IEEE 6th International Conference on Biometrics, BTAS 2013.

Behavioral Biometric Verification of Student Identity in Online Course Assessment and Authentication of Authors in Literary Works

John V. Monaco, John C. Stewart, Sung-Hyuk Cha, and Charles C. Tappert

Seidenberg School of CSIS, Pace University, White Plains, NY 10606

Abstract

Keystroke and stylometry behavioral biometrics were investigated with the objective of developing a robust system to authenticate students taking online examinations. This work responds to the 2008 U.S. Higher Education Opportunity Act that requires institutions of higher learning undertake greater access control efforts, by adopting identification technologies as they become available, to assure that students of record are those actually accessing the systems and taking the exams in online courses. Performance statistics on keystroke, stylometry, and combined keystroke-stylometry systems were obtained on data from 30 students taking examinations in a university course. The performance of the keystroke system was 99.96% and 100.00%, while that of the stylometry system was considerably weaker at 74% and 78%, on test input of 500 and 1000 words, respectively. To further investigate the stylometry system, a separate study on 30 book authors achieved performance of 88.2% and 91.5% on samples of 5000 and 10000 words, respectively, and the varied performance over the population of authors was analyzed.

1. Introduction

The main application of interest in this study is verifying the identity of students in online examination environments, an application that is becoming more important with the student enrollment of online classes increasing, and instructors and administrations becoming concerned about evaluation security and academic integrity. The 2008 federal Higher Education Opportunity Act (HEOA) requires institutions of higher learning to make greater access control efforts for the purposes of assuring that students of record are those actually accessing the systems and taking online exams by adopting identification technologies as they become more ubiquitous [9]. To meet the needs of this act, the keystroke biometric seems appropriate for the student authentication process. Stylometry appears to be a useful addition to the process because the correct student may be keying in the test answers while a coach provides the answers with the student merely typing the coach's words without bothering to convert the linguistic style into his own.

Keystroke biometric systems measure typing characteristics believed to be unique to an individual and difficult to duplicate [4, 10]. The keystroke biometric is a behavioral biometric, and most of the systems developed previously have been experimental in nature. Nevertheless, there has been a long history of commercially unsuccessful implementations aimed at continuous recognition of a typist. While most previous work dealt with short input (passwords or short name strings) [1, 7, 14, 15, 16], some used long free (arbitrary) text input [2, 8, 11, 13, 19, 20]. Free-text input as the user continues typing allows for continuous authentication [5, 12, 13, 17] which can be important in online exam applications [6, 19].

Stylometry is the study of determining authorship from the authors' linguistic styles. Traditionally, it has been used to attribute authorship to anonymous or disputed literary documents. More recently, computer-based communication and digital documents have been the focus of research, sometimes with the goal of identifying perpetrators or other malicious behavior. Recent computer studies have used stylometry to determine authorship of emails, tweets, and instant messaging, in an effort to authenticate users of the more commonly used digital media. A few studies have applied stylometry to the detection of intentional obfuscation or deceptive writing style, and others to the detection of the author's demographics [3]. Appendix A summarizes the prior authorship attribution stylometry studies and lists the associated references.

There are several reasons keystroke and stylometry biometric applications are appealing. First, they are not intrusive to computer users. Second, they are inexpensive since the only hardware required is a computer with keyboard. Third, text continues to be entered for potential repeated checking after an initial authentication phase, and this continuing verification throughout a computer session is referred to as dynamic verification [11].

A number of measurements or features are generally used to characterize an individual. For the keystroke biometric these measurements are typically key press duration (dwell) times, transition (latency) times, and the identity of the keys pressed. Stylometry typically uses statistical linguistic features at the word and syntax level.

The current work addresses some of the limitations of prior work on free-text biometric systems [20]. The current system has several unique aspects. First, it can collect raw keystroke data over the Internet as well as from a key logger on an individual machine. Second, it focuses on free-text input where sufficient keystroke data are available to permit the use of powerful statistical feature measurements – and the number, variety, and strength of the measurements used in the system are much greater than those used by earlier systems reported in the literature. Third, it focuses on applications using arbitrary text input because copy texts are unacceptable for most applications of interest. And, fourth, because of the statistical nature of the features and the use of arbitrary text input, special statistical procedures are incorporated into the system to handle the paucity of data from infrequently used keyboard keys.

Using an open biometric system approach, an earlier student authentication study was conducted on data obtained from students taking actual tests in a university course [19]. In contrast, this paper presents a closed biometric system approach to classification that significantly increases the performance reported in the earlier study. Also, to further analyze the stylometry component of the system, a separate study on 30 book authors was undertaken to evaluate the stylometry performance on text lengths ranging from 250 to 10000 words. Additionally, because the mean population performance does not give the complete picture, the varied performance over the population of users was analyzed on the book-author study.

