Toward a Framework for Continuous Authentication using Stylometry

Marcelo Luiz Brocardo, Issa Traore
Department of Electrical and Computer Engineering
University of Victoria - UVIC
Victoria, British Columbia, Canada
{*marcelo.brocardo, itraore*}*@ece.uvic.ca*

Isaac Woungang
Department of Computer Science
Ryerson University
Toronto, Ontario, Canada
iwoungan@scs.ryerson.ca

***Abstract*—Continuous Authentication (CA) consists of monitoring and checking repeatedly and unobtrusively user behavior during a computing session in order to discriminate between legitimate and impostor behaviors. Stylometry analysis, which consists of checking whether a target document was written or not by a specific individual, could potentially be used for CA. In this work, we adapt existing stylometric features and develop a new authorship verification model applicable for continuous authentication. We use existing lexical, syntactic, and application specific features, and propose new features based on *n-gram* analysis. We start initially with a large features set, and identify a reduced number of user-specific features by computing the information gain. In addition, our approach includes a strategy to circumvent issues regarding unbalanced dataset which is an inherent problem in stylometry analysis. We use Support Vector Machine (SVM) for classification. Experimental evaluation based on the Enron email dataset involving 76 authors yields very promising results consisting of an Equal Error Rate (EER) of 12.42% for message blocks of 500 characters.**

***Keywords*-Continuous authentication, biometrics systems, authorship verification, classification, stylometry, *n-gram* features, short message verification, text mining, writeprint.**

I. Introduction

Continuous authentication (CA) has emerged in the last decade as a mechanism to address the threats posed by masqueraders and session hijackers. CA consists of testing the authenticity of the user repeatedly and unobtrusively throughout the authenticated session as data become available. It has been shown in previous works that behavioral biometrics such as mouse dynamics biometric and keystroke dynamics biometric are good candidates for CA because data can be collected passively using standard computing devices (e.g. mouse and keyboard) throughout a session without any knowledge of the user [1]. We believe that stylometry analysis can achieve the same purpose. Stylometry analysis consists of identifying individual users based on their writing styles, and has so far been studied primarily for the purpose of forensic authorship analysis. Forensic authorship analysis consists of inferring the authorship of a document by extracting and analyzing the writing styles or stylometric features from the document content. Authorship analysis of physical and electronic documents has generated a significant amount of interest over the years and led to a rich body of research literature [2]–[5]. Authorship analysis can be carried out from three different perspectives, including, authorship attribution or identification, authorship verification, and authorship profiling or characterization. Authorship attribution consists of determining the most likely author of a target document among a list of known individuals. Authorship verification consists of checking whether a target document was written or not by a specific individual. Authorship profiling or characterization consists of determining the characteristics (e.g. gender, age, and race) of the author of an anonymous document.

While forensics authorship identification using stylometry has been widely studied, authentication using that modality is still in its infancy. Our goal in this work is to apply authorship analysis technique for continuous user authentication, which is closely related to the problem of forensic authorship verification. Similar to forensic authorship verification, authentication consists of comparing sample writing of an individual against the model or profile associated with the identity claimed by that individual at login time (i.e. 1-to-1 identity matching).

One of the key research challenges faced by continuous authentication approaches, regardless of the data source, is that their accuracy tends to degrade significantly as the amount of data involved in the authentication decreases. However, shorter authentication delay (i.e. smaller data sample) is essential to reduce the window of vulnerability of the system. Therefore, for a continuous authentication scheme to be meaningful it is essential to develop analytical models that will achieve high accuracy while maintaining acceptable authentication delays. We view our proposed approach as a step toward achieving a robust framework for continuous user authentication. Likewise, while some important contributions are made toward achieving our ultimate goal, some open challenges remain that we discuss toward the end of the paper and intend to address in the future.

We evaluate experimentally our approach using the Enron emails dataset and compute the following performance metrics:

- False Acceptance Rate (FAR): measures the likelihood that the system will fail to recognize the genuine person;

- False Rejection Rate (FRR): measures the likelihood that the system may falsely recognize someone as the genuine person;
- Equal Error Rate (ERR): corresponds to the operating point where FAR and FRR have the same value.

Our evaluation yields an EER of 12.42%, which is very encouraging considering the existing works on authorship verification using stylometry.

The rest of the paper is structured as follows. Section II summarizes and discusses related works. Section III introduces our proposed approach. Section IV presents our experimental evaluation by describing the underlying methodology and discussing the obtained results. Section V discusses the strengths and shortcomings of our approach and outlines the ground for future works. Section VI makes some concluding remarks.

II. Related Work

The writing style is an unconscious habit, which varies from one author to another in the way he/she uses words and grammar to express an idea. The patterns of vocabulary and grammar could be a reliable indicator of the authorship. The linguistic characteristics used to identify the author of a text is referred to as stylometry [6], [7]. Although the writing style may change a bit with time [8], each author has a unique stylistic tendency. A large number of studies have used stylometric techniques not only for authorship identification, but also for authorship verification and authorship characterization. Some of the previous studies in authorship identification investigated ways to identify patterns of terrorist communications [9], the author of a particular e-mail for computer forensic purposes [10]–[12], as well as how to collect digital evidence for investigations [13] or to solve a disputed literary, historical [14], or musical authorship [15]–[17] . Work on authorship characterization has targeted primarily gender attribution [18]–[20] and the classification of the author education level [21].

