Active Linguistic Authentication Revisited: Real-Time Stylometric Evaluation towards Multi-Modal Decision Fusion

Ariel Stolerman
Drexel University
Philadelphia, PA
stolerman@cs.drexel.edu

Alex Fridman
Drexel University
Philadelphia, PA
af59@cs.drexel.edu

Rachel Greenstadt
Drexel University
Philadelphia, PA
greenie@cs.drexel.edu

Patrick Brennan
Juola & Associates
Pittsburgh, PA
pbrennan@juolaassoc.com

Patrick Juola
Juola & Associates
Pittsburgh, PA
pjuola@juolaassoc.com

Abstract

Active authentication is the process of continuously verifying a user based on his/her ongoing interaction with the computer. Forensic stylometry is the study of linguistic style, applied to author (user) identification. We evaluate the Active Linguistic Authentication Dataset [Juola et al., 2013], collected from users working individually in an office environment for a period of one week. We consider a battery of stylometric modalities, as a representative collection of high-level behavioral biometrics. As opposed to the initial evaluation presented on this dataset before on 14 users, we consider the fully collected dataset, which consists of data by 67 users. An additional significant difference is in the type of evaluation: instead of a day-based, or data-based (number-of-characters) windows considered for classification, we evaluate time-based, overlapping sliding windows; our evaluation tests the ability to produce authentication decisions every 10–60 seconds, highly applicable to real-world active security systems. We evaluate the different sensors via cross-validation, measuring false acceptance and rejection rates (FAR & FRR). We show that under these realistic settings, stylometric sensors perform with considerable effectiveness down to 0/0.5 FAR/FRR, for decisions produced every 60 seconds, available 95% of the time. This work is considered towards a decision-fusion approach, that undertakes multiple modalities (e.g. keyboard and mouse dynamics) for making centralized, highly accurate authentication decisions.

Keywords. Active Authentication, Forensic Stylometry, Authorship Attribution, Authorship Verification, Biometrics.

1 Introduction

The challenge of identity verification for the purpose of access control is the tradeoff between maximizing the probability of intruder detection, and minimizing the cost for the legitimate user in time and distractions due to false alerts, and extra hardware requirements for physical biometric authentication. In recent years, behavioral biometric systems have been explored extensively in addressing this challenge [Ahmed and Traore, 2007a]. These systems rely on input devices such as the keyboard and mouse that are already commonly available with most computers, and are thus low cost in terms of having no extra equipment requirements. However, their performance in terms of detecting intruders, and maintaining a low-distraction human-computer interaction (HCI) experience has been mixed [Bergadano et al., 2002], showing error rates ranging from 0% [Obaidat and Sadoun, 1997] to 30% [Ord and Furnell, 2000] depending on context, variability in task selection, and various other dataset characteristics.

The bulk of biometric-based authentication work focused on verifying a user based on a static set of data. This type of one-time authentication is not sufficiently applicable to a live multi-user environment, where a person may leave the computer for an arbitrary period of time without logging off. This context necessitates continuous authentication when a computer is in a non-idle state. In particular, to represent this general real-world scenario, we used the Active Linguistic Authentication Dataset [Juola et al., 2013]. This dataset consists of data collected in a simulated office environment, which contains behavioral biometrics associated with typical human-computer interaction (HCI) by an office worker.

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, the field is currently dominated by AI techniques such as neural networks and statistical pattern recognition. State-of-the-art stylometry approaches can identify individuals in sets of 50 authors with over 90% accuracy [Abbasi and Chen, 2008a], and even scaled to over 100,000 authors [Narayanan et al., 2012]. Stylometry is currently used in intelligence analysis and forensics, with increasing interest for digital communication analysis [Wayman et al., 2009]. It is even accurate enough to be admissible as legal evidence [Chaski, 2005, Chaski, 2007]. The application of stylometry as a high-level modality for authenticating users in a continuous user verification system is novel; initial evaluation of authorship attribution technologies are proven promising, reaching more than 90% identification accuracy over 14 users [Juola et al., 2013].

