Active Linguistic Authentication Using Real-Time Stylometric Evaluation for Multi-Modal Decision Fusion

Ariel Stolerman, Alex Fridman, Rachel Greenstadt, Patrick Brennan, Patrick Juola

Chapter 12

ACTIVE LINGUISTIC AUTHENTICATION USING REAL-TIME STYLOMETRIC EVALUATION FOR MULTI-MODAL DECISION FUSION

Ariel Stolerman, Alex Fridman, Rachel Greenstadt, Patrick Brennan and Patrick Juola

Abstract Active authentication is the process of continuously verifying a user based on his/her ongoing interactions with a computer. Forensic stylometry is the study of linguistic style applied to author (user) identification. This paper evaluates the Active Linguistic Authentication Dataset, collected from users working individually in an office environment over a period of one week. It considers a battery of stylometric modalities as a representative collection of high-level behavioral biometrics. While a previous study conducted a partial evaluation of the dataset with data from fourteen users, this paper considers the complete dataset comprising data from 67 users. Another significant difference is in the type of evaluation: instead of using day-based or data-based (number-of-characters) windows for classification, the evaluation employs time-based, overlapping sliding windows. This tests the ability to produce authentication decisions every 10 to 60 seconds, which is highly applicable to real-world active security systems. Sensor evaluation is conducted via cross-validation, measuring the false acceptance and false rejection rates (FAR/FRR). The results demonstrate that, under realistic settings, stylometric sensors perform with considerable effectiveness down to 0/0.5 FAR/FRR for decisions produced every 60 seconds and available 95% of the time.

Keywords: Active authentication, stylometry, authorship verification

1. Introduction

The challenge of identity verification for the purpose of access control is the trade-off between maximizing the probability of intruder detec-

tion and minimizing the cost to the legitimate user in terms of time and distraction due to false alerts, along with the extra hardware requirements for physical biometric authentication. In recent years, behavioral biometric systems have been explored to address this challenge [3]. These systems rely on inexpensive input devices such as a keyboard and mouse. However, their performance in terms of detecting intruders and maintaining a low-distraction human-computer interaction experience has been mixed [8]. In particular, they have error rates ranging from 0% [29] to 30% [30], depending on the context, variability in task selection and other dataset characteristics.

The bulk of biometric-based authentication research has focused on verifying a user based on a static set of data. This type of one-time authentication is not well suited to a live multi-user environment, where a person may leave the computer for an arbitrary period of time without logging off. Such an environment requires continuous authentication when the computer is in a non-idle state. The Active Linguistic Authentication Dataset [18] used in this work was created to represent this general real-world scenario. The dataset, which was generated in a simulated office environment, contains behavioral biometrics associated with typical human-computer interactions by office workers.

Stylometry is a form of authorship recognition that relies on the linguistic information found in a document. While stylometry existed before computers and artificial intelligence (AI), the field is currently dominated by techniques such as neural networks and statistical pattern recognition. State-of-the-art stylometry approaches can identify individuals in sets of 50 authors with more than 90% accuracy [2] and even scaled to more than 100,000 authors [28]. Stylometry is currently used in intelligence analysis and forensics, with increasing applications in digital communication analysis [40]; the results are accurate and reliable enough to be admissible as legal evidence [11, 12]. The application of stylometry as a high-level modality for authenticating users in a continuous user verification system is novel. Initial evaluations of authorship attribution technologies are promising, realizing in excess of 90% identification accuracy over fourteen users [18].

This paper considers a set of stylometric classifiers, also referred to as sensors, as a representative selection of high-level behavioral biometrics. The primary goal is to evaluate authorship attribution approaches in realistic settings for active authentication, which require constant monitoring and frequent decision making about the legitimacy of a user at a computer in a dynamic, time-constrained environment. This work is designed as a preliminary evaluation of one modality among many to consider for an active authentication system. In the future, the sty-

lometric modalities discussed in this paper would be interleaved with other low- and high-level modalities, such as keyboard dynamics [33], mouse movements [3] and web browsing behavior [41], in a centralized decision fusion system. The use of such modalities, including stylometry, may provide a cost-effective alternative to sensors based on physiological biometrics [39].

Although this research focuses on active authentication, a live security application of stylometric analysis and its implications for the usability and configuration of stylometric sensors are relevant to forensic contexts. Consider, for example, a standard post-mortem forensic analysis of user input data aggregated over an entire day. This research seeks to identify the features to consider in such "noisy" settings, which include window sizes, effects of overlapping windows and how idle periods in data input should be considered.