User identity verification is a central aspect of user authentication, however, according to Koppel et al., "using stylometry verification is significantly more difficult than basic attribution and virtually no work has been done on it, outside the framework of plagiarism detection" [4]. Most previous works on authorship verification focus on general text documents. However, authorship verification for online documents can play a critical role in various criminal cases such as blackmailing and terrorist activities, to name a few. To our knowledge, only a handful of studies have been done on authorship verification for online documents. Authorship verification of online documents is difficult because of their relatively short lengths and also because these documents are quite poorly structured or written (as opposed to literary works).

Among the few studies available on authorship verification, are works by Koppel et al. [4], Iqbal et al. [11], Canales et al. [3], and Brocardo et al. [22].

Koppel et al. proposed an authorship verification method named "unmasking" where an attempt is made to quantify the dissimilarity between the sample document produced by the suspect and that of other users (i.e. imposters) [4]. The experimental evaluation, however, shows that the proposed approach can provide trustable results only for documents of at least 5000 words long, which is not realistic in the case of online verification.

Iqbal et al. studied email authorship verification by extracting 292 different features and analyzing these features using different classification and regression algorithms [11]. Experimental evaluation of the proposed approach using the Enron e-mail corpus yielded EER ranging from 17.1% to 22.4%.

Canales et al. extracted keystroke dynamics and stylistic features from sample exam documents for the purpose of authenticating online test takers [3]. The extracted features consisting of timing features for keystroke and 82 stylistic features were analyzed using a K-Nearest neighbor (KNN) classifier. Experimental evaluation involving 40 students with sample document size between 1710 to 70,300 characters yielded (FRR=20.25%, FAR=4.18%) and (FRR= 93.46%, FRR=4.84%) when using separately keystroke and stylometry, respectively. The combination of both types of features yielded EER= 30%.

Brocardo et al. investigated the possibility of using stylometry for authorship verification for short online messages [22]. The technique was based on a combination of supervised learning and *n-gram* analysis. The evaluation used real-life dataset from Enron, where the e-mails were combined to produce a single long message per individual, and then divided into smaller blocks used for authorship verification. The experimental evaluation yielded an EER 14.35% for 87 users for message blocks of 500 characters. The current work built on the previous model by extending significantly the feature set and using Support Vector Machine (SVM) for classification, yielding improved verification accuracy.

III. Proposed Approach

In this section, we present our approach by discussing feature selection and describing in detail our classification model. In a general overview of our approach, we decompose an online document into consecutive blocks of short texts over which (continuous) authentication decisions happen. We extract lexical, syntactic, and application specific features. In addition, we compute new features based on *n-gram* analysis [22], and use information gain technique for feature selection. In order to balance the dataset, we define a weight for the instances based on the proportion of positive and negative training samples. Finally, we use SVM for classification.

A. Initial Features

Stylometry consists of the quantification of the writing style characteristics or style markers of a document in order to create a writeprint that represents the style of its author. As shown in Figure 1, in this study we selected our set of features by combining lexical character frequency (50 features), lexical character *n-gram* (16 features), lexical word (25 features), syntactic (251 features) and application specific features (7 features).

Lexical features are related to the words or vocabulary of a language. Lexical analysis consists of breaking a text into a single atomic unit of language called token. A token can be a word or a character [23]. While earlier studies used a set of 100 frequent words to determine the author of a document [24], recent studies have used more than 1000 frequently used words to represent the style of an author [25]. However, lexical features encompass not only the frequency of characters or words found in a text but also vocabulary richness, sentence/line length, word length distribution, n-grams and lexical errors [3], [26].

Some lexical features measure the frequency of characters, which include letters (uppercase and lowercase), digits, and special characters (e.g. '@', '#', '$', '%', '(', ')', '{', '}', etc.). Other lexical features are obtained by extracting *n*-grams from a text. *N*-grams are tokens formed by a contiguous sequence of *n* items. The most frequent *n*-grams constitute the most important feature for stylistic purposes. Importantly, *n*-grams are noise tolerant since their representation is not affected dramatically by factors such as misspelling [26].

Vocabulary richness measures the diversity of vocabulary in a text by quantifying the total number of unique vocabulary, the number of *hapax legomenon* (i.e., a word which occurs only once in a text) and the number of *hapax dis legomenon* (e.g., dis legomenon or tris legomenon, referring to double or triple occurrences). This metric is computed by dividing the total number of unique vocabulary (hapax legomenon or dis legomenon) by the total number of tokens (each token is a word).

Syntactic features can be divided into average of punctuation and part-of-speech (POS). Syntactic pattern is an unconscious characteristic and it is considered to be more reliable than lexical information [27]. Punctuation is an important rule to define boundaries and identify meaning (quotation, exclamation, etc.) by splitting a paragraph into sentences and each sentence into various tokens. However, it is not sufficient to analyze only the punctuation of a document, as certain words such as 'Ph.D.' or 'uvic.ca' include punctuation characters too. Therefore, it is necessary to format the text before analyzing it.

Application specific features can easily be extracted from documents by analyzing structural and content-specific characteristics [28]–[31]. Structural characteristics are related to the organization and format of a text and are usually more flexible in online documents such as e-mail. These features can be categorized at the message-level, paragraph-level or according to the technical structure of the document [2].