In this paper, we consider a set of stylometric classifiers, also referred to as sensors, as a representative selection of high-level behavioral biometrics. This work aims to evaluate authorship attribution approaches in more realistic settings for active authentication, which require constant monitoring and frequent decision making about the legitimacy of the user at the computer in a dynamic and time-constrained environment. Moreover, this work is designed as a preliminary evaluation of one modality among many to consider for an active authentication system. The main goal for the stylometric modalities presented here in future work is to be interleaved with other low- and high-level modalities, such as keyboard dynamics [Shanmugapriya and Padmavathi, 2009], mouse movements [Ahmed and Traore, 2007b], web browsing behavior [Yampolskiy, 2008] and the like, in one centralized decision fusion system. Usage of such modalities, stylometry included, may provide a cost-effective alternative to sensors based on physiological biometrics [James, 2001].

Although this work is targeted for active authentication, a live security application of stylometric analysis, its implications on the usability and configuration of stylometric sensors are relevant for forensic contexts as well: consider a standard post-mortem forensic analysis of user input data aggregated throughout an entire day; this work lays grounds for what features to look at in such "noisy" settings, the size of windows to look at, the effects of looking at overlapping windows, how idle periods in data input should be considered, etc.

The remainder of the paper is structured as follows. Sec. 2 reviews background and related work. Sec. 3 discusses the simulated work environment dataset that we used for evaluation. The stylometry biometrics applied on the dataset are discussed in Sec. 4, followed by evaluation in Sec. 5. We conclude the paper in Sec. 6 and discuss directions for future work.

2 Related Work

A defining problem of active authentication arises from the fact that a verification of identity must be carried out continuously on a sample of sensor data that varies drastically with time. The classification therefore has to be made based on a "window" of recent data, dismissing or heavily discounting the value of older data outside that window. Depending on what task the user is engaged in, some of the biometric sensors may provide more data than others. For example, as the user browses the web, mouse and web activity sensors will be actively flooded with data, while keystroke dynamics or stylometric sensors may only get a few infrequent key presses. This motivates the recent work on multimodal authentication systems where the decisions of multiple classifiers are fused together [Sim et al., 2007]. In this way, the verification process is more robust to the dynamic mode of real-time HCI. In this paper we examine only the effectiveness of stylometry sensors under active authentication settings. The main goal of this work is to be combined in a multi-modal biometric system. The idea of decision fusion is motivated by the work in [Ali and Pazzani, 1995] that greater reduction in error rates is achieved when the classifiers are distinctly different (i.e. using different behavioral biometrics), with several fusion approaches available to be applied [Kittler et al., 1998, Hashem and Schmeiser, 1995, Chair and Varshney, 1986].

Authorship attribution based on linguistic style, or Stylometry, is a well-researched field [Argamon et al., 2009, Rudman, 1998, Juola, 2006, Koppel et al., 2009, Stamatatos, 2009, Jockers and Witten, 2010]. The main domain it is applied on is written language – identifying an anonymous author of a text by mining it for linguistic features. The theory behind stylometry is that everyone has a unique linguistic style ("stylome" [van Halteren et al., 2005]) that can be quantified and measured in order to distinguish between different authors. The feature space is potentially endless, with frequency measurements or numeric evaluations based on features across different levels of the text, including function words [Mosteller and Wallace, 1964, Binongo, 2003], grammar [Kukushkina et al., 2001], character n-grams [Stamatatos, ming] and more. Although stylometry has not been used for active user authentication, its application to this sort of task brings higher level inspection into the process, compared to other lower level biometrics like mouse movements or keyboard dynamics [Zheng et al., 2011, Bakelman et al., 2012].

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 some unique author.

In an active authentication setting, authorship verification is applied, where unknown text is classified by a unary author-specific classifier. The text is attributed to an author if and only if it is stylistically close enough to that 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, it should be considered as an intruder.

Another usage of stylometry is in author profiling [Koppel et al., 2005, Argamon et al., 2009, van Halteren, 2007, Gray and Juola, 2011, Juola et al., 2011] rather than recognition. Writings are mined for linguistic features in order to identify characteristics of their author, like age, gender, native language etc.

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[1], some methods achieved more than 80% accuracy on a set of 241 documents, sometimes with added distractor authors.