2. Related Work

A defining problem of active authentication is that the verification of an identity must be carried out continuously on a sample of sensor data that varies drastically with time. Therefore, the classification has to be made using a "window" of recent data, dismissing or discounting the value of older data outside the window. Depending on the user task being performed, some biometric sensors may provide more data than others. For example, when a user browses the web, mouse and web activity sensors are flooded with data, while keystroke and stylometric sensors may only receive a few infrequent key presses.

This motivates recent work on multimodal authentication systems where the decisions of multiple classifiers are fused together [34]; the resulting verification process is more robust to the dynamic nature of real-time human-computer interactions. This paper examines only the effectiveness of stylometric sensors under active authentication settings, the goal being to eventually construct a multi-modal biometric system. The notion of decision fusion is motivated by the research of Ali and Pazzani [4], which achieved reduced error rates using distinctly different classifiers (i.e., using different behavioral biometrics) with several fusion options [10, 15, 20].

Authorship attribution based on linguistic style, or stylometry, is a well-researched field [6, 16, 17, 23, 32, 35]. Its principal application domain is written language – identifying an anonymous author of a text by mining it for linguistic features. The theory behind stylometry is that every person has a unique linguistic style or "stylome" [38] that can be quantified and measured in order to distinguish between different

authors. The feature space is potentially endless, with frequency measurements and numeric evaluations based on features across different levels of the text, including function words [9, 27], grammar [25], character n-grams [36] and more. Although stylometry has not been used for active user authentication, its application to this task brings higher-level inspection into the process compared with lower-level biometrics such as mouse movements and keyboard dynamics [7, 42].

The most common practice of stylometry is in supervised learning, where a classifier is trained on texts of candidate authors and used to attribute the stylistically closest candidate author to unknown writings. In an unsupervised setting, a set of writings whose authorship is unknown are classified into style-based clusters, each representing texts of a unique author.

In an active authentication setting, authorship verification is applied where unknown text is to be classified by a unary author-specific classifier. The text is attributed to an author only if it is stylistically close enough to the author. Although pure verification is the ultimate goal, standard authorship attribution as a closed-world problem is an easier (and sometimes sufficient) goal. In either case, classifiers are trained in advance and used for real-time classification of processed sliding windows of input keystrokes. If enough windows are recognized as an author other than the real user, the presence of an intruder is indicated.

Another use of stylometry is in author profiling [6, 19, 37]. This application is quite different from author recognition because writings are mined for linguistic features to identify characteristics of their authors such as age and gender [5], native language [24] and personality characteristics [13].

In a pure authorship attribution setting, where classification is done off-line on complete texts (rather than sequences of input keystrokes) and in a supervised setting where all candidate authors are known, state-of-the-art stylometry techniques perform very well. For instance, at PAN-2012 (`pan.webis.de`), some methods achieved more than 80% accuracy on a set of 241 documents, sometimes with added distractor authors.

Two key challenges arise in an active authentication setting. First, open-world stylometry is a much harder problem, with a tendency to yield high false negative (false reject) rates. The unmasking technique [22] has been shown to be effective – 95.7% accuracy – on a dataset of 21 books by ten nineteenth-century authors. However, the amount of data collected by sliding windows of sufficiently small durations requires efficient authentication and the lack of quality coherent literary writings render this approach infeasible for active linguistic authentication. Second, the inconsistent frequency nature of keyboard input along with the

relatively large amount of data required for good performance of stylometric techniques render a large portion of the input windows unusable for learning writing style.

On the other hand, an active authentication setting offers some advantages with regard to potential features and analysis method. Since the raw data consists of keystrokes, some linguistic and technical idiosyncratic features can be extracted; these include misspellings caught before they are auto-corrected and vanish from the dataset and patterns of deletions such as selecting a sentence and hitting delete as opposed to repeatedly hitting backspace to delete individual characters. Additionally, in an active authentication setting, it is intuitive to consider overlaps between consecutive windows, resulting in a large dataset and providing grounds for local voting based on a set of windows and controlling the frequency at which decisions are made.

3. Evaluation Dataset

The complete Active Linguistic Authentication Dataset [18] is used in the evaluation. The dataset, which contains data collected in a simulated work environment, is designed specifically for behavioral biometric evaluation. The data collection utilized an office space, which was allocated, organized and supervised by some of the authors of this paper. The office space contained five desks, each with a laptop, mouse and headphones. The equipment and supplies were chosen to be representative of a standard office environment. However, one of the important properties of the dataset is uniformity. Because the computers and input devices in the simulated office environment were identical, the variations in behavioral biometrics data can be attributed to variations in the characteristics of the users instead of the effects of variations in the physical environment.