The paper organization is as follows: section 2 describes the system procedures, section 3 the student online testing studies, section 4 the stylometry study on short novels, and section 5 the conclusion and suggestions for future work.

2. Keystroke and Stylometry Systems

The keystroke and the stylometry systems consist of a data collector, a feature extractor, and a pattern classifier. The frontends of both systems, up through the feature extractor, were used from earlier studies, the keystroke frontend from [20] and the stylometry frontend from [19], and these frontend systems are described only briefly below. A third combined keystroke-stylometry system simply concatenates the feature vectors from the first two systems. A generic classification system operates on feature-vector input from the keystroke, stylometry, or the combined system. This classification system was improved significantly over those in the earlier mentioned studies and is one of the important contributions of this study.

The input system captures the keystroke timings and full input text in an XML file. The feature extractor parses each file creating both keystroke and stylometry feature vectors for later processing.

2.1. Keystroke System

Outlier removal preprocessing, performed iteratively until no change, eliminates key-press duration (dwell) and key transition (latency) times greater than two standard deviations from the mean over the whole dataset. This is particularly important for eliminating long transitions due to typing pauses from phone calls and other interruptions.

The 239 employed features include means and standard deviations of the timings of key press durations and transitions, and percent use of certain keys, grouped as follows [20]:

- 78 duration features (39 means and 39 standard deviations) of individual letter and non-letter keys, and of groups of letter and non-letter keys (Figure 1)
- 70 type-1 transition features (35 means and 35 standard deviations) of the transitions between letters or groups of letters, between letters and non-letters or groups thereof, between non-letters and letters or groups thereof, and between non-letters and non-letters or groups thereof (Figure 2)
- 70 type-2 transition features (35 means and 35 standard deviations) identical to the type-1 transition features except for the method of measurement (Figure 2)
- 19 percentage features that measure the percentage of use of the non-letter keys and mouse clicks
- 2 keystroke input rates: the unadjusted input rate (total time to enter the text / total number of keystrokes and mouse events) and the adjusted input rate (total time to enter the text minus pauses greater than ½ second / total number of keystrokes and mouse events)

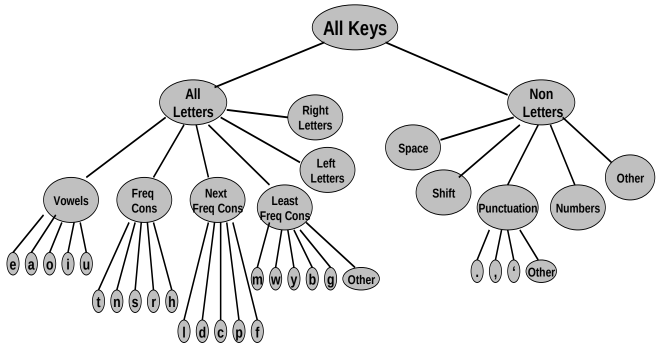

Figure 1. Hierarchy tree for the 39 duration categories (each oval).

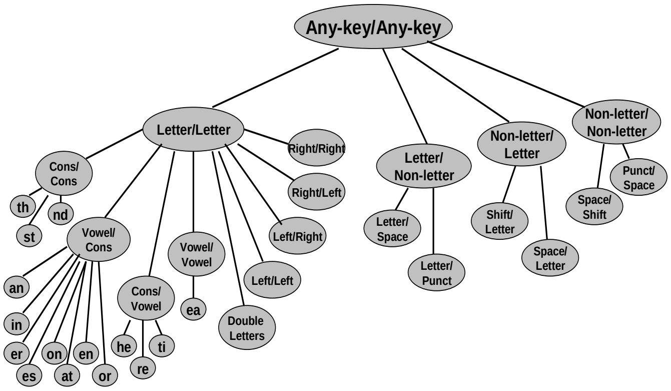

Figure 2. Hierarchy tree for the 35 transition categories (each oval) for type 1 and type 2 transitions.

Finally, to give each measurement roughly equal weight the features are standardized into the range 0-1 by converting raw measurement x to x' by the formula $x' = (x–x_{min})/(x_{max}–x_{min})$, where x_{min} and x_{max} are set to plus and minus two standard deviations from the mean, and *x'* is clamped between 0 and 1.

2.2. Stylometry System

The stylometry system uses a set of 228 linguistic features – 49 character-based, 13 word-based, and 166 syntax-based features (Appendix B). The features were normalized to be relatively independent of the text length – for example, *the number of different words (vocabulary) / total number of words* was used rather than simply *the number of different words.* The features were also chosen to show reasonable variation over a population of users – for example, some students use a large vocabulary and others a small one. As in the keystroke system, the features are standardized into the range 0-1.