In an active authentication setting, a few challenges arise. First, open-world stylometry is a much harder problem, with a tendency to high false-negative (false reject) rates. The unmasking technique [Koppel and Schler, 2004b] has been shown effective on a dataset of 21 books of 10 different 19^{th}-century authors, obtaining 95.7% accuracy. However, the amount of data collected by sliding windows of sufficiently small durations required for an efficient authentication system, along with the lack of quality coherent literary writings make this method perform insufficiently for our goal. Second, the inconsistent frequency nature of keyboard input along with the relatively large amount of data required for good performance of stylometric techniques make a large portion of the input

[1] http://pan.webis.de

windows unusable for learning writing style.

On the other hand, this type of setting allows some advantages in potential features and analysis method. Since the raw data consists of all keystrokes, some linguistic and technical idiosyncratic features can be extracted, like misspellings caught prior to being potentially auto-corrected and vanished from the dataset, or patterns of deletions (selecting a sentence and hitting delete versus repeatedly hitting backspace deleting character at-a-time). In addition, it is more intuitive in this kind of setting to consider overlap between consecutive windows, resulting with a large dataset, grounds for local voting based on a set of windows and control of the frequency in which decisions are outputted by the system.

3 Dataset

We utilize the complete Active Linguistic Authentication Dataset (presented initially in [Juola et al., 2013] while still being collected), a dataset designed specifically for the purpose of behavioral biometrics evaluation, based on data collected in a simulated work environment. For the collection of the data, an office space was allocated, organized and supervised by a subset of the authors. The space contained 5 desks, each with a laptop, mouse and headphones. This equipment and supplies were chosen to be representative of a standard office workplace. One of the important properties of this dataset is that of uniformity. Due to the fact that the computers and input devices in the simulated office environment were identical, the variation in behavioral biometrics data can be more confidently attributed to variation in characteristics of the users, rather than effects of variations in physical environmental settings.

The complete dataset contains data collected from 80 users. Due to crashes in the mouse, keyboard, web browser tracking software, or sick days taken, a few more than 80 subjects participated, to cover the missing data and reach the 80 users goal. However, within the final 80 users data, some users had significantly less data than the rest. In order to eliminate user activity variance effects on our evaluation, we set a threshold of 60,000 seconds minimum activity (16.67 hours). This filtering left us with 67 qualifying users for the evaluation presented in this paper.

During each week of the data collection, 5 temporary employees were hired for a total of 40 hours of work. Each day they were assigned two tasks. The first was an open-ended blogging task, where they were instructed to write blog-style articles related in some way to the city in which the testing was carried out. This task was allocated 6 hours of the 8 hour workday. The second task was less open-ended. Each employee was given a list of topics or web articles to write a summary of. The articles were from a variety of reputable news sources, and were kept consistent between users except for a few broken links due to the expired lifetime of the linked pages. This second task was allocated 2 hours of the 8 hour workday.

Min. per user	17,027
Max. per user	263,165
Avg.	84,206
Total	5,641,788

Table 1: Character count statistics on the 67-user dataset across all 5 work days.

Both tasks encouraged the workers to do extensive online research by using the web browser. They were allowed to copy and paste content, but they were instructed that the final work they produced was to be of their own authorship. As expected, the workers 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 throughout the work day, with the purpose of simulating the settings which a real-time authentication system will have to work under.

The 67-user dataset is further parsed in order to provide one large stream of mouse/keyboard events. For every user, the entire 5 days of data were concatenated into one stream (in JSON format), and marked to be divided into 5 equally sized folds, for later cross-validation evaluation. In addition, any inactivity period exceeding 2 minutes is marked, to be considered as idle. For the purposes of this paper, a subset of events including only keyboard strokes are kept (whereas mouse events are to be used by other sensors in future work). The format of one continuous stream allows to utilize the data to its full in evaluation of a real-time, continuous active authentication system. Keystroke events statistics for the parsed 67-user dataset are summarized in Tab. 1. The keystroke events include both the alpha-numeric keys and also special keys such as `shift`, `backspace`, `ctrl`, `alt`, etc. In counting the key presses in Tab. 1, we count just the down press and not the release.

4 Methodology

4.1 Challenges and Limitations

An active authentication system presents a few concerns. First, a potential performance overhead is expected to accompany deployment of such a system, as 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 may be consumed by language processing tools (e.g. dictionary based features, part-of-speech taggers etc.), therefore a careful configuration should be applied to balance the tradeoff between the accuracy of the system and its expected resource consumption behavior. This issue becomes more prominent in a multi-modal system, where multiple sensors are used.