Features			Characteristics	Total
Lexical	Character	Frequency	Number of characters (C)	1
			Number of lower character/C	1
			Number of upper characters/C	1
			Number of white-space characters/C	1
			Total number of vowels (V)/C	1
			Vowels (a, e, i, o, u) / V	5
			Alphabets (A-Z) / C	26
			Number of special characters (S) /C	1
			Special Characters (%,&,etc.) / S	13
		n-grams	Character 5 and 6-grams (Ru and Decision)	16
				66
	Word	Total number of words (N)		1
		Average sentence length in terms of words /N		10
		Frequency	Words longer than 6 characters/N	1
			Total number of short words (1-3 characters)/N	1
			Average word length	1
			Average syllable per word	1
			Ratio of characters in words to N	1
			Replaced words / N	6
		Vocabulary richness	Hapax legomena	1
			Hapax dislegomena	1
			Vocabulary richness (total different words/N)	1
				25
Syntatic		Total number of punctuation (P)		1
		Frequency	single quotes, commas, periods, colons, semi-colons, question marks, exclamation marks divided by P	8
		Functional words	Total number of auxiliary verbs, conjunction, prepositions, pronouns, determiners each one divide by N	6
			Ratio of functional word divide by the respective total word group	236
				251
Application-specific	Structural	Paragraph-Level	Total number of sentences	1
			Total number of paragraphs	1
			Average of characters, words and sentences in a paragraph	3
			Average of sentences beginning with upper case	1
			Average of sentences beginning with lower case	1
				7

Figure 1. List of stylometry features used in our work.

B. N-gram Model

Our classification model consists of a collection of profiles generated separately for individual users. Our proposed system operates in two modes: enrollment and verification. The enrollment process uses sample data to compute the behavioral profile of the user. Each sample is an instance composed by a label and a set of features.

The feature extraction is performed in two steps. During the first step, the frequency and average of lexical, syntactic and application specific features are computed. In the second step, we calculate the character n-grams.

The n-gram calculation is adapted from our previous work [22], where we reduced the number of n-grams features to one feature. Our method differs from our previous work as we calculate not only all unique n-grams, but also all n-grams with frequency equal or higher than some number f.

Given a user U, we divide her training data into two subsets, denoted $T(f)_1^U$ and T_2^U. Let $N(T(f)_1^U)$ denote the set of all unique *n-grams* occurring in $T(f)_1^U$ with frequency f.

Let m denote a binary variable (i.e., $m \in \{1, 0\}$) that represents the mode of calculation of the n-grams.

Given a block b, let $N_m(b)$ denote the following:

$$\begin{cases} \text{- the set of all unique } \textit{n-grams} \text{ occurring in } b, \text{ if } m = 0 \\ \text{- the set of all } \textit{n-grams} \text{ occurring in } b, \text{ otherwise.} \end{cases} \tag{1}$$

We define the similarity $r_U(b)$ between a sample data block b and the profile of user U, where the similarity varies between 0 and 1, as the percentage of unique *n-grams* shared by block b and training set $T(f)_1^U$, giving[1]:

$$r_U(b) = \frac{|N_m(b) \cap N(T(f)_1^U)|}{|N_m(b)|} \tag{2}$$

We also define a binary similarity metric, denoted $d_U(b)$ and referred to as *decision*, which captures the closeness of a block b to the profile of user U, as follows:

$$\begin{cases} d_U(b) = 1 \text{ if } |r_U(b)| \geq \epsilon_U \\ d_U(b) = 0, \text{otherwise} \end{cases} \tag{3}$$

where ϵ_U is a user-specific threshold derived from the training data.

We derive the value of ϵ_U for user U using a supervised learning technique outlined by $Algorithm$ 1. Given a user U, we divide the training subset T_2^U into p blocks of characters of equal size: $b(m)_1^U, ..., b(m)_p^U$. Our model approximates the actual (but unknown) distribution of the ratios $(r_U(b_1^U), ..., r_U(b_p^U))$ (extracted from T_2^U) by computing the sample mean denoted μ_U and the sample variance σ_U^2 during the training. In the algorithm, the threshold is initialized (i.e. $\epsilon_U = \mu_U - (\sigma_U/2)$), and then varied incrementally by minimizing the difference between FRR and FAR values for the user, the goal being to obtain an operating point that is as close as possible to the EER.

For each test block b, we derive 2 new features corresponding to $r_U(b)$ and $d_U(b)$. In this study, we consider only 5-grams and 6-grams, and cover two different values for the frequency f (i.e. $f = 1$ and $f = 2$) and for the mode of calculation of the n-grams (i.e. $m = 0$ and $m = 1$). Therefore, the number of new features created from the above n-gram model is 2 (for f) $\times$2 (for m) $\times$2 (for *n-gram* types) $\times$2 (for r_U and d_U) = 16.

C. Features Selection

Continuous authentication occurs by performing authentication decisions repetitively over consecutive blocks of data captured during a session. For each block of text, we extract all features represented as a vector of features values. Our next step is to normalize the feature values to range between 0 and 1. Since, most of the candidate features values fall in the above range, normalization is applied only for the features that have absolute values, which are the "total number of characters", the "total number of words", the "average number of words per sentence", and the "average word size". We normalize these features using *maximum normalization* scheme, in which case a given feature value will be replaced by its ratio with the maximum value for the same feature over the training set.

