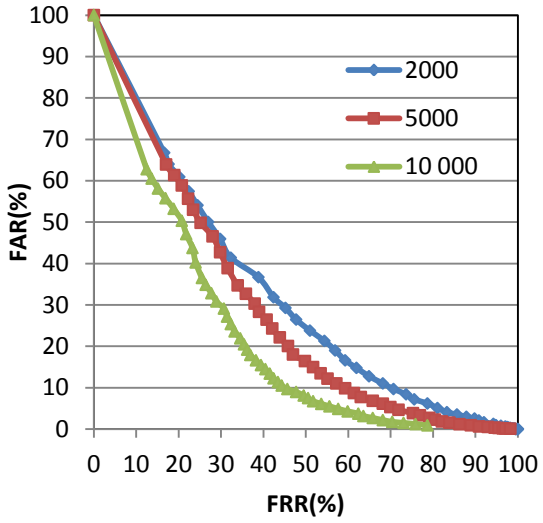


Fig. 6. Exp. C ROC Curves: 2K, 5K, 10K words.

The 1-nearest-neighbor authentication results for all eleven 30-author experiments and the three additional 15-author experiments are shown in Table 2.

Table 2. Authentication Results, k = 1.

Experiment	FRR	FAR	Performance	EER
Main Experiment				
250 words	94.00% (282/300)	4.11% (447/10875)	93.48% (10446/11175)	
500 words	92.67% (278/300)	5.33% (580/10875)	92.32% (10317/11175)	
750 words	90.33% (271/300)	4.63% (503/10875)	93.07% (10401/11175)	
1000 words	89.67% (269/300)	5.64% (613/10875)	92.11% (10293/11175)	
1500 words	83.33% (250/300)	6.09% (662/10875)	91.84% (10263/11175)	~37%
2000 words	79.33% (238/300)	6.52% (709/10875)	91.53% (10228/11175)	~34%
2500 words	75.33% (226/300)	7.34% (798/10875)	90.84% (10151/11175)	~34%
3K words	71.67% (215/300)	8.78% (955/10875)	89.53% (10005/11175)	~30%
4K words	75.33% (226/300)	7.18% (781/10875)	90.99% (10168/11175)	~32%
5K words	70.00% (210/300)	7.33% (797/10875)	90.99% (10168/11175)	~30%
10K words	55.00% (165/300)	7.13% (775/10875)	91.59% (10235/11175)	~25%
Additional Experiments				
Exp. A 2K words	68.67% (103/150)	12.53% (329/2625)	84.43% (2343/2775)	~30%
Exp. A 5K words	54.67% (82/150)	13.03% (342/2625)	84.72% (2351/2775)	~24%
Exp. A 10K words	36.67% (55/150)	9.30% (244/2625)	89.23% (2476/2775)	~20%
Exp. B 2K words	79.33% (119/150)	9.07% (238/2625)	87.14% (2418/2775)	~36%
Exp. B 5K words	58.67% (88/150)	9.68% (254/2625)	87.68% (2433/2775)	~25%
Exp. B 10K words	52.00% (78/150)	9.03% (237/2625)	88.65% (2460/2775)	~20%
Exp. C 2K words	70.07% (473/675)	17.40% (1827/10500)	79.42% (8875/11175)	~37%
Exp. C 5K words	53.48% (361/675)	19.66% (2064/10500)	78.30% (8750/11175)	~34%
Exp. C 10K words	38.52% (260/675)	21.93% (2303/10500)	77.06% (8612/11175)	~30%

The False Rejection Rate (FRR) is the rate the system fails to recognize an authentic user, while the False Acceptance Rate (FAR) is the rate the system falsely identifies an unauthorized user. In previous short-text experiments on this stylometry system [2], the FRR was particularly high with a decreasing trend with increased text length. For the main experiments here, FAR initially increases with text length, but the 10K-word experiment shows a significantly decreased FRR in comparison with the 5K-word experiment, while FAR remains roughly constant. The overall performance of the additional experiments is lower than that of the main experiments. The false acceptance rate is higher for the additional experiments, especially for experiment C (strong training). Validation experiments gave similar results for the main experiments, but the improvement for experiments A (~33%, ~30%, ~24% respectively) and B (~36%, ~31%, ~25%) was less significant, and

the results were slightly better for experiment C (~39%, ~29%, ~28%).