Another concern with this type of authentication system is its user input requirements. In non-active authentication schemes, the user is required to provide credentials only when logging in, and perhaps when certain operations are to be executed. The provided credentials consist of some sort of personal key (password, private key etc.), dedicated for the purpose of identifying the system's users. In active authentication systems based on stylometric modalities, all of the user keyboard input is required. In a multi-modal system, all interaction may be required, including mouse events and web browsing behavior. The precise sequence and timing of keyboard events is essential for the system's performance. However, this type of input is not designed for stylometric analysis and authentication, and in most probability contains sensitive and private information, collected when the user types in passphrases to log into accounts, writes something personal s/he wishes to keep confidential, or simply browses the web. To cope with these security and privacy issues, some actions can be taken in the design of such a system: the collected data should be managed carefully, by avoiding storage of raw collected data (i.e. save only parsed feature vectors extracted from the data) and use encrypted storage for the data that is stored. The privacy issue specifically applies to stylometric modalities, where the contents of the user input is of importance. Other modalities that can be applied may void these issues by not targeting the content, like mouse movement biometrics that focus on physical characteristics of the user rather than the possibly sensitive semantics of the generated input.

4.2 Previous Evaluation

In [Juola et al., 2013] we presented an initial evaluation of part of the Active Linguistic Authentication Dataset, when data of only 14 users was available (as the dataset was still in collection). 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 (5 days for 14 users, minus a missing day by one user). One-vs-all analysis was applied, using a simple nearest-neighbor classifier with Manhattan or Intersection distance metric, using character n-grams as features (n between 1 and 5). The best result achieved was as high as 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, to better simulate the performance of a realistic active stylometric authentication system. The generated windows were non-overlapping, with window sizes set to 100, 500 and 1,000 words (tokens separated by whitespace). The motivation for requiring a minimum window size is in order to allow sufficient data required for stylistic profiling of the window. An extensive linguistic feature set, inspired by that used in the Writeprints [Abbasi and Chen, 2008b] stylometry method, was used, along with a linear SVM and a nearest-neighbor classifier. The best result achieved was 93.33% accuracy with 0.009/0.067 FAR/FRR.

These reported results were sufficient to determine that using stylometric biometrics for active authentication is beneficial; however, the approach taken in this analysis, although satisfactory as preliminary results, lacks addressing a few key requirements in an active authentication system. First, only 14 subjects were used for the initial analysis, and its performance over a large set of users remains unknown. Stylometry research has thus far provided solutions for large author sets, but even those were never attempted on a dataset with such incoherent and noisy qualities; the performance of the approaches above may certainly be proven inefficient when the author set increases in size.

Perhaps the main issue with this 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 alert as quickly as possible (in a time frame of minutes, perhaps seconds). Even the second data-based-windows analysis is insufficient: each window may have an arbitrary length in time on which it spans, and collecting the minimum amount of words may allow an intruder enough time to apply his attack. Moreover, due to the possibility of time-wise long windows, which may cross idle periods, data of different users can be mixed (e.g. first half of the window is the legitimate user input, whereas the second half, an idle-period later, is by an intruder) causing contamination of "bad" windows with "good" data, which may throw off the classifier and cause it to miss an alert.

In this paper we provide an analysis in a more realistic setting of the authentication system. We focus on a time-wise sliding window (rather than data-wise), and allow overlapping windows in order to provide the system the ability to output frequent decisions. With this approach, the system is compelled to decide whether to accept/reject the latest window in a timely manner, based on the data it has acquired thus far, or to determine it cannot make a decision. A balance between the amount of collected data, the required time-wise size of windows and desired decision frequency is inspected 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: we use classifiers trained on the closed set of all 67 users, where each classification results with one of those users 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 verification approach is more naturally suited for this open-world scenario, where possible imposters can originate outside the set of legitimate users (e.g. an intruder from outside an office that takes over an unlocked computer, rather than a vicious colleague); however, in this paper we consider the case of a closed set of possible users, as a baseline for future verification-based classifiers.