The dataset contains data collected from 80 users. Due to equipment and software crashes and sick days taken by the subjects, a little more than 80 subjects were used for data collection to reach the 80 user goal. However, when examining the data obtained from the 80 users, it was observed that some users had produced significantly less data than the others. In order to eliminate user activity variance effects, a threshold of 60,000 seconds (16.67 hours) of minimum activity was set. This threshold left 67 qualifying users for the evaluation presented in this paper.

Five temporary employees (subjects) were hired during each week of the data collection. The subjects had to work for a total of 40 hours. Each subject was assigned two tasks each day. The first, for six hours of the eight-hour workday, was an open-ended task to write blog-style

Table 1. Character count statistics over five workdays.

Minimum per user	17,027
Maximum per user	263,165
Average	84,206
Total	5,641,788

articles related to the city in which the testing was carried out. The second task, involving two hours of the workday, was less open-ended. Each subject was asked to write summaries from a list of topics or web articles. The articles were from various reputable news sources and were kept consistent between users.

Both tasks encouraged the subjects to conduct extensive online research using web browsers. They were allowed to copy and paste content, but they were told that the final work product had to be their own authorship. As expected, the subjects almost exclusively used two applications: Microsoft Word 2010 for word processing and Internet Explorer for browsing the web. Although the user-generated documents are available in the dataset, the evaluation in this paper is based on the stream of keystrokes recorded during the workday, with the purpose of simulating the settings with which a real-time authentication system would have to operate.

The 67-user dataset was further parsed in order to produce one large stream of mouse/keyboard events. For every user, the entire five days of data was concatenated into one stream (in JSON format), and marked to be divided into five equally-sized folds for later cross-validation evaluation. In addition, periods of inactivity lasting more than two minutes were marked as idle. For the purposes of this research, a subset of events that only included keyboard strokes was maintained (the mouse events will be used by other sensors in future work). The format of one continuous stream permitted the complete utilization of the data in a real-time, continuous active authentication system. Table 1 summarizes the keystroke event statistics for the parsed 67-user dataset. The keystroke events include the alphanumeric keys as well as special keys such as `shift`, `backspace`, `ctrl` and `alt`. The keystroke counts only include the down presses, the key releases were ignored.

4. Methodology

This section describes the methodology in detail, including the challenges and limitations.

4.1 Challenges and Limitations

An active authentication system presents a few concerns. First, a potential performance overhead is expected to accompany the deployment of such a system because it requires constant monitoring and logging of user input, and on-the-fly processing of all its sensor components. With stylometric sensors, large amounts of memory and computation power are often consumed by language processing tools (e.g., dictionary-based features and part-of-speech taggers). Therefore, the system should be carefully configured to balance accuracy and resource consumption. This issue becomes more acute in a multi-modal system that uses multiple sensors.

A second concern with an active authentication system is the user input requirements. In a non-active authentication scheme, a user is required to provide credentials only when logging in and perhaps when certain operations are to be executed. The credentials include some sort of personal key (e.g., password or private key) that identifies the user. However, in the case of an active authentication system based on stylometric modalities, the user keyboard input is required. In a multi-modal system, all user interactions may be required, including mouse events and web browsing behavior. The precise sequence and timing of keyboard events are essential to enhance system performance. However, such input is not designed for stylometric analysis and authentication. Additionally, it often contains sensitive and private information, which is collected when the user types in passphrases to log into accounts, writes something personal or simply browses the web. Some actions may be taken during system design to address security and privacy concerns. For example, the collected data could be managed carefully, storage of raw collected data could be avoided (except for parsed feature vectors extracted from the data) and all stored data could be encrypted. The privacy issue specifically applies to stylometric modalities, where the content of user input is of importance. Other modalities may void these issues by not targeting content, such as mouse movement biometrics that focus on physical characteristics of the user instead of the possibly sensitive semantics of the generated input.

4.2 Previous Evaluation

In earlier work [18], we presented an initial evaluation of a portion of the Active Linguistic Authentication Dataset (i.e., data for fourteen users). Two methods of evaluation were applied.

First, each day's worth of work was analyzed as one unit or document, for a total of 69 documents (five days for fourteen users, minus a missing

day by one user). One-vs.-all analysis was applied using a simple nearest-neighbor classifier with the Manhattan or intersection distance metric and character n-grams as features ($1 \leq n \leq 5$). The best result achieved was 88.4% accuracy.

In the second analysis, a number-of-characters-based sliding window technique was applied to generate the input segments to be classified; this provides a better simulation of the performance of a realistic active stylometric authentication system. The generated windows were non-overlapping, with the window sizes set to 100, 500 and 1,000 words (tokens separated by whitespaces). The minimum window size was used to allow sufficient data for the stylistic profiling of the window. An extensive linguistic feature set, inspired by the one used in the Writeprints [2] stylometry method, was employed along with a linear SVM and a nearest-neighbor classifier. The best result achieved was 93.33% accuracy with 0.009 FAR and 0.067 FRR.