2.3. Common Classification System

The classification procedure is based on a vector-difference authentication model which transforms a multi-class problem into a two-class problem [20]. The resulting two classes are *within-person* ("you are authenticated") and *between-person* ("you are not authenticated"). This dichotomy model is a strong inferential statistics method found to be particularly effective in large open biometric systems where it is not possible to train the system on all individuals in the population.

The application of interest here, however, involves a closed population of students where it is possible to train the system on all of the authorized users. Therefore, a more accurate "engineering" closed-system procedure was developed for these and similar applications. Two performance enhancing modifications were made in converting the open to the closed-system procedure. First, the new procedure matches the claimed user's sample against all the enrollment samples from that user for authentication rather than just one as in the open system. Second, the new procedure is user-focused in that only the claimed user's enrollment samples and their relationships to the other users' enrollment samples are utilized in the classification process.

In the simulated authentication process, a claimed user's keystroke sample requiring authentication is first converted into a feature vector. The differences between this feature vector and all the earlier-obtained enrollment feature vectors from this user are computed. The resulting query difference vectors are then classified as within-person (authentication) or between-person (non-authentication) by comparing them to the previously computed difference vectors for the claimed user. A k-nearest-neighbor algorithm with Euclidean distance is used to classify the unknown difference vectors, with a reference set composed of the differences between all combinations of the claimed user's enrolled vectors (within-person) and the differences between the claimed user and every other user (between-person). Thus, *differences of difference vectors are being calculated*.

A leave-one-out cross fold validation (LOOCV) is used in order to obtain system performance. The LOOCV procedure simulates many true users trying to get authenticated and many imposters trying to get authenticated as other users. For n users each supplying m samples, $m \times n$ positive (one for each sample) and $m \times n \times (n-1)$ negative (each sample versus the other users) tests can be performed.

Receiver operating characteristic (ROC) curves characterize the performance of a biometric system and show the trade-off between the False Accept Rate (FAR) and the False Reject Rate (FRR). In this study, the ROC curves were obtained using a linear-weighted decision procedure of the k nearest neighbors with k=21. Each neighbor is assigned a weight, from k to 1, with the closest neighbor weighted by k, the second by k-*1*, …, and the farthest by 1. With k fixed, another parameter, l, is varied from 0 to $k(k+1)/2$, resulting in 232 points on the ROC curve. At each point, the query sample is accepted as *within* if the weighted sum is greater than or equal to l and *between* otherwise. The error rates are then calculated as $FAR = FP/(FP + CN)$ and $FRR = FN/(FN + CP)$, where FP = # false positives, FN = # false negatives, CP = # correct positives, and CN = # correct negatives.

3. Student Online Testing Studies

3.1. Data Collection

The data employed in this study were obtained from an earlier study [19]. The data were collected from 40 students of a spreadsheet modeling course in the business school of a four-year liberal arts college. The classes met in a desktop computer laboratory where the exams were administered. Although this study investigated an online examination application, the data were captured in a classroom setting for greater experimental control. The 40 students took four online short-answer tests of 10 questions each, the tests spaced at approximately two week intervals. The students were unaware that their data were being captured for experimental analysis.

There were several problems with the keystroke data collection system. It was run from a weak server which ran slowly with 40 students accessing the system. The software was designed so students would click the NEXT button to go to the next question after completing the current one, not allowing a return to previous questions. However, due to the slowness of the system response to the click, some students would click on the NEXT button several times when there was not an immediate response and this would result in skipped questions. Also, some students could not remember the usernames and passwords they created on the first test and consequently could not log into the second; for the third and fourth tests this problem was resolved by the instructor providing the usernames and passwords when requested. As a result of these data collection problems, data were removed from students not completing all four tests or answering a sufficient number of questions per test, resulting in complete data sets from 30 students, 17 male and 13 female.

The text lengths of the answers to a test ranged from 433 to 1831 words per test, with a mean of 966 and a median of 915 words. An average word length of five characters (six with spaces between words) yields roughly 6000 keystrokes per test as input to the keystroke system.

All the tests were taken on classroom Dell desktop computers with associated Dell keyboards. Training and testing on the same type of keyboard is optimal because it is known that keystroke data tends to vary for different keyboards, different environmental conditions, and different types of texts [8, 20].

3.2. Experimental Design and Results

Two closed-system experiments were conducted on the data from the 30 students on each of the keystroke and stylometry systems using the leave-one-out procedure. Because the answers to the test questions could be short, several answers were combined to obtain reasonably sized biometric samples. In the first experiment, five test answers (half the test answers) were combined to obtain each sample, resulting in eight samples per student since each of the four tests contained ten questions for a total of 40 questions. In the second experiment, ten answers (all the answers of a test) were combined to obtain each sample, resulting in four samples per student. The experimental design and results are summarized in Table 1.

Table 1. Experimental design and results summary.