In the preprocessing phase, we parsed the keystrokes data files to produce a list of documents (text windows) consisting of overlapping windows for each user, with the following time-based sizes in seconds: 10, 30, 60, 300, 600 and 1,200. For the first 3 settings we advanced a sliding window with steps of 10 seconds of the stream of keystrokes, and for the last 3 – steps of 60 seconds. The step size determines how often a decision can be made by the sensor. In addition, although the window generation is configured with a fixed time-wise size and step, e.g. $\{300, 60\}$, in practice the timestamps of the generated windows correlate with the keystroke events, by relaxing the generation to $\{\leq 300, \geq 60\}$ (empty windows are discarded.) In a live system a similar approach is expected to be used: a window is "closed" and a decision is made for it when the size limitation time is up, hence ≤ 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 simply 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 will be applied, hence ≥ 60.

In this paper, the generated windows are set to ignore idle periods, as if the stream of data is continuous with no more than 2 minutes delay between one input character and the next. This is applied in the dataset by preprocessing the keystroke timestamps, such that any idle period longer than 2 minutes is artificially narrowed down to precisely 2 minutes. Furthermore, the data is aggregated and divided into 5 equally-sized folds for analysis purposes, thus potentially contains windows originally contain an idle period between days. Although this preprocessing suffers from the issues of possible mixed legitimate/non-legitimate user-input windows, or mixed time-of-day windows (e.g. end of one day and beginning of the next) if applied in a real system, in our analysis it is applied to allow generating as many windows as possible. Since the analysis presented here is not applied on legitimate/non-legitimate mixed windows, idle-crossing windows are reasonable for the purposes of this paper.

Specifically for stylometry-based biometrics, selecting the size of the window affects a delicate tradeoff between the amount of captured text (and probability for correct stylistic profiling of that window) and response time of the system, whereas other biometrics can perform satisfactorily with small windows (even the size of seconds). This is somewhat overcome by using small steps (and overlapping windows), leaving this as a problem only at the beginning of the day (until the first window is generated). Similar to the analysis in [Juola et al., 2013], during preprocessing only keystrokes were taken (key releases were filtered out) and all special keys were converted to unique single-character placeholders. For instance `BACKSPACE` was converted to β and `PRINTSCREEN` was converted to π. Any representable special keys like `\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, denoted the AA feature set hereinafter, is a variation of the *Writeprints* [Abbasi and Chen, 2008b] feature set, which includes a vast range of linguistic features across different levels of text. A summarized description of the features is presented in Tab. 2. By using a rich linguistic feature set we are able to better capture the user's writing style. With the special-character placeholders, some features capture aspects of the user's style usually not found in standard authorship problem settings. For instance, frequencies of backspaces and deletes provide some evaluation of the user's typo-rate (or lack of decisiveness).

The features were extracted using the *JStylo* framework [2] [McDonald et al., 2012], an open-source authorship attribution platform developed in the Privacy, Security and Automation Laboratory at Drexel University. JStylo was chosen for analysis since 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 (the core of the feature), feature post-processing tools and normalization/factoring options. The features available in JStylo are either frequencies of a class of related features (e.g. frequencies of "a", "b", ..., "z" for the "letters" feature class) or some numeric evaluation of the input document (e.g. average word length, or Yule's Characteristic K). Its output is compatible with the data mining and machine learning platform Weka [Hall et al., 2009], which we utilized for the classification process. Definition and implementation of all the features the AA feature set consists of is available in JStylo, making it easily reproducible.

Two important processing procedures were applied in the feature extraction phase. First, every word-based feature (e.g. the function words class, or different word-grams) was applied a tailor-made preprocessing tool developed for this unique dataset, that applies the relevant special characters on the text. For instance, 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 amount of collected data as in [Juola et al., 2013], normalization is crucial for all frequency-based features (which consist the majority of the feature set). These features were simply divided by the most relevant measurement related to the feature. For instance, character bigrams were divided by the total character count of the window.

Group	Features
Lexical	Avg. word-length
	Characters
	Most common character bigrams
	Most common character trigrams
	Percentage of letters
	Percentage of uppercase letters
	Percentage of digits
	Digits
	2-digit numbers
	3-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

Table 2: The AA feature set. Inspired by the *Writeprints* [Abbasi and Chen, 2008b] feature set, includes features across different levels of the text. Some features are normalized frequencies of feature classes; others are numerical evaluations of the input text.