The results demonstrate that it is beneficial to use stylometric biometrics for active authentication. However, the analysis approach, although satisfactory as a preliminary study, omitted some key requirements for an active authentication system. First, only fourteen subjects were used; thus, the system performance for a large set of users remains unknown. Stylometry research has thus far provided solutions for large author sets, but no research to date has focused on a dataset with incoherent and noisy qualities. Indeed, the approaches discussed above may prove to be inefficient for larger author sets.

Perhaps the main issue with the method of analysis is the units determined for learning/classification. Day-based windows are certainly not useful for active authentication, which aims to provide intruder alerts as quickly as possible (in minutes, if not seconds). Even the second data-based-windows analysis is insufficient: each window may have an arbitrary length in terms of its time span and collecting the minimum amount of words may allow an intruder enough time to apply an attack. Moreover, due to the possibility of large windows that may cross idle periods, data associated with different users could be mixed. For example, the first half of a window could contain legitimate user input while the second half, an idle-period later, could be supplied by an intruder. This causes the contamination of "bad" windows with "good" data that could throw off the classifier and cause it to miss an alert.

This paper provides an analysis of the authentication system in a more realistic setting. The focus is on a time-wise (instead of a data-wise) sliding window, and overlapping windows are allowed so that the system can output decisions with increased frequency. With this approach, the system is compelled to decide whether to accept or reject the latest

window in a timely manner based on the data it has acquired thus far, or to determine that it cannot make a decision. The issue of balancing the amount of collected data, the required time-wise size of windows and the desired decision frequency is examined in the following section.

4.3 Real-Time Approach

The stylometric classifiers, or sensors, presented in this section are based on the simplest settings of closed-world stylometry: the classifiers are trained on the closed set of 67 users, where each classification results with one of the users being identified as the author. A more sophisticated approach would use open-world verifiers, where each legitimate user is paired to its own classifier in a one-class/one-vs.-all formulation. Such a verification approach is naturally suited to the open-world scenario, where possible imposters can originate outside the set of legitimate users (e.g., an intruder from outside an office who takes over an unlocked computer, instead of a malicious colleague). However, this paper considers the case of a closed set of possible users as a baseline for future verification-based classifiers.

In the preprocessing phase, the keystroke data files were parsed to produce a list of documents (text windows) consisting of overlapping windows for each user with time-based sizes of 10, 30, 60, 300, 600 and 1,200 seconds. For the first three settings, a sliding window was advanced with steps of 10 seconds of the stream of keystrokes; the last three settings used steps of 60 seconds. The step size determines how often a decision can be made by the sensor. In addition, although the window generation was configured with fixed parameters (e.g., a time-wise size of 300 and step of 60), in practice, the timestamps of the generated windows correlated with the keystroke events by relaxing the generation to no greater than 300 and no less than 60 with empty windows being discarded. In a live system, a similar approach is expected to be used: a window is "closed" and a decision is made when the size limitation time is up; hence, it is no greater than 300. In addition, when determining the beginning of a window followed by another window, a difference of at least one character is expected (otherwise the second window is a subset of the first). Therefore, if the time span between the first character in a window and the one that follows is greater than the determined step size, effectively a greater step size is applied; hence, it is no less than 60.

In this research, the generated windows were set to ignore idle periods, as if the data stream was continuous with no more than two minutes of delay between one input character and the next. This was applied in the dataset by preprocessing the keystroke timestamps such that an idle pe-

riod longer than two minutes was artificially narrowed down to precisely two minutes. Furthermore, the data was aggregated and divided into five equally-sized folds for analysis purposes, thus potentially containing windows with an idle period between days. Although this preprocessing suffers from the problems associated with mixed legitimate/non-legitimate user-input windows or mixed time-of-day windows (e.g., end of one day and beginning of the next day) if applied in a real system, in our analysis, the processing was applied to allow the generation of as many windows as possible. Since the analysis presented does not involve legitimate/non-legitimate mixed windows, idle-crossing windows are reasonable for the purpose of this work.

In the case of stylometry-based biometrics, selecting a window size involves a delicate trade-off between the amount of captured text (and the probability of correct stylistic profiling for the window) and the system response time. In contrast, other biometrics can perform satisfactorily with small windows (with sizes in the order of seconds). The problem is somewhat overcome using small steps (and overlapping windows), leaving the problem only at the beginning of a day (until the first window is generated). Similar to the analysis in [18], only keystrokes were considered during preprocessing (key releases were filtered) and all special keys were converted to unique single-character placeholders. For example, `BACKSPACE` was converted to β and `PRINTSCREEN` was converted to π. Representable special keys such as `\t` and `\n` were taken as is (i.e., tab and newline, respectively).