6. Conclusion

The main conclusion is that performance as measured by the Equal Error Rate generally improves with increased text length. All sample sizes have high false rejection rates in the majority-decision authentication tests (Table 2), but show an EER on the ROC curves at points corresponding to less than a majority decision, one to three nearest neighbors out of the nine. The additional experiments show that performance improves with fewer test subjects, and that the best performance results are obtained using strong training.

With eight of the 300 books having slightly less than 10,000 words, these experiments have reached the upper limit on the text length for the books selected. Therefore, further increases in performance would require better features and/or better pattern classification algorithms. This seems to indicate that the linguistic style of authors is not very unique and thus an extremely weak biometric. It is well known that physiological biometrics such as iris and fingerprint have high performance rates and EER's down to small fractions of a percent. Even most of the other behavioral biometrics have much higher performance rates - for example, a recent study on the keystroke biometric attained an EER of 0.5% on a population of 30 users with text lengths ranging from 433 to 1831 words of input [11]. This raises the question of how accurate previous studies in stylometry can have been.

7. Further Work

In order to test the relationship between the stylometric features and the accuracy of the system, we propose a series of experiments on the same samples using one feature group at a time. This should give a good indication of which features contribute more to the accuracy. According to the results in Grieve[5, 6], we might expect function word frequency, punctuation, and two-grams to prove especially useful, whereas the frequency of other categories of words may prove less useful as a characteristic of an author's style. We have provided a suggestion in Appendix B for additional features for future work on the PSBS.

References

[1] Ahmed Abbasi, Stephen France, Zhu Zhang, Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks,"*IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 447-462, Mar. 2011, doi:10.1109/TKDE.2010.110

[2] Omar Canales, Vinnie Monaco, Thomas Murphy, Edyta Zych, John Stewart, Charles Tappert ,Alex Castro, Ola Sotoye, Linda Torres, Greg Truly. "A Stylometry System for Authenticating Students Taking Online Tests". Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 6th, 2011.

[3] Malcolm Walter Corney. *Analysing E-mail Text Authorship for Forensic Purposes*. Queensland University of Technology, Brisbane, Australia, 2003. Master of Information Technology thesis.

[4] "Equal Error Rate (EER)." *Biometrics Glossary*. NSTC Subcommittee on Biometrics. http://www.expertglossary.com/biometrics/definition/equal-error-rate-eer. Accessed October 2011.

[5] Jack Grieve. Quantitative Authorship Attribution: A History and an Evaluation of Techniques. Simon Fraser University, Burnaby, BC, Canada, 2005. MA thesis.

[6] Jack Grieve. "Quantitative authorship attribution: an evaluation of techniques." Literary and Linguistic Computing 22: 251-270. 2007

[7] k-nearest Neighbor Algorithm. http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm Accessed October 2011

[8] Elizabeth Lex, Alexander Juffinger, Michael Granitzer. "A Comparison of Stylometric and Lexical Features for Web Genre Classification and Emotion Classification in Blogs". dexa, pp.10-14, 2010 Workshops on Database and Expert Systems Applications, 2010

[9] Project Gutenberg. http://www.gutenberg.org/. Accessed October 2011.

[10]Project Gutenberg http://en.wikipedia.org/wiki/Project_Gutenberg. Accessed October 2011.

[11] John C. Stewart, John V. Monaco, Sung-Hyuk Cha, and Charles C. Tappert, "An Investigation of Keystroke and Stylometry Traits for Authenticating Online Test Takers," Proc. IEEE Int. Joint Conf. Biometrics, Washington D.C., October 2011.

[12] Benno Stein, Nedim Lipka, Peter Prettenhofer. "Intrinsic Plagiarism Analysis". *Language Resources and Evaluation*, January 2010. http://www.uni-weimar.de/medien/webis/publications/papers/stein_2011a.pdf

[13] "Stylometry." Oxford Dictionaries. April 2010. Oxford University Press. http://oxforddictionaries.com/definition/stylometry?region=us. Accessed October 2011.