Experiment	Data Samples	Keystr EER	Stylo EER
1	8 samples/student 5 answers combined	0.04%	~26%
2	4 samples/student 10 answers combined	0.00%	~22%

Figure 3 presents the ROC curves for the keystroke and stylometry systems for the two experiments. For both the keystroke and stylometry systems, performance improved in going from experiment 1 to experiment 2 with the doubling of the data sample size.

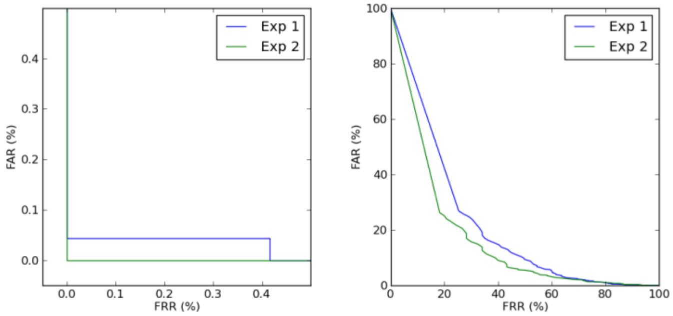

Figure 3. Online test ROC curves, 30 students: keystroke (left) and stylometry (right).

4. Stylometry Study on Short Novels

The stylometry results on the student tests were considered weak and the combined keystroke-stylometry system did not result in increased performance over that of the keystroke system alone. Therefore, considering that stylometry could require considerably more text input than keystroke analysis, a more extensive stylometry study was performed on short novels to determine system performance as a function of text length.

4.1. Data Collection

Text samples, 10 from each of 30 authors for a total of 300 samples, were retrieved from Project Gutenberg (http://en.wikipedia.org/wiki/Project_Gutenberg). The text samples were taken from books published between 1880 and 1930. This period was chosen based on the availability of books with expired copyrights and the period was restricted to fifty years to ensure that linguistic differences between authors would be more related to personal style than to the time of writing. The samples were not restricted geographically – authors were included from Great Britain, Ireland, and the United States. The samples from each author also span a variety of text types. For example, Oscar Wilde's samples include an essay, *De Profundis*, a novel, *The Picture of Dorian Gray*, and a play, *The Importance of Being Earnest*. All texts were longer than 5,000 words and originally written in English. 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), shown in Table 2.

Table 2. Overview of the 30 Authors.

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

The 300 text samples were cut into 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. Of the 10000 word samples, 8 had slightly less than 10000 words due to the size of the original text file.

4.2. Experimental Results

Figure 4 presents the ROC curves for the various sample lengths and Figure 5 shows the ERR as a function of the sample sizes in words. The EER was 8.5% for the 10K and

11.8% for the 5K word samples. As expected, performance gradually increased (lower EER) with increasing text length.

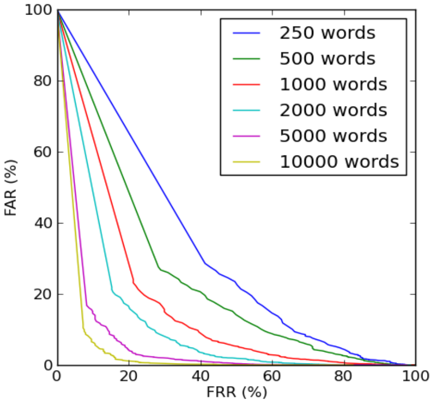

Figure 4. Book stylometry ROC Curves, 30 authors.

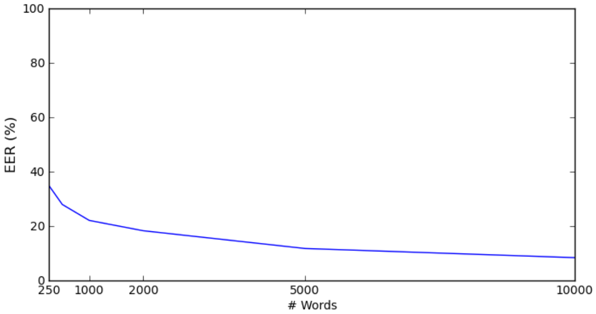

Figure 5. EER as a function of sample sizes in words.