The chosen feature set is probably the most crucial part of the configuration. The constructed feature set, which we refer to as the AA feature set, is a variation of the Writeprints [2] feature set that includes a vast range of linguistic features across different levels of text. Table 2 summarizes the features of the AA feature set. This rich linguistic feature set can better capture a user's writing style. With the special-character placeholders, some features capture aspects of a user's style that are not found in standard authorship problem settings. For example, frequencies of backspaces and deletes provide an evaluation of a user's typo-rate (or lack of decisiveness).

The features were extracted using the JStylo framework [26]. JStylo was chosen for analysis because it is equipped with fine-grained feature definition capabilities. Each feature is uniquely defined by a set of its own document preprocessing tools, one unique feature extractor (core of the feature), feature post-processing tools and normalization/factoring options. The features available in JStylo are frequencies of a class of related features (e.g., frequencies of "a," "b," ..., "z" for the "letters" feature class) or some numeric evaluation of an input document (e.g., av-

Table 2. AA feature set.

Group	**Features**
Lexical	Average word length
	Characters
	Most common character bigrams
	Most common character trigrams
	Percentage of letters
	Percentage of uppercase letters
	Percentage of digits
	Digits
	Two-digit numbers
	Three-digit numbers
	Word length distribution
Syntactic	Function words
	Part-of-speech (POS) tags
	Most common POS bigrams
	Most common POS trigrams
Content	Words
	Word bigrams
	Word trigrams

erage word length or Yule's Characteristic K). Its output is compatible with the Weka platform [14], which was employed during the classification process. The definition and implementation of all the features in the AA feature set are available in JStylo, making it easily reproducible.

Two important procedures were applied in the feature extraction phase. First, every word-based feature (e.g., function words class or different word-grams) was assigned a tailor-made preprocessing tool developed for the dataset; each tool applied the relevant special characters on the text. For example, the character sequence `ch`$\beta\beta$`Cch`$\beta\beta$`hicago` becomes `Chicago`, where β represents backspace. Second, since the windows are determined by time and not by the amount of collected data as in [18], normalization is crucial for all frequency-based features (which constitute the majority of the feature set). Each of these features was simply divided by the most relevant measurement related to the feature. For example, character bigrams were divided by the total character count of the window.

The classification used sequential minimal optimization support vector machines [31] with a linear kernel and complexity parameter $C = 1$ available in Weka. Support vector machines are commonly used for authorship attribution [1, 21, 43] and are known to provide high performance and accuracy. As mentioned earlier, they are closed-world

classifiers, i.e., they classify each window to one of the known candidate users (with the legitimate user as the true class). No acceptance thresholds were integrated in the classification process.

Finally, the data was analyzed with the stylometric sensors using a varying threshold for minimum characters-per-window spanning from 100 to 1,000 with steps of 100. For every threshold set, all the windows with less than the threshold amount of characters were discarded and the sensors outputted no decisions for these windows. The different thresholds help assess the trade-off in sensor performance in terms of accuracy and availability: as the threshold increases, the window has richer data and is potentially classified with higher accuracy, but the portion of total windows that pass the threshold decreases, making the sensor less available. Note that even the largest threshold (1,000 characters) is considerably smaller than that used in most previous stylometry analyses (minimum of 500 words). After filtering, only configurations with training data available for all users were retained; as expected, this resulted in the removal of sensors configured to small windows with high minimum number of character thresholds.

After eliminating sensors according to the rule mentioned above, 37 stylometric sensors were retained. These sensors spanned a variety of time-wise window sizes and minimum character-wise window sizes. In the remainder of this paper, the stylometric sensors are denoted as $S_{n,m}$, where n denotes the time-wise window size in seconds and m denotes the minimum characters-per-window configuration.

5. Evaluation

The generated user data streams, divided into five equally-sized folds, are intended to be evaluated in a multi-modal decision fusion active authentication system. Such a system requires knowledge of the expected FAR and FRR rates of its sensors in order to make a cumulative weighted decision. Therefore, the intended evaluation is based on five-fold cross-validation where, in each of the five validations, three folds were used for training, one fold for characterization of the FAR and FRR of the sensor, and the last fold for testing. Thus, each of the five validations outputted a decision for each test instance (from the last fold) and a global FAR and FRR characterization of the sensor. Eventually, the results of all five validations were averaged to determine the performance of the system. The configuration of the validations was cyclic, such that, in the first validation, folds 1, 2 and 3 were used for training, fold 4 for characterization and fold 5 for testing; in the second, folds 2, 3 and 4