[1] $|X|$ denote the cardinality of set X.

```
/* U a user for whom the threshold is being calculated */
/* I_1, ..., I_m: a set of other users (I_k ≠ U)        */
/* ε_U: threshold computed for user U                   */
Input: Training data for U, I_1, ..., I_m
Output: ε_U
1  begin
2      up ← false;
3      down ← false;
4      δ ← 1;
5      ε_U ← μ_U − (σ_U/2);
6      while δ > 0.0001 do
           /* Calculating FAR and FRR for user U       */
7          FRR_U, FAR_U = calculate(U, I_1, ..., I_m, ε_U, γ);
               /* Minimizing the difference between FAR
               and FRR                                  */
8          if (FRR_U − FAR_U) > 0 then
9              down ← true;
10             ε_U ← ε_U − δ;
11         end
12         if (FAR_U − FRR_U) > 0 then
13             up ← true;
14             ε_U ← ε_U + δ;
15         end
16         if (up & down) then
17             up ← false;
18             down ← false;
19             δ ← δ/10;
20         end
21     end
22     return ε_U;
23 end
```

Algorithm 1: Threshold calculation for a given user.

Analyzing a large number of features does not necessarily provide the best results, as some features provide very little or no predictive information. Being able to keep only the most discriminating features individually per user allows reducing the data size by removing irrelevant attributes and improve the processing time for training and classification. This can be achieved by applying feature selection measures, which allows finding a minimum set of features that represent the original distribution obtained using all the features.

Although features selection by an expert is commonly used, it is complex and sometime inefficient because it is easy to select irrelevant attributes while omitting important attributes. Other method that could be applied is an exhaustive search. Such brute-force feature selection method could evaluate all possible feature combinations, but it is time consuming and impractical. A probabilistic approach is an alternative for speeding up the processing time and selecting optimal subset of features.

In order to evaluate the worth of an attribute with the highest discrimination, we apply in this work a ranking

strategy based on the information gain. Prior to computing the information gain, it is necessary to discretize the numeric feature values into binary values (0 and 1). The discretization process consists of finding a cut-point or split-point that divides the range into two intervals, one interval being less or equal than the cut-point while the other is greater [32]. We use the entropy-based discretization method proposed by Fayyad and Irani [33], which is a supervised discretization method and has been known to achieve some of the best performances in the literature.

After discretizing the features, we calculate the information gain for each of them by computing their entropy with respect to the training sample.

Let T be a set of training samples $(y_j, x_1, ..., x_n)$, where y_j is the corresponding class label ($y_j = 1$ for genuine sample; $y_j = -1$ for impostor sample), and $x = (x_1, ..., x_n)$ is a feature vector. The information gain $IG(T, x_i)$ for a given feature x_i measures the expected reduction in entropy computed by the following equation:

$$IG(T, x_i) = H(T) - H(T|x_i) \tag{4}$$

$H(T)$ denotes the information entropy, which is a measure of the uncertainty in a random variable as given by:

$$H(T) = -\sum_{j=1}^{n} p(y_j, x) log_2 p(y_j, x) \tag{5}$$

where $p(y_j, x)$ denote the probability mass function for the fraction in x having class y_j.

$H(T|x_i)$ represents the information entropy given a feature x_i and it is calculated by the following formula:

$$H(T|x_i) = \sum_{v \in values(x_i)} \frac{|T_{x_i(v)}|}{|T|} \times H(T_{x_i(v)}) \tag{6}$$

For the purpose of feature selection, we retain only features with non-zero information gain. As a result using sample data, the number of features is reduced from 349 to 50 on average.

D. SVM Classifier

We use SVM for classification. SVM is a supervised learning method used for classification and regression analysis that performs a non-probabilistic binary linear classification [34]. SVM is based on the idea of mapping the original finite-dimensional space X into a much higher-dimensional space F. SVM constructs a hyperplane separating points of the two classes and finds a dividing straight line (optimal hyperplane). The decision boundary is the maximum-margin hyperplane or the largest minimum distance to the training examples, as illustrated in Figure 2. The hyperplane can be modeled using only the samples on the margin, called the support vectors, and training an SVM consists of identifying the support vectors within the training samples.

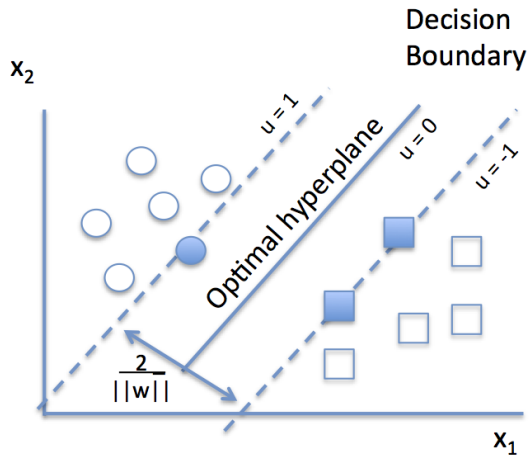


Figure 2. Decision boundary separating two classes; samples on the margin are called the support vectors.

Assume that $s_i \in X$ are the support vectors, each associated with a class label $y_j \in \{+1, -1\}$ (for positive and negative examples, respectively). Given an unlabeled sample $x \in X$, classification consists of predicting the corresponding label y_j. This is performed using a decision function as follows.

$$f(x) = \sum_i y_i \alpha_i K(x, s_i) + b \tag{7}$$

Where α_i are Lagrangian multipliers, and K is a kernel function that measures the similarity or distance between the unlabeled sample x and the support vector s_i. The kernel function $K(x, s_i)$ maps the sample space X into a high-dimensional feature space F. Examples of kernel functions include linear, polynomial, Gaussian, and hyperbolic tangent [34].