For classification we used sequential minimal optimization (SMO) support vector machines [Platt, 1998] with a linear kernel and complexity parameter $C = 1$, available in Weka. Support vector machines are commonly used for authorship attribution [Abbasi and Chen, 2005,Koppel and Schler, 2004a,Zheng et al., 2006] and known to achieve high performance and accuracy. As mentioned earlier, these are closed-world classifiers, i.e. classify each window to one of the known candidate users (with the legitimate user as the true class). No acceptance thresholds are integrated in the classification process.

Finally, the data was analyzed with the stylometry sensors using a varying threshold for minimum characters-per-window to consider, spanning from 100 to 1000 with steps of 100. For every threshold set, all windows with less than that amount of characters were thrown away, and for those windows the sensors output no decision. The different thresholds allow assesing the tradeoff in the sensor's performance in terms of accuracy and availability: as the threshold increases, the window is richer with data and will po-

[2] http://psal.cs.drexel.edu/

tentially be 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 (1000 *characters*) is considerably smaller than used in most previous stylometry analyses – a minimum of 500 *words*. After filtering, only configurations with training data available for *all* users are kept, which expectedly yielded removal of sensors configured to small windows with high minimum number of characters thresholds.

After removal according to the rule above, 37 stylometry sensors are kept that span over a variety of time-wise window size and minimum character-wise window size. For the rest of the paper, the stylometry 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 5 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/FRR rates of its different sensors, in order to make a cumulative weighted decision. Therefore the intended evaluation is based on 5-fold cross validation, where in each of the 5 validations, 3 folds are used for training, 1 fold is used for characterization of the sensors' FAR/FRR, and the last fold is used for testing. Thus each of the 5 validations outputs a decision for each test instance (from the last fold) and a global FAR/FRR characterization of the sensor in that validation. Eventually, the results of all 5 validations are averaged to determine the performance of the system. The configuration of the validations is cyclic, such that in the first folds 1, 2 and 3 are used for training, 4 for characterization and 5 for testing; in the second, 2, 3 and 4 are used for training, 5 for characterization and 1 for testing, and so on.

To evaluate the performance of the different stylometry sensors, we use the results averaged over all train-characterize-test configurations described above, in order to provide performance evaluation of the sensors when combined in a centralized decision fusion algorithm. Since the false reject rate (FRR) and false accept rate (FAR) produced in the characterization phase of the main experiments provide an evaluation of the reliability of the decisions made in the test phase, it is reasonable to use them to evaluate the standalone performance of the stylometry sensors. Averaged FRR and FAR results are shown in Fig. 1 and Fig. 2, respectively. Fig. 3 illustrates the averaged percentage of remaining windows, after removing all those that do not pass the minimum characters-per-window threshold.

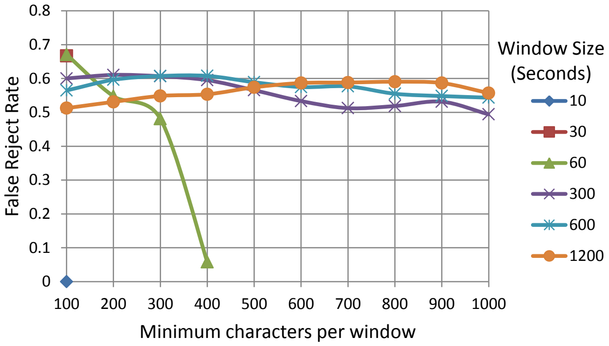

Figure 1: Averaged false reject rate (FRR) for all characterization phases using the stylometry sensors with varying time-wise window sizes and varying threshold for minimum number of characters per window. Only windows that pass the threshold (i.e. contain at least that many characters) participated in the analysis. This measurement accounts for the portion of legitimate user's windows that were not detected as the user's, i.e. false alarms.

The high FRR and low FAR suggest that the majority of the sensors are rather strict, i.e. they almost never falsely identify an intruder as legitimate, but in the price of a high false-alarm rate. The FRR results indicate that as the window size (in seconds) increases, the less the minimum characters-per-window threshold affects performance. Same trend is seen with the FAR results: the large windows (300, 600 and 1,200) show insignificant differences across varying minimum characters-per-window thresholds.