[14] C.C. Tappert, S. Cha, M. Villani, and R.S. Zack, "Keystroke Biometric Identification and Authentication on Long-Text Input," Int. Journal Information Security and Privacy (IJISP), 2010.

[15] R.S. Zack, C.C. Tappert, and S. Cha, "Performance of a Long-Text-Input Keystroke Biometric Authentication System Using an Improved k-Nearest-Neighbor Classification Method," Proc. IEEE 4th Int. Conf. Biometrics, Washington D.C., September 2010.

[16] Rong Zheng, Jiexun Li, Hsinchun Chen, Zan Huang. "A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques". *Journal of the American Society for Information Science and Technology*, American Society for Information Science and Technology, February 2006.

[17] S. Yoon, S-S Choi, S-H Cha, Y. Lee, and C.C. Tappert. On the individuality of the iris biometric. Int. J. Graphics, Vision & Image Processing, 5(5): 63-70, 2005

Appendix A. Stylometry Features

Character-based features:
1. number of alphabetic characters/number of characters
2. number of uppercase alphabetic characters/ number of alphabetic char
3. number of digit characters/number of characters
4. number of space characters/number of characters
5. number of vowel (a,e,i,o,u) characters/number of alphabetic characters
6. number of "a" (upper or lowercase) characters/number of vowel characters
7. number of "e" characters/number of vowel characters
8. number of "i" characters/number of vowel characters
9. number of "o" characters/number of vowel characters
10. number of "u" characters/number of vowel characters
11. number of most frequent consonants (t,n,s,r,h)/number of alph char
12. number of "t" characters/number of (t,n,s,r,h)
13. number of "n" characters/number of (t,n,s,r,h)
14. number of "s" characters/number of (t,n,s,r,h)
15. number of "r" characters/number of (t,n,s,r,h)
16. number of "h" characters/number of (t,n,s,r,h)
17. number 2nd most frequent consonants (l,d,c,p,f)/number of alph char
18. number of "l" characters/number of (l,d,c,p,f)
19. number of "d" characters/number of (l,d,c,p,f)
20. number of "c" characters/number of (l,d,c,p,f)
21. number of "p" characters/number of (l,d,c,p,f)
22. number of "f" characters/number of (l,d,c,p,f)
23. number 3rd most frequent consonants (m,w,y,b,g)/number of alph char
24. number of "m" characters/number of (m,w,y,b,g)
25. number of "w" characters/number of (m,w,y,b,g)
26. number of "y" characters/number of (m,w,y,b,g)
27. number of "b" characters/number of (m,w,y,b,g)
28. number of "g" characters/number of (m,w,y,b,g)
29. number of least frequent consonants (j,k,q,v,x,z) / number of alph char
30. number of consonant-consonant digrams/number alph digrams
31. number of "th" digrams/consonant-consonant digrams
32. number of "st" digrams/number consonant-consonant digrams
33.number of "nd" digrams/number consonant-consonant digrams
34. number of vowel-consonant digrams/number alph digrams
35. number of "an" digrams/number of vowel-consonant digrams
36. number of "in" digrams/number of vowel-consonant digrams
37. number of "er" digrams/number of vowel-consonant digrams
38. number of "es" digrams/number of vowel-consonant digrams
39. number of "on" digrams/number of vowel-consonant digrams
40. number of "at" digrams/number of vowel-consonant digrams
41. number of "en" digrams/number of vowel-consonant digrams
42. number of "or" digrams/number of vowel-consonant digrams
43. number of consonant-vowel digrams/number of alphabet digrams
44. number of "he" digrams/number of consonant-vowel digrams
45. number of "re" digrams/number of consonant-vowel digrams
46. number of "ti" digrams/number of consonant-vowel digrams
47. number of vowel-vowel digrams/number of alphabet letter digrams
48. number of "ea" digrams/number of vowel-vowel digrams
49. number of double-letter digrams/number of alphabet letter digrams
Word-based features:
1. number of one-letter words/number of words
2. number of two-letter words/number of words
3. number of three-letter words/number of words
4. number of four-letter words/number of words
5. number of five-letter words/number of words
6. number of six-letter words/number of words
7. number of seven-letter words/number of words
8. number of long words (eight or more letters)/number of words
9. number of short words (one to three letters)/number of words
10. average word length = number letters in all words/number of words
11. number of different words (vocabulary)/number of words
12. number of words occurring once/number of words
13. number of words occurring twice/number of words
Syntax-based features:
1. number of the eight punctuation symbols (.,?!;:'")/number of char
2. number of periods (.)/number of the eight punctuation symbols
3. number of commas (,)/number of the eight punctuation symbols
4. number of "?" and "!"/number of the eight punctuation symbols
5. number of semicolons (;) and colons (:)/number punctuation symbols
6. number of single (') and double quotes (")/punctuation symbols
7. number of non-alphabetic, non-punctuation, and non-space characters (0,1,2,3,4,5,6,7,8,9,@,#,$,%,etc.)/number of characters
8. number of digit char/number of non-alph, non-punct, and non-space char
9. number of common conjunctions/number of words
10. number of common interrogatives/number of words
11. number of common prepositions/number of words
12. number of first-person personal pronouns/number of personal pronouns
13. number of 2nd-person personal pronouns/number personal pronouns
14. number of 3rd-person personal pronouns/number of personal pronouns
15. number of personal pronouns (from above)/number of words
16. number of common nouns number of words
17. number of common verbs/number of words
18. number of common auxiliary verbs/number of words
19. number of "can" words/number of common auxiliary verbs
20. number of "did", "do", "does" words/number of common auxiliary verbs
21. number of "had", "has", "have" words/number of common auxiliary verbs
22. number of could, should, would/number of common auxiliary verbs
23. number of "will" words/number of common auxiliary verbs
24. number of common auxiliary verbs/number of common verbs
25. number to-be verbs (am,are,be,been,being,is,was,were)/number words
26. number of to-be verbs/number of common verbs
27. number of "am" words/number of to-be verbs
28. number of "are" words/number of to-be verbs
29. number of "be", "been", and "being" words/number of to-be verbs
30. number of "is" words/number of to-be verbs
31. number of "was" words/number of to-be verbs
32. number of "were" words/number of to-be verbs
33. number of common adjectives/number of words
34. number of articles (a, an, the)/number of words
35. number of articles (a, an, the)/number of common adjectives
36. number of "the" articles/number of articles
37. number of "a" or "an" articles/number of articles
38. number of indefinite personal pronouns/number of words
39. number of determiners/number of words
40-64. number of each of 25 most common words/number of words
65-164. number of each of 100 most common words /number of most common words of that major category (e.g., number "the"/num adjectives)
165. average number of characters per sentence
166. average number of words per sentence