IV. Experimental Evaluation

A. Dataset and Data Preprocessing

In order to validate our system, we performed experiments on a real-life dataset from Enron e-mail corpus[2]. Enron was an energy company (located in Houston, Texas) that was bankrupt in 2001 due to white collar fraud. The e-mails of Enron's employees were made public by the Federal Energy Regulatory Commission during the fraud investigation. The e-mail dataset contains more than 200 thousands messages from about 150 users. The average number of words per e-mail is 200. The e-mails are plain texts and cover various topics ranging from business communications to technical reports and personal chats.

In order to obtain the same structural data and improve classification accuracy, we performed several preprocessing steps to the data as follows:

[2]available at http://www.cs.cmu.edu/~enron/

- E-mails from the folders "sent" and "sent items" within each user's folder were selected, with all duplicate e-mails removed;
- JavaMail API was used to parse each e-mail and extract the body of the message;
- Remove messages where the average of digit per total of character was higher than 25%;
- Since different texts must be logically equivalent, (i.e., must have the same canonical form), the following filters were applied:
  - Replace phone number for a single phone word;
  - Replace currency for $XX;
  - Replace percentage for XX%;
  - Strip e-mail replay;
  - Replace e-mail address for a single e-mail word;
  - Replace htpd address for a single http word;
  - Replace information between tags ("< information >") for a single TAG word;
  - Replace date for a single date word;
  - Replace time for a single time word;
  - Delete content when have the following information: Date:, Time:, Location:;
  - Replace numbers for the single "numb" word;
  - Replace information among quotes ("information") for a single quote word;
  - Normalize the document to printable ASCII;
  - Convert the document to lowercase characters;
  - Strip white space.
- All messages, per author, were grouped creating a long text or stream of characters that was divided into blocks.

The dataset involves an imbalance class distribution where we have more negative samples than positive ones. Balanced classification can be achieved by changing the class distribution through under-sampling the majority class or over-sampling the minority class. Our approach to deal with this situation is to assign a weight to the negative class corresponding to the ratio between the total number of positive samples and the total number of negative samples.

B. Evaluation Method and Results

We implemented our framework in Java and utilized the WEKA (Waikato Environment for Knowledge Analysis)[3] machine learning framework and libraries for our classification algorithms [35]. We used a SVM learner called Sequential Minimal Optimization (SMO) in WEKA [36]. We tried linear and Gaussian kernels on sample data, but we did not get good results when compared to polynomial. Using sample data with a polynomial kernel, we explored different configurations, and obtained the best results by setting the degree of the polynomial kernel to one.

[3]available at http://weka.wikispaces.com

The testing was conducted with a block size of 500[4] characters[5] on average and 50 blocks or instances per user, since this configuration showed the best results in our previous work.

To evaluate the accuracy of the proposed approach, we performed a 10-fold cross-validation test. We randomly sorted the dataset, and allocated in each (validation) round 90% of the dataset for training and the remaining 10% for testing. For each user U, we computed a corresponding profile by using their training data and training data from other users considered as impostors.

After the preprocessing phase, the dataset was reduced from 150 authors to 76 authors to ensure that only users with 50 instances and 500 characters per instance were involved in our analysis.

Each individual user profile was built using 45 positive instances and 3375 $(= 75 \times 45)$ negative instances. The remaining instances consisting of 5 positive instances and 375 $(= 75 \times 5)$ negative instances were used for testing. The test was repeated 76 times by considering each time one of the users in our experiment as a legal user while the remaining users were considered as impostors.

We computed the FRR for user U by comparing each of the instances from her test data against her profile. A false rejection (FR) is counted when the system rejects one of these instances. The FAR is computed by comparing each of the test instances from the other users (i.e. the impostors) against the profile of user U. A false acceptance (FA) occurs when the system categorizes any of these instances as belonging to user U. By repeating the above process for each of the users, we compute the overall FAR and FRR by averaging the individual measures.

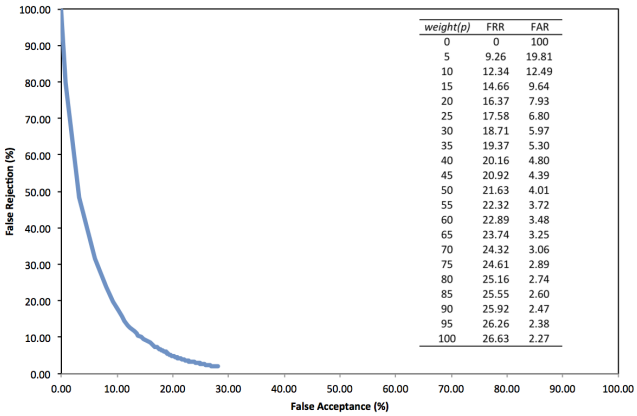
Figure 3 shows the receiver operating characteristic (ROC) curve for the experiment. The curve shows the relation between the FAR and FRR when varying the weight assigned to the negative samples denoted $weight(P)$ from 0 to 100. The optimal performance achieved by our system was obtained when setting the $weight(P)$ limit to 10, with a FAR of 12.49% and a FRR of 12.34%. The equal error rate (EER) was calculated as 12.42%.