Because the mean population performance does not give the complete picture, the varied performance at the EER over the population of authors was analyzed and described using the biometric animal designations. The FRR of each individual user was analyzed in order to find users which have trouble authenticated as themselves (goats). A distinction between two different types of FAR must be made though. When the true identity of the query sample is different from what is claimed during authentication, and a decision has been made to accept the query, then a false acceptance occurs. This false acceptance may contribute to either a weak template or a strong imposter. A distinction is made between the rate at which a template falsely accepts query samples and the rate at which an attacking query sample is falsely accepted. This distinction allows weak templates in the model to be identified (lambs), as well as attackers who may be skilled at imitating the identity of others (wolves). Over the author population, Figure 4 shows histograms of:

- FRR – identifying those easily verified, *sheep*, and those difficult to verify, *goats*.
- $FAR_{template}$ of how easily the true authors were imitated – identifying those easily attacked, *lambs.*
- $FAR_{attacker}$ of how easily imitators attacked true authors – identifying the strong attackers, *wolves.*

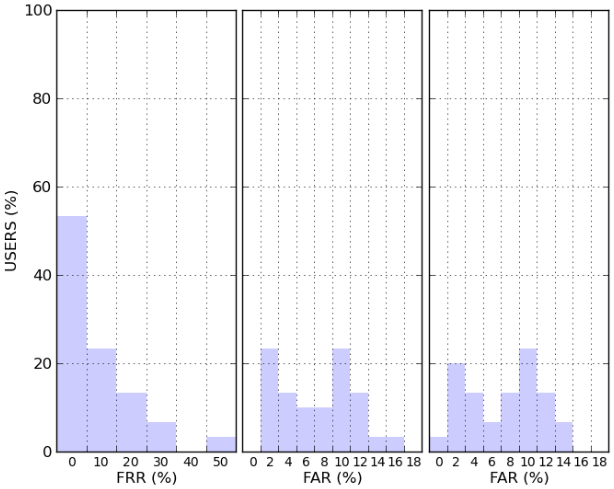

Figure 6. Histograms: FRR (left), FAR of receivers (middle), FAR of attackers (right), over the 30 author 10000 word samples.

Significant variation in performance over the population was demonstrated. For example, one author had 50% FRR (5 of 10 samples rejected) while all others had 30% or less (Figure 6 left). This author, Oscar Wilde, can be considered difficult to verify, a *goat*. Oscar Wilde's samples – which included an essay, a novel, and a play, as noted earlier – were not as homogeneous as those of the other authors.

5. Conclusion

The keystroke system performance results on the student test data were 100% on the 6000-keystroke full-test and 99.96% on the 3000-keystroke half-test samples. Although the results were obtained on a relatively small database, 30 students is a reasonable class size. These results were an improvement over the 99.45% performance on the 3000-keystroke half-test samples previously reported on the same data [19] Note that the leave-one-out procedure used in this study permitted the full test evaluations which were not possible using the procedure of the earlier study. High keystroke performance was anticipated in this study for such large volume keystroke input because high performance was also achieved in the earlier study on the same data [19] and a 98.3% performance was achieved on a 30-user, 750-keystroke-sample experiment in a recent study [13].

The performance of the keystroke biometric system is far superior to that of the stylometry one. While the keystroke and stylometry biometrics are both behavioral biometrics, they operate at different cognitive levels. The keystroke biometric operates at essentially an automatic motor control level. Stylometry, however, operates at a higher cognitive level, and because it primarily involves word and syntax-level units, much longer text passages are required relative to those required by the keystroke biometric.

To obtain system performance in this study we simulated the authentication process of many true users trying to get authenticated and of many zero-effort imposters trying to get authenticated as other users. Although authentication of online examination participants in real time would not be possible with the described technique due to the significant amount of input required (half or full test), delayed authentication with batch processing should be sufficient for university and HEOA requirements.

Important parameters in authorship attribution methods are the length and number of training and testing texts, and the number of potential authors [18]. Another important factor discovered in this stylometry study was the relationship between the texts under study and how the texts are produced. For example, in an earlier study it was discovered that a relatively strong correlation existed between the test answers and the test questions producing the answers [19]. Content-specific terminology inherent to the course subject matter, and used by a majority of the participants, confounded the results. Therefore, better performance results would likely be obtained from student essays on a variety of topics, as might be obtained from students in an English class, although two students who happen to choose the same or similar topic may present a problem.

Future work on improving stylometry in student examination applications might investigate the use of idiosyncratic features like the fraction of misspelled words, typing speed, and sequences of characters such that would be found in short words like “the” [6]. The use of longer text passages and those on different topics, such as essays in English classes, might also be explored, as well as different ways of fusing the keystroke and stylometry results. Finally, while the student examination experiments reported here used actual test data, the authentication process itself was simulated, so future work might explore an actual authentication process in a student assessment environment.

References

[1] S.S. Bender and H.J. Postley. Key sequence rhythm recognition system and method. U.S. Patent 7,206,938, 2007.

[2] F. Bergadano, D. Gunetti, and C. Picardi. User authentication through keystroke dynamics. *ACM* ***Trans. Info. & System Security,*** 5(4): 367-397, 2002.

[3] S. Bergsma, M. Post, and D. Yarowsky. Stylometric analysis of scientific articles. *Conf. North Am. Chap. Assoc. Comp.Ling: Human Language Tech. Assoc. Comp. Ling.*, PA, 327-337, 2012.