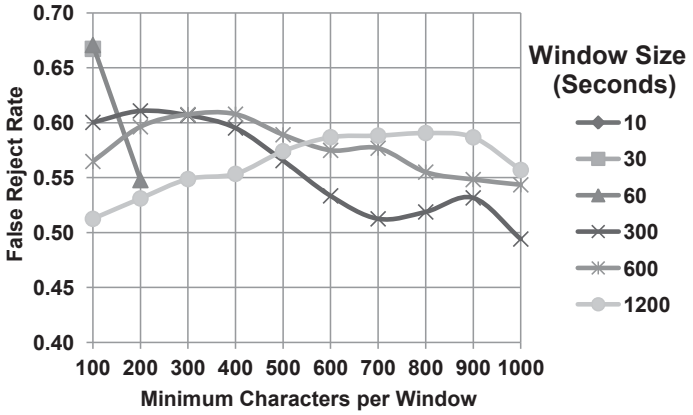

Figure 1. Averaged FRR for all characterization phases.

were used for training, fold 5 for characterization and fold 1 for testing, and so on.

To evaluate the performance of the stylometric sensors, we use the results averaged over all train-characterize-test configurations described above to provide performance evaluations of the sensors when combined in a centralized decision fusion algorithm. Since the FRR and FAR produced in the characterization phase of the main experiments provide an evaluation of the reliability of the decisions made during the test phase, it is reasonable to use them to evaluate the standalone performance of the stylometric sensors.

Figures 1 and 2 show the averaged FRR and FAR results. In particular, Figure 1 shows the averaged FRR for all characterization phases using the stylometric sensors with varying time-wise window sizes and varying thresholds for the minimum number of characters per window. Only windows passing the threshold (i.e., with at least that many characters) participated in the analysis. This measurement accounts for the portion of legitimate user windows that were not detected as belonging to the user (i.e., false alarms).

Figure 2 shows the averaged FAR for all characterization phases using the stylometric sensors with the same configurations. Note that the FAR accounts for the portion of intruder windows that were classified as belonging to a legitimate user (i.e., security breaches).

Figure 3 shows the averaged percentage of the remaining windows after all the windows that did not pass the minimum characters-per-window threshold were removed.

The high FRR and low FAR suggest that the majority of the sensors are rather strict: they almost never falsely identify an intruder as le-

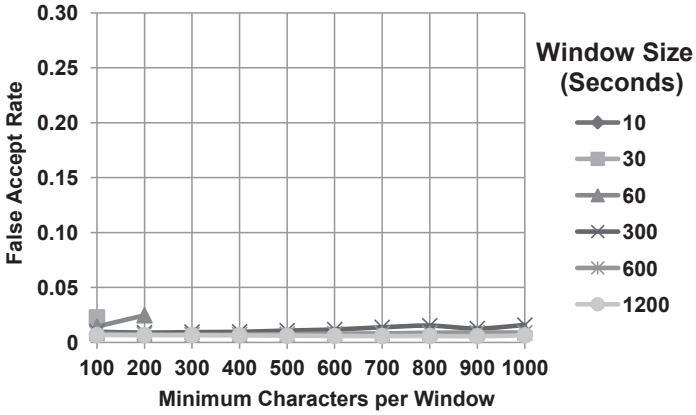

Figure 2. Averaged FAR for all characterization phases.

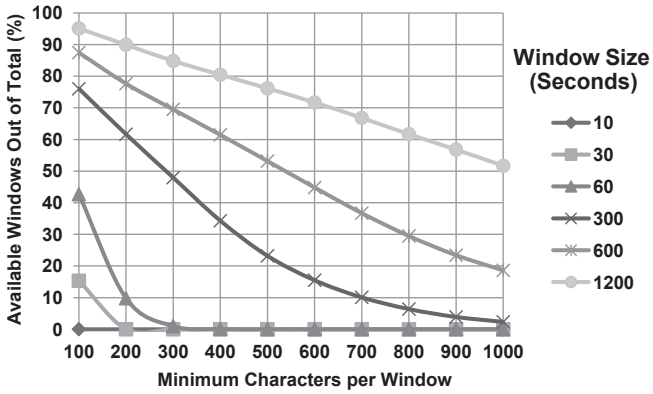

Figure 3. Percentage of remaining windows.

gitimate, but a cost is paid in terms of a high FAR. The FRR results indicate that as the window size (in seconds) increases, the less the minimum characters-per-window threshold affects performance. The same trend is seen with the FAR results: the large windows (300, 600 and 1,200) show insignificant differences for different minimum characters-per-window thresholds.