V. Discussions

Despite significant progress in identifying an author among a few candidates (e.g. 3 to 10), it is still challenging to identify an author when we have a large number of candidates or when the text is short like an e-mail or an online message (e.g. in twitter messages). Most of the previous work on stylometry have included a combination of lexical, semantic, syntactic, and application-specific features,

[4]The block size starts after punctuation and ends with an entire word, which means if 500 characters corresponds exactly to the middle of a word, then the block will have 500 characters plus the rest of the characters involved in the complete word.

[5]include A..Z, a..z, 0..9, punctuation, white space, special characters



weight(p)	FRR	FAR
0	0	100
5	9.26	19.81
10	12.34	12.49
15	14.66	9.64
20	16.37	7.93
25	17.58	6.80
30	18.71	5.97
35	19.37	5.30
40	20.16	4.80
45	20.92	4.39
50	21.63	4.01
55	22.32	3.72
60	22.89	3.48
65	23.74	3.25
70	24.32	3.06
75	24.61	2.89
80	25.16	2.74
85	25.55	2.60
90	25.92	2.47
95	26.26	2.38
100	26.63	2.27

Figure 3. Receiver Operating Characteristic curve for the experiment and sample performance values for different weights.

but there is no consensus among researchers regarding what is the best set of features. Stylometry is considered a behavioral biometric and although many studies have employed stylometric techniques for authorship attribution and characterization, fewer studies have focused on verification, and to our knowledge there is no study on using stylometry for continuous authentication.

The work presented in this paper is a step toward implementing continuous authentication using stylometry. Three key challenges must be addressed in order to fully reach that objective:

1) High accuracy;
2) Low authentication delay;
3) Ability to withstand forgery.

The results achieved in this work are very encouraging and better than those obtained so far by similar work in the literature in terms of verification accuracy. Table I summarizes the performances, block sizes, and population sizes of previous stylometry studies.

Some studies provide the block size in number of words and other in number of characters. According to Sanderson and Guenter the average word length is about 5.6 characters. The accuracy tends to degrade when the block size becomes smaller [5], [22]. Smaller block size means shorter authentication delay, which is important for CA.

As illustrated by Table I, an important limitation of many previous stylometry studies is that their performances were computed using only classification accuracy which covers only one side of the story. Classification accuracy actually corresponds to the true match rate (TMR) and allows deriving only one type of error, namely, $FAR = 1 - Classification Accuracy$. Nothing is said about FRR in these studies, which makes it difficult to judge their real strength in terms of accuracy. As shown by Table I, only few studies have provided both types of errors, among which our work can be considered as the strongest in terms of sample population size, block size, and accuracy.

When comparing this research with our previous work [22], the sample population size decreased from 87 to 76 because we are applying new filters in order to have the same canonical form. By expanding our feature set beyond *n-grams*, we obtain an improvement of the verification accuracy. Our proposed approach achieves EER of 12.42% which is better compared to the accuracy obtained using similar techniques in the literature and in our previous work [22].

Although the accuracy of the authentication mechanism is an important performance metric in CA, the authentication delay plays an important role as well, since it is a measure of the window of vulnerability of the system. However, attempting to reduce at the same time the authentication delay and the verification error rates is a difficult task in the sense that these attributes are loosely related to each other. A quicker authentication decision may lead to increased identity verification error rates, and vice-versa.

In continuous authentication, the authentication delay is related to the block size. However, existing stylometry analysis approaches use overwhelmingly large documents sizes for identity verification, varying from several hundreds to several thousands words. We investigated in this work a block size of 500 characters, which represents significantly shorter messages compared to the messages used so far in the literature for identity verification.

Sanderson and Guenter achieved similar results using block size of 500 characters, although with a relatively smaller dataset (i.e. 50 users) [5]. However, it is important to mention that their dataset consisted of newspapers articles, which are known to be well structured compared to e-mail messages.

Another important issue that we need to address to achieve a robust CA system is to assess and strengthen the approach against forgeries. Stylometry analysis can be the target of automated attacks, also referred to as generative attacks, where high quality forgeries can be generated automatically using a small set of genuine samples [37]. An adversary having access to writing samples of a user may be able to effectively reproduce many of the existing stylometric features. It is essential to integrate specific mechanisms in the authentication system that would mitigate forgery attacks. Furthermore, our evaluation methodology and dataset must be upgraded by generating and incorporating forgery samples. We intend to tackle these issues in our future work.

VI. Conclusion

We have presented in this paper a new framework for continuous authentication using stylometry analysis. Our feature set consists in the first place of existing lexical, syntactic, and application specific features. In addition, we derived 16 new features through *n-gram* analysis. In order

Table I
COMPARATIVE PERFORMANCES, BLOCK SIZES AND, POPULATION SIZES FOR STYLOMETRY STUDIES.