The availability of decisions as a function of the minimum characters-per-window thresholds illustrated in Fig. 3 completes the image of how the stylometry sensors perform. For instance, $S_{1200,100}$, triggered every 60 seconds (the step configuration of the 1200-second-windows sensors), will produce a decision 95% of the time, with accuracy of approx. 0.5/0 FRR/FAR.

6 Conclusion

In [Juola et al., 2013], a proof of concept is given for the effectiveness of stylometric biometrics in an active authentication system; however, the shortcom-

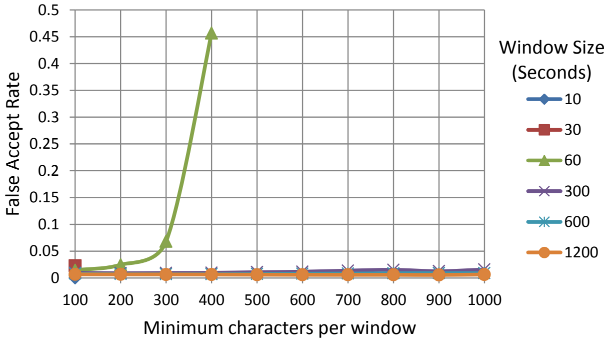

Figure 2: Averaged false accept rate (FAR) for all characterization phases using the stylometry sensors, with the same configurations as described in Fig. 1. The FAR accounts for the portion of intruder windows that were classified as the legitimate user's, i.e. security breaches.

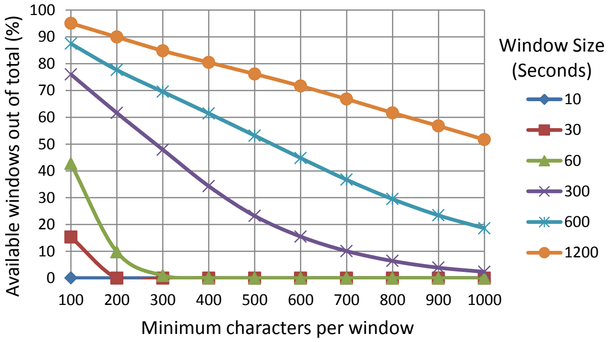

Figure 3: Percentage of remaining windows out of the total windows after filtering by the minimum characters-per-window threshold.

ings of this preliminary evaluation are put to the test here, with settings simulating a more realistic active authentication environment, with many users and high frequency decision making constraints. We have shown that under such settings, the effectiveness of stylometric sensors deteriorates drastically, down to 0.5 false rejection and 0 false acceptance rates. Nevertheless, these results are yet promising, and if not on their own, such sensors may be used in a mixture-of-experts approach that fuses multi-modal sensors.

Moreover, although the configuration of the data with overlapping sliding windows is more realistic than presented before, the classification methodology is still rather limited, focused on closed-world SVM classifiers with an extensive linguistic feature set. Future analysis must include other classifiers, specifically open-world verifiers that can be applied in scenarios where the set of suspects is not closed. In addition, due to the nosiness of the data, other feature sets should be attempted, perhaps those that focus less on high linguistic characteristics of the text (like POS-taggers), and more on typing patterns. Perhaps a mixture of writing style and *typing* style quantification can achieve better profiling of this type of data.

The immediate next step of evaluation is interleaving the results presented in this paper in a multimodal fusion system, that undertakes multiple sensors for a centralized decision. Other sensors to be considered include mouse movements, keyboard dynamics (i.e. low-level key patterns) and web-browsing behavior, all collected and available in the Active Linguistic Authentication dataset. Configuration of such a system based on closed-world sensors and data-based windows, evaluated on a subset of 19 users, has been able to achieve ≈ 1% FAR/FRR, however its performance in open-world settings with the complete dataset is yet unknown, and is currently work in progress.

References

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

[Abbasi and Chen, 2008a] Abbasi, A. and Chen, H. (2008a). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):1–29.

[Abbasi and Chen, 2008b] Abbasi, A. and Chen, H. (2008b). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):7:1–7:29.

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

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

[Ali and Pazzani, 1995] Ali, K. and Pazzani, M. (1995). *On the link between error correlation and* *error reduction in decision tree ensembles.* Citeseer.