Appendix B. Suggested Additional Features

Character-based features:
1. number non alphabetic+space+digit characters/number characters
2. number alphabetic+space digrams/number digrams
3. number non alphabetic+space digrams/number digrams
4. number space-consonant digrams/number alph+space digrams
5. number space-vowel digrams/number alph+space digrams
6. number consonant-space digrams/number alph+space digrams
7. number vowel-space digrams/number alph+space digrams
Word-based features:
10. number "not" contracted words/number common contracted words
11. number "am/is/are" contracted words/num common contracted words
12. number other contracted words/number common contracted words
13. number common contracted words/number words
14. number certain abbreviated words/number common abbreviated words
15. number common abbreviated words/number words
16. number certain slang words/number common slang words
17. number common slang words/number words
18. number common French words/number common foreign words
19. number common Ital/Span/Latin words/number common foreign words
20. number common other language words/number common foreign words
21. number common foreign words/number words
22. number certain hyphenated words/number hyphenated words
23. number common hyphenated words/number words
24. number 1-10 number words/number common number words
25. number 11-90 number words/number common number words
26. number 100+ number words/number common number words
27. number common number words/number words
28. number certain word pairs/number common word pairs
29. number common word pairs/number word pairs
Syntax-based features:
1. number sentences with 1-10 words/number sentences
2. number sentences with 11-20 words/number sentences
3. number sentences with 21-30 words/number sentences
4. number sentences with 31 or more words/number sentences