[4] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to biometrics*. NY: Springer, 2004.

[5] J. Ferreira and H. Santos. Keystroke dynamics for continuous access control enforcement. *Proc. Int.Conf. on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 216-223, 2012

[6] E. Flior and K.Kowalski. Continuous biometric user authentication in online examinations. *Proc. 7th Int. Conf. Info. Tech.: New Generations.* IEEE Computer Society, Wash. DC, 488-492, 2012.

[7] R.Giot, M. El-Abed, and C. Rosenberger. Keystroke dynamics with low constraints svm based passphrase enrollment. *Proc. IEEE Int. Conf. Biometrics: Theory, Applications, and Systems* (BTAS), 2009.

[8] D. Gunetti and C. Picardi. Keystroke analysis of free text. *ACM Trans. Info. and System Security*, 8(3):312-347, 2005.

[9] Higher Education Opportunity Act (HEOA) of 2008. http://www2.ed.gov/policy/highered/leg/hea08/index.html, accessed May 2012.

[10] L. Jin, X Ke, R. Manuel, and M. Wilkerson. Keystroke dynamics: a software based biometric solution. In *13th USENIX Security Symposium*, 2004.

[11] J. Leggett, G. Williams, M. Usnick, and M. Longnecker. Dynamic identity verification via keystroke characteristics. *Int. J. Man Machine Studies*, 35(6): 859-870, 1991.

[12] A. Messerman, T. Mustafic, S. Camtepe, and S. Albayrak. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. *Proc. Int. Joint Conf. Biometrics (IJCB 2011)*, Wash. D.C., 2011.

[13] J.V. Monaco, N. Bakelman, S.-H. Cha, and C.C. Tappert. Recent advances in the development of a long-text-input keystroke biometric authentication system for arbitrary test input. *Proc. Euro. Intelligence and Security Informatics Conf. (EISIC)*, Sweden, 2013.

[14] F. Montrose, M.K. Reiter, and S. Wetzel. Password hardening based on keystroke dynamics. *Int. J. Info. Security*, 1(2): 69-83, 2002.

[15] K. Revett. Chap 4: Keystroke dynamics, 73-136. in *Behavioral Biometrics: A Remote Access Approach*, Wiley, 2008.

[16] R. N. Rodrigues, G.F.G. Yared, C.R. Costa, J.B.T. Yabu-Uti, F. Violaro, and L.L. Ling. Biometric access control through numerical keyboards based on keystroke dynamics. *Lecture Notes Comp.Sci.*, 3832: 640-646, 2006.

[17] T. Shimshon, R. Moskovitch, L. Rokach, and Y. Elo,vici. Continuous verification using keystroke dynamics. *Proc. Int. Conf. Comp. Intel. and Security*. IEEE Computer Soc., Washington, DC, 411-415, 2010.

[18] E. Stamatatos. A survey of modern authorship attribution methods. *J. Am. Soc. Info. Science and Tech*., 60(3): 538–556, 2009.

[19] J. Stewart, J. Monaco, S. Cha, and C. Tappert. An Investigation of Keystroke and Stylometry Traits. Proc. *Int. Joint Conf. Biometrics (IJCB 2011)*, Washington D.C., Oct 2011.

[20] C. Tappert, S. Cha, M. Villani, and R.S. Zack. A keystroke biometric system for long-text input. *Int. J. Info. Security and Privacy (IJISP)*, 4(1): 32-60, 2010.

Appendix A. Summary of Prior Authorship Attribution Stylometry Studies.