Category	Ref	Sample Population Size	Block Size	Number of Features	Technique	Accuracy* (%)	EER (%)
Attribution	[2]	100 **	277 words	L(25065), Sy(2766), A(128)	Karhunen-Loeve (K-L) - (Principal Component Analysis)	83.10	- -
	[13]	10	200 words	L(1), Sy(10)	Discriminant Function Analysis (DFA)	95.70	- -
	[38]	2 - 4	60,000 words	Se (many)	Synonym-based features through statistical classification	93.8 - 97.8	- -
	[39]	3	20 sentences	L(28820), Sy(4117), Se(1896)	SVM	87.63	- -
	[30]	3 **	200 words	400 features including lexical, syntactic, and structural features	SVM*** and C4.5 (Decision Tree)	69 -83	- -
	[25]	87	287 words	L(not specified the quantity, but include top-k word frequencies, 4-grams characters), Sy(8)	Logic Fuzzy	50 - 60	- -
	[40]	3 - 10 **	200 words	L(82), Sy(311), A(26)	Expectation Maximization (EM), k-means, and bisecting k-means***	80 - 90	- -
	[12]	4 - 20 **	300 words	L(105), Sy(159), Se(10), A(28)	Frequent pattern	69.75 - 88.37	- -
	[10]	6 - 10 **	200 words	Not specify the quantity, but include L, Sy, A	Frequent pattern	77 - 80.5	- -
	[41]	1000	500 words	L(250K - space- free character 4-grams.)	Naïve	42.2 - 93.2	- -
	[6]	20	169 words	L(87), Sy(158), Se(11), A(14)	SVM	99.01	- -
	[42]	20	600 words	Sy(171)	Prediction by Partial Matching (PPM)	84.30	- -
	[5]	50	500 characters	????	Hidden Markov Model (HMM)	- -	8.08 - 30.88
Characterization	[9]	5	76 words	L(79), Sy (262), Se(15), A(62)	C4.5 decision tree and SVM***	90.1 - 97	- -
	[20]	108 **	50 - 200 words	L(130), Sy(402) , A(13)	SVM*** , Bayesian logistic regression, and AdaBoost decision tree	73 - 82.23	- -
	[19]	114 **	50 - 200 words	L(130), Sy(402), A(13)	Decision Tree, SVM***	80.08 - 82.20	- -
	[43]	325	50 - 200 words	L(69), Sy(122), A(30)	SVM	70.20	- -
	[12]	4 - 20 **	300 words	L(105), Sy(159), A(15), Se(23)	Frequent Pattern	39.13% - 60.44%	- -
	[44]	100	300 words	L(89), Sy(119), A(3)	k-NN, Naïve Bayesian ***, Patient rule induction method, SVM	39.0 - 99.70	- -
	[18]	10 - 40	450 words	L (several - unigram, bigram, and trigram)	Probabilistic Context-Free Grammar (PCFG)	68.3 - 91.5	- -
Verification	[3]	40	1710 - 70300 characters	L(62), Sy(20)	k-NN	- -	30
	[31]	25 - 40 **	30 - 50 words	L(40), Sy(76), Se(25), A(9)	SVM	83.90 - 88.31	
	[45]	8	628 - 1342 words	L(100K), Sy(900K)	Weighted Probability Distribution Voting (WPDV)	- -	3
	[4]	10	500 words	L(250)	SVM	95.70	- -
	[46]	29	2400 words	L(40)	Linear Discriminant Analysis (LDA)	- -	22
	[22]	87 **	500 character	L(n-gram)	Supervised	- -	14.35
	Current paper	76 **	500 character	L(91), Sy(251), A(7)	SVM	- -	12.42

* The accuracy is measured by the percentage of correctly matched authors in the testing set.
** Used Enron dataset for testing.
*** Best result achieved.
(L) = Lexical, (Sy) = Syntactic, (Se) = Semantic, (A) = Application

to select the best set of features to represent individual user profile, we compute and analyze the information gain. This allows reducing our feature set from 349 to 50 on average. We used SVM classifer to build and train users profiles. Experimental evaluation of our approach using the Enron dataset yields very encouraging results compared to the existing literature, consisting of EER 12.42% for 76 users for a relatively small block size of 500 characters. While the obtained results are promising, it is clear that more work must be done for the proposed scheme to be fully usable for continuous authentication. We discussed above some of the limitations of our existing approach and plan to address them in our future work.

VII. ACKNOWLEDGMENTS

This research has been enabled by the use of computing resources provided by WestGrid and Compute/Calcul Canada. The research is funded by a Vanier schorlarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) and by a research grant from the National Council for Scientific and Technological Development (CNPq) of Brazil.

REFERENCES

[1] I. Traore and A. A. E. Ahmed, Eds., *Continuous Authentication Using Biometrics: Continuous Authentication Using Biometrics: Data, Models, and Metrics*. IGI Global, 2012.

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

[3] O. Canales, V. Monaco, T. Murphy, E. Zych, J. Stewart, C. T. A. Castro, O. Sotoye, L. Torres, and G. Truley, "A stylometry system for authenticating students taking online tests," P. of Student-Faculty Research Day, Ed., CSIS. Pace University, May 6 2011.

[4] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in *Proceedings of the 21st international conference on Machine learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 62–.

[5] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, markov chains and author unmasking: an investigation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 482–491.

[6] J. Li, R. Zheng, and H. Chen, "From fingerprint to writeprint," *Commun. ACM*, vol. 49, pp. 76–82, April 2006.

[7] J. L. Hilton, *On verifying wordprint studies: Book of Mormon authorship*, ser. Reprint (Foundation for Ancient Research and Mormon Studies). F.A.R.M.S., 1991. [Online]. Available: http://books.google.ca/books?id=gVQKYgEACAAJ

[8] F. Can and J. M. Patton, *Change of Writing Style With Time*. Kluwer Academic Publishers, 2004, vol. 38.

[9] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems*, vol. 20, pp. 67–75, September 2005.

[10] F. Iqbal, R. Hadjidj, B. C. Fung, and M. Debbabi, "A novel approach of mining write-prints for authorship attribution in e-mail forensics," *Digital Investigation*, vol. 5, Supplement, no. 0, pp. S42 – S51, 2008, the Proceedings of the Eighth Annual DFRWS Conference.