The availability of decisions as a function of the minimum characters-per-window thresholds (shown in Figure 3) completes the evaluation of the performance of the stylometric sensors. For example, $S_{1200,100}$ triggered every 60 seconds (the step configuration of the 1,200-second-windows sensors) produces a decision 95% of the time with an accuracy of approximately 0.5/0 FRR/FAR.

6. Conclusions

A preliminary study by Juola, *et al.* [18] indicates the effectiveness of stylometric biometrics for active authentication. This paper puts the shortcomings of the preliminary study to the test by using settings that simulate a more realistic active authentication environment, with many users and high frequency decision making constraints. The results obtained under these settings demonstrate that the effectiveness of stylometric sensors deteriorates drastically, down to 0.5 FRR and 0 FAR. Nevertheless, the results are promising because the sensors may be used in a mixture-of-experts approach that fuses multi-modal sensors.

Although the configuration of data with overlapping sliding windows is realistic, the classification methodology is still limited and focused on closed-world SVM classifiers with an extensive linguistic feature set. Future analysis must include other classifiers, especially open-world verifiers that can be applied in scenarios where the set of suspects is not closed. In addition, because of the noisiness of the data, other feature sets should be considered, including sets that focus less on high linguistic characteristics of the text (like POS-taggers) and more on typing patterns. A mixture of writing style and typing style quantification could, perhaps, achieve better profiling with this type of data.

The immediate next step in evaluation is to use a multi-modal fusion system employing multiple sensors. Sensors to be considered include mouse movements, keyboard dynamics (i.e., low-level key patterns) and web-browsing behavior. The configuration of such a system based on closed-world sensors and data-based windows evaluated on a subset of nineteen users achieves approximately 1% FAR/FRR, but its performance in open-world settings with the complete dataset is as yet unknown and is currently the subject of investigation.

References

[1] A. Abbasi and H. Chen, Identification and comparison of extremist-group web forum messages using authorship analysis, *IEEE Intelligent Systems*, vol. 20(5), pp. 67–75, 2005.

[2] A. Abbasi and H. Chen, Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace, *ACM Transactions on Information Systems*, vol. 26(2), pp. 7:1–7:29, 2008.

[3] A. Ahmed and I. Traore, A new biometric technology based on mouse dynamics, *IEEE Transactions on Dependable and Secure Computing*, vol. 4(3), pp. 165–179, 2007.

[4] K. Ali and M. Pazzani, On the Link Between Error Correlation and Error Reduction in Decision Tree Ensembles, Department of Information and Computer Science, University of California at Irvine, Irvine, California, 1995.

[5] S. Argamon, M. Koppel, J. Pennebaker and J. Schler, Mining the blogosphere: Age, gender and the varieties of self-expression, *First Monday*, vol. 12(9), 2007.

[6] S. Argamon, M. Koppel, J. Pennebaker and J. Schler, Automatically profiling the author of an anonymous text, *Communications of the ACM*, vol. 52(2), pp. 119–123, 2009.

[7] N. Bakelman, J. Monaco, S. Cha and C. Tappert, Continual keystroke biometric authentication on short bursts of keyboard input, *Proceedings of the Student-Faculty Research Day*, Seidenberg School of Computer Science and Information Systems, Pace University, New York, 2012.

[8] F. Bergadano, D. Gunetti and C. Picardi, User authentication through keystroke dynamics, *ACM Transactions on Information Systems Security*, vol. 5(4), pp. 367–397, 2002.

[9] J. Binongo, Who wrote the 15th Book of Oz? An application of multivariate analysis of authorship attribution, *Chance*, vol. 16(2), pp. 9–17, 2003.

[10] Z. Chair and P. Varshney, Optimal data fusion in multiple sensor detection systems, *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22(1), pp. 98–101, 1986.

[11] C. Chaski, Who's at the keyboard: Authorship attribution in digital evidence investigations, *International Journal of Digital Evidence*, vol. 4(1), 2005.

[12] C. Chaski, The keyboard dilemma and forensic authorship attribution, in *Advances in Digital Forensics III*, P. Craiger and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 133–146, 2007.

[13] C. Gray and P. Juola, Personality identification through on-line text analysis, presented at the *Chicago Colloquium on Digital Humanities and Computer Science*, 2012.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, The Weka Data Mining Software: An update, *SIGKDD Explorations Newsletter*, vol. 11(1), pp. 10–18, 2009.

[15] S. Hashem and B. Schmeiser, Improving model accuracy using optimal linear combinations of trained neural networks, *IEEE Transactions on Neural Networks*, vol. 6(3), pp. 792–794, 1995.