Author - Year	#Subjects	Samples/Subj	Sample Size	Feature Types	#Features	Classification	Accuracy
Afroz, et. al. 2012	68	documents	500 words	Lexical, Syntactic, Content	707	SVM	96.6%
Alison & Guthrie 2008	9	174-706 Emails	~75 words	Bigrams, Trigrams Word frequency	Not given	Multimodal Hierarchical SVM	78.46% 87.05% 86.74%
Christani, et. al. 2012	77	60-100 IMs	615 words avg	Lex, Syntactic, Struct, Topic	Not given	Cumulative match	89.5%
Corney, et al. 2002	4	253 Emails	50-200 words	Stylistic, Struct, Func words	184	SVM	70.2%
de Vel 2000	5	18-87 Emails	3-680 words	Func words, Struct, Stylistic	38	SVM	85.7%
de Vel, et al. 2001	3	156 Docs	~12000 words	Stylistic	191	SVM	100%
Feiguina & Hirst 2007	11	4-10	2000 words	POS, Lexical, Syntactic	194	SVM	91.2% lex feat 88.7% all feat
Feng, et. al. 2012	10 5	8 sci papers 5 Novels	Variable 3000 sentences	Syntactic, Lexical,style	11 11	PCFG Trees & SVM	96% 95.2%
Gamon 2004	3	20 Sentences	Sentence	Func words, POS, Semantic	6018	SVM	85%
Goldman & Allison 2008	5	3 Novels	Variable length	POS, bigrams, Word freq	Not given	Chi Square	80%
Hirst & Feiguina 2007	2	250	1000 words	POS, Lexical, Vocab richness	194	SVM	99.2%
Hoover 2001	10	17 Novels	Variable length	Most frequent words	500	Cluster analysis	70%
Hoover 2003	8	16 Books	~24000 words	Vocabulary richness	10	Cluster analysis	37%
Iqbal, et al. 2008	6	20 Emails	Email	Lexical, Vocab richness	Not given	Data mining	86-90%
Iqbal, et al. 2010	158	200 Emails	Enron Corpus	Lex, Syntactic, Struct, Topic	292	SVM	82.9%
Kelselj, et al. 2003	8	Books	Variable length	Character 4-8-grams	Not given	Dissimilarity	100%
Koppel & Schler 2003	11	480 Emails	~200 words	Lexical, POS, Idiosyncrasies	358	C4.5 Trees, SVM	71.8%
Koppel & Schler 2004	10	21 Books	Variable length	Most frequent words	250	SVM	95.7%
Layton, et al. 2010	50	120 Tweets	=<140 char	Character n-grams	Not given	SCAP	70%
Li, et al. 2006	10	30-40	~169 words	Lexical, Structural, Syntactic	270	SVM	99.01%
Luyckx & Daelemans 2005	2	100 Articles	Variable length	n-grams, Syntactic, POS	91	ANN	71.3%
Luyckx & Daelemans 2008	145	Student Essays	~1400 words	Word, POS, n-grams, Lexical	91	k-NN, SVM	34%
Mustafa, et al. 2009	3	8 Books	Variable length	Frequent words, Word pairs	42	Correlation	0.99 correlation
Narayanan, et. al. 2012	100,000	24 blogs avg	Avg 305 words	Lex, Syntactic, Struct, Topic	1,188	kNN/RLSC	20%
Pavelec, et al. 2009	20	30 Articles	Variable length	Conjunctions (Portuguese) Adverbs (Portuguese)	77 94	Partial match SVM	84.3% 83.2%
Popescu & Dinu 2009	10	21 Books	Variable length	Function words	Not given	PCA Clustering	100%
Raghavan et al. 2010	6	14-28 Docs	7-24 k words	Syntactic	Not given	Naïve Bayes	95%
Stanczk & Cyan 2007	2	70 and 98	Short works	Function words, Punctuation	17	ANN	100%
Sun, et al. 2010 Sun, et al. 2010	20	30 Messages	~1383 char	Character n-grams	575 645	SVM GA	96.67% 93.67%
Tan & Tsai 2010	2	Novels	~60000 words	Syntactic	13	Bayesian	88.71%
Timboukakis & Tambouratzis 2009	5	1004	Variable length	Word freq, POS, Structural	85	MLP-NN SVM	89,71% 91.43%
Van Haltern 2004	8	9	628-1342words	Lexical and Syntactic	1050	Weighted voting	97%
Zheng, et al. 2003	9 3 3	153 70 70	News group Email Messages	Style Structural Content-specific	18	SVM	97% 91% 84%
Zheng, et al. 2006	20	30-92 Emails	84-346 words	Lexical, Syntactic, Structural	270	C4.5 Trees / SVM	93.36% / 97.69%

S. Afroz, M. Brennan, and R. Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. *Proc. 2012 IEEE Sym. Security and Privacy*. IEEE Computer Soc., Wash. DC, 461-475, 2012.

B. Allison and L. Guthrie. Authorship attribution of e-mail comparing classifiers cver a new corpus for evaluation, *Proc. LREC'08*, 2008.

M.Cristani, et al. Conversationally-inspired stylometric features for authorship attribution in instant messaging. Proc. 20th ACM Int. Conf. Multimedia. NY, 1121-1124, 2012.

M. Corney, O. de Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. *Proc. 18th Annual Computer Security App. Conf.*, Las Vegas, NV, Dec 2002.

O. de Vel. Mining e-mail authorship. *Proc. KDD-2000 Workshop on Text Mining*, Boston, Aug 2000.

O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, 30(4):55, Dec 2001.

O.Feiguenia and G.Hirst. Authorship attribution for small texts: literary and forensic experiments. *Proc. 30th Int.Conf. Special Int. Group. Info Retrieval* (SIGIR), 2007.

S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. *Proc. 50th Annual Meeting Assoc. Comp. Linguistics: Short Papers*, Assoc. Comp. Ling., Stroudsburg, PA, 2: 171-175, 2012.

M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. *Proc. 20th Int. Conf. Comp. Ling. (COLING '04). Assoc. Comp. Ling.*, Morristown, NJ, 611-617, 2004.