[11] F. Iqbal, L. A. Khan, B. C. M. Fung, and M. Debbabi, "E-mail authorship verification for forensic investigation," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, ser. SAC '10. New York, NY, USA: ACM, 2010, pp. 1591–1598.

[12] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi, "A unified data mining solution for authorship analysis in anonymous textual communications," *Information Sciences*, 2011.

[13] C. E. Chaski, "Who's at the keyboard: Authorship attribution in digital evidence investigations," *International Journal of Digital Evidence*, vol. 4, no. 1, Spring 2005.

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

[15] J. Burrows, "Delta: a measure of stylistic difference and a guide to likely authorship," *Literary and Linguistic Computing*, vol. 17, no. 3, pp. 267–287, 2002.

[16] E. Backer and P. van Kranenburg, "On musical stylometry pattern recognition approach," *Pattern Recognition Letters*, vol. 26, no. 3, pp. 299 – 309, 2005.

[17] Y. Zhao and J. Zobel, "Searching with style: authorship attribution in classic literature," in *Proceedings of the thirtieth Australasian conference on Computer science - Volume 62*, ser. ACSC '07. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2007, pp. 59–68.

[18] K. G. Ruchita Sarawgi and Y. Choi, "Gender attribution: tracing stylometric evidence beyond topic and genre," in *Proceedings of the 15th Conference on Computational Natural Language Learning*, ser. CoNLL '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 78–86.

[19] N. Cheng, X. Chen, R. Chandramouli, and K. Subbalakshmi, "Gender identification from e-mails," in *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, 30 2009-april 2 2009, pp. 154 –158.

[20] N. Cheng, R. Chandramouli, and K. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, vol. 8, no. 1, pp. 78 – 88, 2011.

[21] P. Juola and R. H. Baayen, "A controlled-corpus experiment in authorship identification by cross-entropy," *Literary and Linguistic Computing*, vol. 20, no. Suppl, pp. 59–67, 2005.

[22] I. S. S. Brocardo, Marcelo Luiz; Traore and I. Woungang, "Authorship verification for short messages using stylometry," in *In Proceedings of the International Conference on Computer, Information and Telecommunication Systems (CITS)*. Piraeus-Athens, Greece, 2013.

[23] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2011.

[24] J. F. Burrows, "Word patterns and story shapes: The statistical analysis of narrative style," *Literary and Linguistic Computing*, vol. 2, no. 1, pp. 61–70, 1987.

[25] N. Homem and J. Carvalho, "Authorship identification and author fuzzy fingerprints," in *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American*, march 2011, pp. 1 –6.

[26] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, pp. 538–556, March 2009.

[27] H. Baayen, H. van Halteren, and F. Tweedie, "Outside the cave of shadows: using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing*, vol. 11, no. 3, pp. 121–132, 1996.

[28] O. de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD RECORD*, vol. 30, pp. 55–64, 2001.

[29] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, pp. 378–393, February 2006.

[30] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," *Digital Investigation*, vol. 5, no. 3-4, pp. 124 – 137, 2009.

[31] X. Chen, P. Hao, R. Chandramouli, and K. P. Subbalakshmi, "Authorship similarity detection from email messages," in *Proceedings of the 7th international conference on Machine learning and data mining in pattern recognition*, ser. MLDM'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 375–386.

[32] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," 2006.

[33] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuousvalued attributes for classification learning," in *Thirteenth International Joint Conference on Articial Intelligence*, vol. 2. Morgan Kaufmann Publishers, 1993, pp. 1022–1027.

[34] V. Vapnik, *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982.

[35] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations," 1999.

[36] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998. [Online]. Available: http://research.microsoft.com/\~jplatt/smo.html

[37] L. Ballard, "Biometric authentication revisited: Understanding the impact of wolves in sheep's clothing," in *In Proceedings of the 15 th Annual Usenix Security Symposium*, 2006, pp. 29–41.

[38] J. H. Clark and C. J. Hannon, "A classifier system for author recognition using synonym-based features," in *Proceedings of the 6th Mexican international conference on Advances in artificial intelligence*, ser. MICAI'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 839–849.

[39] M. Gamon, "Linguistic correlates of style: authorship classification with deep linguistic analysis features," in *Proceedings of the 20th international conference on Computational Linguistics*, ser. COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[40] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *Digital Investigation*, vol. 7, no. 1-2, pp. 56 – 64, 2010.

[41] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Lang. Resour. Eval.*, vol. 45, pp. 83–94, March 2010.

[42] D. Pavelec, L. Oliveira, E. Justino, F. Neto, and L. Batista, "Author identification using compression models," in *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, july 2009, pp. 936 –940.

[43] M. Corney, O. de Vel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *Proceedings of the 18th Annual Computer Security Applications Conference*, 2002, pp. 282 – 289.

[44] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining: Predicting user and message attributes in computer-mediated communication," *Information Processing Management*, vol. 44, no. 4, pp. 1448 – 1466, 2008.

[45] H. V. Halteren, "Author verification by linguistic profiling: An exploration of the parameter space," *ACM Trans. Speech Lang. Process.*, vol. 4, pp. 1:1–1:17, February 2007.

[46] I. Krsul and E. H. Spafford, "Authorship analysis: identifying the author of a program," *Computers and Security*, vol. 16, no. 3, pp. 233 – 257, 1997.