Journal of Computer and System Sciences 81 (2015) 1429–1440

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Authorship verification of e-mail and tweet messages applied for continuous authentication ☆

Marcelo Luiz Brocardo [a,*], Issa Traore [a,**], Isaac Woungang [b]

[a] Department of Electrical and Computer Engineering, University of Victoria, Victoria, British Columbia, V8W 3P6, Canada
[b] Department of Computer Science, Ryerson University, Toronto, Ontario, M5B 2K3, Canada

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form 14 December 2014
Accepted 14 December 2014
Available online 29 December 2014

Keywords:
Continuous authentication
Stylometry
Short message verification
n-Gram features
Unbalanced dataset
SVM classifier

ABSTRACT

Authorship verification using stylometry consists of identifying a user based on his writing style. In this paper, authorship verification is applied for continuous authentication using unstructured online text-based entry. An online document is decomposed into consecutive blocks of short texts over which (continuous) authentication decisions happen, discriminating between legitimate and impostor behaviors. We investigate blocks of texts with 140, 280 and 500 characters. The feature set includes traditional features such as lexical, syntactic, application specific features, and new features extracted from *n*-gram analysis. Furthermore, the proposed approach includes a strategy to circumvent issues related to unbalanced dataset, and uses Information Gain and Mutual Information as a feature selection strategy and Support Vector Machine (SVM) for classification. Experimental evaluation of the proposed approach based on the Enron email and Twitter corpuses yields very promising results consisting of an Equal Error Rate (EER) varying from 9.98% to 21.45%, for different block sizes.

1. Introduction

Continuous authentication (CA) consists of re-authenticating a user repeatedly and unobtrusively throughout an authenticated session as data becomes available. CA is considered as a remedy against session hijacking, where an intruder seizes control of a legitimate user session [1]. Ideally, CA should be conducted unobtrusively, by enabling transparent user identity data collection and verification. Stylometric analysis, which consists of the identification of a user based on his writing style [2,3], could potentially be used for CA.

Stylometry has so far been studied primarily for the purpose of forensic authorship analysis, generating a significant amount of interest over the years and led to a rich body of research literature [4–7]. Three different kinds of authorship analysis using stylometry have been studied in the literature, including authorship attribution, authorship characterization, and authorship verification.

The focus of this paper is on authorship verification, which is the most relevant to continuous authentication. Authorship verification consists of checking whether a target document was written or not by a specific (i.e. known) individual (with

☆ This paper is the journal version of the conference paper (Marcelo Luiz Brocardo, Issa Traore, Isaac Woungang [23]).

* Corresponding author.
** Principal corresponding author.
E-mail addresses: marcelo.brocardo@ece.uvic.ca (M.L. Brocardo), itraore@ece.uvic.ca (I. Traore), iwoungan@scs.ryerson.ca (I. Woungang).

http://dx.doi.org/10.1016/j.jcss.2014.12.019
0022-0000/

some claimed identity). In this setting CA is performed by 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).

CA involves several challenges including the need for low authentication delay, high accuracy, and the ability to withstand forgery. This paper focuses on the two first challenges. Low authentication delay is simulated by analyzing short texts. Attempting to reduce at the same time the text size and the verification error rates is a difficult task in the sense that these attributes are loosely related to each other. A smaller verification block may lead to increased verification error rates and vice-versa.

In this work, a set of new features are derived from *n*-gram analysis, and two classification schemes are studied: a Support Vector Machine (SVM) classifier and a hybrid SVM with Logistic Regression (LR) classifier. Authorship verification is investigated as a two-class problem. Furthermore a weighting strategy for unbalanced dataset, different SVM kernels, and different features selection strategy are tested. The proposed approach is evaluated experimentally by computing the following performance metrics:

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

Different block sizes of characters (140, 280 and 500) are tested on Enron and Twitter datasets, yielding EER ranging from 9.98% to 21.45% for different block sizes. The results are very encouraging considering the existing works on authorship verification using stylometry.

The remainder of the paper is organized as follows. Section 2 summarizes and discusses related works. Section 3 presents an outline of the proposed approach. Section 4 presents the experimental evaluation of the proposed approach and the obtained results. Section 5 discusses the strengths and shortcomings of the proposed approach. Section 6 makes some concluding remarks and discusses future work.

2. Related work

As mentioned above, research on authorship analysis covers three different areas: authorship identification, authorship characterization, and authorship verification. A considerable number of studies have been conducted on authorship identification and characterization. For instance, previous studies on authorship identification investigated ways to identify patterns of terrorist communications [8], the author of a particular e-mail for computer forensic purposes [9–11], as well as how to collect digital evidence for investigations [12] or solve a disputed literary, historical [13], or musical authorship [14–16]. Work on authorship characterization has targeted primarily gender attribution [17–19] and the classification of the author education level [20]. However, there are few papers on authorship verification outside the framework of plagiarism detection [6], and most of them focus on general text documents. In addition, the performance of authorship verification for online documents is affected by the text size, the number of candidate authors, the size of the training set, and the fact that these documents are in general quite poorly structured or written (as opposed to literary works).

Among the few studies available on authorship verification are works by Koppel et al. [6], Iqbal et al. [10], Canales et al. [5], and Chen and Hao [21].

Koppel et al. used SVM with linear kernel and addressed the authorship verification as a one-class classification problem, ignoring negative samples. The corpus used in their study was composed by 21 English books written by 10 different authors. They divided the text into approximately equal sections of 500 words, preserving the paragraphs. The feature set was composed by the 250 most frequent words. They introduced a technique named “unmasking” where they quantify the dissimilarity between the sample document produced by the suspect and that of other users (i.e. imposters). Although the overall accuracy was 95.7%, they concluded that the use of negative examples could improve the results.

Iqbal et al. experimented with two different approaches [10]. The first approach conducts verification using classification; three different classifiers are investigated, namely, Adaboost.M1, Bayesian Network, and Discriminative Multinomial Naive Bayes (DMNB). The second approach conducts verification by regression; three different classifiers were studied including linear regression, SVM with Sequential Minimum Optimization (SMO), and SVM with RBF kernel. The feature set was composed of 292 attributes, which included lexical (collected either in terms of characters or words), syntactic (punctuation and function words), idiosyncratic (spelling and grammatical mistakes) and content-specific (keywords commonly found in a specific domain). Experimental evaluation of the proposed approach, using the Enron e-mail corpus and by analyzing 200 e-mails per author, yielded EER ranging from 17.1% to 22.4%.

Canales et al. trained a K-Nearest Neighbor (KNN) classifier with 82 stylistic features including 49 character-based, 13 word-based, and 20 syntactic features [5]. In addition, they combined stylometry and keystroke dynamics analysis for the purpose of authenticating online test takers. They experimented with 40 students with sample document size ranging between 1710 and 70 300 characters, and obtained as performances (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%. They concluded that the feature set must be extended and certain type of punctuations may not necessarily represent the style of students when taking online exams.