E. Goldman and A. Allison. Using grammatical markov modes for stylometric analysis. Stanford Univ. Tech. Report.

G. Hirst and O. Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4): 405-417, 2007.

D. Hoover. Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, 18(4): 341-60, 2003.

D. Hoover. Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing*, 16: 421-44, 2001.

F. Iqbal, R. Hadjidj, B. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digtal Investigation*, 5: 42-51, 2008.

F. Iqbal, L. Khan, C. Benjamin, and M. Debbabi. E-mail authorship verification for forensic investigation. *Proc. 2010 ACM Symposium Applied Computing* (SAC '10). ACM, New York, NY, 1591-1598, 2010.

V. Keselj, F. Peng, N. Cerone, and C. Thomas. N-gram-based author profiles for authorship attribution. *Proc. Conf. Pacific Assoc. Comp. Ling., PACLING'03*, Nova Scotia, 255-264, 2003.

M. Koppel and J. Schler. Authorship verification as a one-class classification problem. *ICML '04 Proc. .21st Int. Conf. Machine Learning*, New York, 2004.

M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. *Proc. IJCAI'03 Workshop Comp. Approaches to Style Analysis and Synthesis*, 69-72, 2003.

R. Layton, P. Watters, and R. Dazeley. Authorship attribution for twitter in 140 characters or less. *Second Cybercrime and Trustworthy Comp. Workshop*, 1-8, 2010

J. Li, R. Zheng, and H. Chen. From fingerprint to writeprint. *Comm. ACM*, 49(4): 76-82, 2006.

K. Luyckx and W. Daelemans. Authorship attribution and verification with many authors and limited data. *Proc. 22nd Int. Conf. Comp. Ling., COLING '08, Assoc. Comp. Ling.*, NJ, 1: 513-520, 2008.

K. Luyckx and W. Daelemans. Shallow text analysis and machine learning for authorship attribution. *Proc. 15th Meeting Comp. Ling. of the Netherlands*, 2005.

T. Mustafa, N. Mustapha, M. Azmi, and N. Sulaiman. Computational stylometic approach based on frequent word and frequent pair in text mining authorship attrib. *IJCSNS Int. J. Comp. Sci. Net. Sec.*, 9-3, 2009.

A. Narayanan, A. Paskov, H. Gong, N.Z. Bethencourt, J. Stefanov, E. Shin, E.C.R. Song, D. On the feasibility of internet-scale author identification. *IEEE Symp. Security and. Privacy*, 300-314, 2012.

D. Pavelec, L.S. Oliveira, E. Justino, F.D. Nobre Neto, and L.V. Batista. Compression and stylometry for author identification. *Int. Joint Conf. Neural Networks*, 2445-2450, 2009.

M. Popescu and L. Dinu. Comparing statistical similarity measures for stylistic multivariate analysis. *Proc. RANLP 2009*, Borovets, Bulgaria, 2009.

S. Raghavan, A. Kovashka, and R. Mooney. Authorship attribution using probabilistic context-free grammars. *Proc. ACL 2010 Conf. Short Papers, CLShort '10, Assoc. Comp. Ling.,* PA, 38-42, 2010.

U. Stanczyk and K. Cyran. Machine learning approach to authorship attribution of literary texts. *Int. J. Applied Mathematics and Informatics*, 1(4), 2007.

J. Sun, Z. Yang, P. Wang, and S. Liu. Variable length character n-gram approach for online writeprint identification. *Int.Conf. Multimedia Info. Netw. Sec. (MINES),* Nanjing, Jiangsu, 486-490, 2010.

J. Sun, Z. Yang, P. Wang, L. Liu, and S. Liu. Feature selection for online writeprint identification using hybrid genetic algorithm. *Int. Symp. Comp. Intel. and Design (ISCID),* Hangzhou, 76-79, 2010.

F. Tan and R. Tsai. Authorship identification for online text. *Proc. 2010 Int. Conf. Cyberworlds (CW '10). IEEE Computer Society*, Washington, DC, 155-162, 2010.

N. Tsimboukakis and G. Tambouratzis. A comparative study on authorship attribution classification tasks using both neural network and statistical methods. *Neural Comp. Appl.*, 19(4): 573-582, 2009.

H. Van Halteren. Linguistic profiling for author recognition and verification. *Proc. 42nd Annual Meeting Assoc. Comp.Ling*, Stroudsburg, PA, 199-206, 2004.

R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: writing-style features and classification techniques. *J. Am. Soc. Info. Science and Tech*., Feb 2006.

R. Zheng, Y. Qin, Z. Huang, and H. Chen. Authorship Analysis in Cybercrime Investigation. *Intel. Security Informatics*, Hsinchun Chen et al., Eds., Springer Berlin Heidelberg, 2665: 59-73, 2003.

Appendix B. 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 char
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