[16] M. Jockers and D. Witten, A comparative study of machine learning methods for authorship attribution, *Literary and Linguistic Computing*, vol. 25(2), pp. 215–223, 2010.

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

[18] P. Juola, J. Noecker, A. Stolerman, M. Ryan, P. Brennan and R. Greenstadt, Towards active linguistic authentication, in *Advances in Digital Forensics IX*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 385–398, 2013.

[19] P. Juola, M. Ryan and M. Mehok, Geographically localizing tweets using stylometric analysis, presented at the *American Association for Corpus Linguistics Conference*, 2011.

[20] J. Kittler, M. Hatef, R. Duin and J. Matas, On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20(3), pp. 226–239, 1998.

[21] M. Koppel and J. Schler, Ad-hoc authorship attribution competition – Approach outline, presented at the *Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[22] M. Koppel and J. Schler, Authorship verification as a one-class classification problem, *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

[23] M. Koppel, J. Schler and S. Argamon, Computational methods in authorship attribution, *Journal of the American Society for Information Science and Technology*, vol. 60(1), pp. 9–26, 2009.

[24] M. Koppel, J. Schler and K. Zigdon, Determining an author's native language by mining a text for errors, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 624–628, 2005.

[25] O. Kukushkina, A. Polikarpov and D. Khmelev, Using literal and grammatical statistics for authorship attribution, *Problemy Peredachi Informatii*, vol. 37(2), pp. 96–198, 2001; Translated in *Problems of Information Transmission*, vol. 37(2), pp. 172–184, 2001.

[26] A. McDonald, S. Afroz, A. Caliskan, A. Stolerman and R. Greenstadt, Use fewer instances of the letter "i:" Toward writing style anonymization, in *Privacy Enhancing Technologies*, S. Fischer-Hubner and M. Wright (Eds.), Springer-Verlag, Berlin, Germany, pp. 299–318, 2012.

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

[28] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin and D. Song, On the feasibility of Internet-scale author identification, *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 300–314, 2012.

[29] M. Obaidat and B. Sadoun, Verification of computer users using keystroke dynamics, *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 27(2), pp. 261–269, 1997.

[30] T. Ord and S. Furnell, User authentication for keypad-based devices using keystroke analysis, *Proceedings of the Second International Network Conference*, pp. 263–272, 2000.

[31] J. Platt, Fast training of support vector machines using sequential minimal optimization, in *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. Burges and A. Smola (Eds.), MIT Press, Cambridge, Massachusetts, pp. 185–208, 1999.

[32] J. Rudman, The state of authorship attribution studies: Some problems and solutions, *Computers and the Humanities*, vol. 31, pp. 351–365, 1998.

[33] D. Shanmugapriya and G. Padmavathi, A survey of biometric keystroke dynamics: Approaches, security and challenges, *International Journal of Computer Science and Information Security*, vol. 5(1), pp. 115–119, 2009.

[34] T. Sim, S. Zhang, R. Janakiraman and S. Kumar, Continuous verification using multimodal biometrics, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29(4), pp. 687–700, 2007.

[35] E. Stamatatos, A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, vol. 60(3), pp. 538–556, 2009.

[36] E. Stamatatos, On the robustness of authorship attribution based on character n-gram features, *Brooklyn Law School Journal of Law and Policy*, vol. 21(2), pp. 421–439, 2013.

[37] H. van Halteren, Author verification by linguistic profiling: An exploration of the parameter space, *ACM Transactions on Speech and Language Processing*, vol. 4(1), pp. 1:1–1:17, 2007.

[38] H. van Halteren, R. Baayen, F. Tweedie, M. Haverkort and A. Neijt, New machine learning methods demonstrate the existence of a human stylome, *Journal of Quantitative Linguistics*, vol. 12(1), pp. 65–77, 2005.

[39] J. Wayman, Fundamentals of biometric authentication technologies, *International Journal of Image and Graphics*, vol. 1(1), pp. 93–113, 2001.

[40] J. Wayman, N. Orlans, Q. Hu, F. Goodman, A. Ulrich and V. Valencia, Technology Assessment for the State of the Art Biometrics Excellence Roadmap, Volume 2, MITRE Technical Report v1.3, The MITRE Corporation, Bedford, Massachusetts, 2009.

[41] R. Yampolskiy, Behavioral modeling: An overview, *American Journal of Applied Sciences*, vol. 5(5), pp. 496–503, 2008.

[42] N. Zheng, A. Paloski and H. Wang, An efficient user verification system via mouse movements, *Proceedings of the Eighteenth ACM Conference on Computer and Communications Security*, pp. 139–150, 2011.

[43] R. Zheng, J. Li, H. Chen and Z. 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*, vol. 57(3), pp. 378–393, 2006.