[Argamon et al., 2009] Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *CACM*, 52(2):119–123.

[Bakelman et al., 2012] Bakelman, N., Monaco, J. V., Cha, S.-H., and Tappert, C. C. (2012). Continual keystroke biometric authentication on short bursts of keyboard input. In *Proceedings of Student-Faculty Research Day, CSIS, Pace University.*

[Bergadano et al., 2002] Bergadano, F., Gunetti, D., and Picardi, C. (2002). User authentication through keystroke dynamics. *ACM Trans. Inf. Syst. Secur.*, 5(4):367–397.

[Binongo, 2003] Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17.

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

[Chaski, 2005] Chaski, C. E. (2005). Who's at the keyboard: Authorship attribution in digital evidence invesigations. *International Journal of Digital Evidence*, 4(1):n/a. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

[Chaski, 2007] Chaski, C. E. (2007). The keyboard dilemma and forensic authorship attribution. Advances in Digital Forensics III.

[Gray and Juola, 2011] Gray, C. and Juola, P. (2011). Personality identification through on-line text analysis. In *Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, IL.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

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

[James, 2001] James, L. (2001). Fundamentals of biometric authentication technologies. *International Journal of Image and Graphics*, 1(01):93–113.

[Jockers and Witten, 2010] Jockers, M. L. and Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *LLC*, 25(2):215–23.

[Juola, 2006] Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

[Juola et al., 2013] Juola, P., Noecker, Jr., J., Stolerman, A., Ryan, M. V., Brennan, P., and Greenstadt, R. (2013). A dataset for active linguistic authentication. In *Proceedings of the Ninth Annual IFIP WG 11.9 International Conference on Digital Forensics*, Orlando, Florida, USA. National Center for Forensic Science.

[Juola et al., 2011] Juola, P., Ryan, M., and Mehok, M. (2011). Geographically localizing tweets using stylometric analysis. In *Proceedings of the American Association of Corpus Linguistics 2011*, Atlanta, GA.

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

[Koppel and Schler, 2004a] Koppel, M. and Schler, J. (2004a). Ad-hoc authorship attribution competition approach outline. In Juola, P., editor, *Ad-hoc Authorship Attribution Contest*. ACH/ALLC 2004.

[Koppel and Schler, 2004b] Koppel, M. and Schler, J. (2004b). Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 62–, New York, NY, USA. ACM.

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

[Koppel et al., 2005] Koppel, M., Schler, J., and Zigdon, K. (2005). Determining an author's native language by mining a text for errors (short paper). In *Proceedings of KDD*, Chicago,IL.

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

[McDonald et al., 2012] McDonald, A. W. E., Afroz, S., Caliskan, A., Stolerman, A., and Greenstadt, R. (2012). Use fewer instances of the letter "i": Toward writing style anonymization. In *Lecture Notes in Computer Science*, volume 7384, pages 299–318. Springer.

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

[Narayanan et al., 2012] Narayanan, A., Paskov, H., Gong, N., Bethencourt, J., Stefanov, E., Shin, R., and Song, D. (2012). On the feasibility of internet-scale author identification. In *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*. IEEE.

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

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

[Platt, 1998] Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In Schoelkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

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

[Shanmugapriya and Padmavathi, 2009] Shanmugapriya, D. and Padmavathi, G. (2009). A survey of biometric keystroke dynamics: Approaches, security and challenges. *Arxiv preprint arXiv:0910.0817*.

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

[Stamatatos, 2009] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–56.

[Stamatatos, ming] Stamatatos, E. (Forthcoming). Title not available at press time. *Brooklyn Law School Journal of Law and Policy*.

[van Halteren, 2007] van Halteren, H. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing*, 4:n/a.

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

[Wayman et al., 2009] Wayman, J., Orlans, N., Hu, Q., Goodman, F., Ulrich, A., and Valencia, V. (2009). Technology assessment for the state of the art biometrics excellence roadmap. MITRE Technical Report Vol. 2, FBI.

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

[Zheng et al., 2011] Zheng, N., Paloski, A., and Wang, H. (2011). An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security*, CCS '11, pages 139–150, New York, NY, USA. ACM.

[Zheng et al., 2006] Zheng, R., Li, J., Chen, H., , and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393.