Authorship Verification for Short Messages using Stylometry

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

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

***Abstract*—Authorship verification can be checked using stylometric techniques through the analysis of linguistic styles and writing characteristics of the authors. Stylometry is a behavioral feature that a person exhibits during writing and can be extracted and used potentially to check the identity of the author of online documents. Although stylometric techniques can achieve high accuracy rates for long documents, it is still challenging to identify an author for short documents, in particular when dealing with large authors populations. These hurdles must be addressed for stylometry to be usable in checking authorship of online messages such as emails, text messages, or twitter feeds. In this paper, we pose some steps toward achieving that goal by proposing a supervised learning technique combined with *n-gram* analysis for authorship verification in short texts. Experimental evaluation based on the Enron email dataset involving 87 authors yields very promising results consisting of an Equal Error Rate (EER) of 14.35% for message blocks of 500 characters.**

Keywords—Authentication and access control, biometrics systems, authorship verification, classification, stylometry, n-gram features, short message verification, text mining, writeprint.

I. Introduction

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], [3], [16], [23]. Authorship analysis can be carried 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.

According to Koppel et al., "verification is significantly more difficult than basic attribution and virtually no work has been done on it, outside the framework of plagiarism detection" [16]. 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).

We address the above challenge by proposing a new supervised learning technique combined with a new *n-gram* analysis approach to check the identity of the author of a short online document. 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 may falsely recognize someone as the genuine author of a document while they are not;
- False Rejection Rate (FRR): measures the likelihood that the system will fail to recognize the genuine author of a document;
- Equal Error Rate (ERR): corresponds to the operating point where FAR and FRR have the same value.

Our evaluation yields an EER of 14.35%, 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

In this section, we present related works on stylometry for authorship attribution, characterization, and verification.

A. Authorship Attribution or Identification

Despite significant progress achieved on the identification of an author within a small group of individuals, it is still challenging to identify an author when the number of candidates increases or when the sample text is short as in the case of e-mails or online messages.

For instance, Chaski (2005) reported 95.70% accuracy in their work on authorship identification, the evaluation sample consisted of only 10 authors [4]. Similarly, Iqbal et al. (2010) achieved when using k-means for author identification, classification accuracy of 90% with 3 authors; the rate decreased to 80% when the number of authors increased to 10 [13]. Iqbal et al. (2008) also proposed another approach named AuthorMiner [15], which consists of an algorithm that captures frequent lexical, syntactical, structural and content-specific patterns. The experimental evaluation used a subset of the Enron dataset, varying from 6 to 10 authors, with 10 to 20 text samples per author. The authorship identification accuracy decreased from 80.5% to 77% when the authors population size increased from 6 to 10.

Hadjidj et al. (2009) used the C4.5 and SVM classifiers to determine authorship [10], and evaluated the proposed approach using a subset of three authors from the Enron dataset. They obtained as correct classification rates 77% and 71% for sender identification, 73% and 69% for sender-recipient identification, and 83% and 83% for sender-cluster identification, for C4.5 and SVM, respectively.

B. Authorship Characterization

Works on authorship characterization have targeted the determination of various traits or characteristics of an author such as gender and education level.

Cheng et al. (2011) investigated the author gender identification from text by using Adaboost and SVM classifiers to analyze 29 lexical character-based features, 101 lexical word-based features, 10 syntactic, 13 structural, and 392 functional words. Evaluation of the proposed approach involving 108 authors from the Enron dataset yielded classification accuracies of 73% and 82.23%, for Adaboost and SVM, respectively [6].

Abbasi and Chen (2005) analyzed the individual characteristics of participants in an extremist group web forum using decision tree and SVM classifiers. Experimental evaluation yielded 90.1% and 97% success rates in identifying the correct author among 5 possible individuals for decision tree and SVM, respectively [1].

Kucukyilmaz et al. (2008) used k-NN classifier to identify the gender, age, and educational environment of a user. Experimental evaluation involving 100 participants grouped in gender (2 groups), age (4 groups), and educational environment (10 groups), yielded accuracies of 82.2%, 75.4% and 68.8%, respectively [19].

C. Authorship Verification

Among the few studies available on authorship verification, are works by Koppel et al. [16], Iqbal et al. [13], Chen and Hao's[5], and Canales et al. [3] .

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) [16]. The experimental evaluation of the approach yields 95.70% of correct verification, but shows that the proposed approach can provide trustable results only for documents of at least 500 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 [13]. Experimental evaluation of the proposed approach using the Enron e-mail corpus yielded EER ranging from 17.1% to 22.4%.

Chen and Hao's (2011) extracted 150 stylistic features from e-mail messages for authorship verification [5]. Experimental evaluation involving 40 authors from the Enron dataset yielded varying classification accuracy rates based on the number of e-mails analyzed. More specifically, 84% and 89% classification accuracy rates were obtained for 10 and 15 short e-mails, respectively.

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 keystrokes and 82 stylistic features were analyzed using a K-Nearest neighbor (KNN) classifier. Experimental evaluation involving 40 students with sample document size between 1710 and 70,300 characters yielded ERR of 30%.

III. Authorship Verification Approach

In this section, we present our approach by discussing feature selection and describing in detail our classification model.

A. Feature Selection

Over a thousand stylistic features have already been identified and used in the literature along with a wide variety of analysis methods. However, there is no agreement among researchers regarding which features yield the best results. As a matter of fact, analyzing a large number of features does not necessarily provide the best results, as some features provide very little or no predictive information. Our approach is to build on previous works by identifying and keeping only the most discriminating features. According to Abbasi and Chen [2], existing stylistic features can be categorized as lexical, syntactic, structural, content-specific, and idiosyncratic style markers.

Previous studies yielded encouraging results with lexical features [3], [12]. In particular, since *n-gram* features are noise tolerant and effective, and e-mails are non-structured documents, we will focus in this paper only on these types of features.

Although *n-gram* features have been shown to be effective, classification based on such feature is complex while the data processing is time consuming. While the approach used so far in the literature has consisted of computing *n-gram* frequency in given sample document, we propose an innovative approach that analyzes exclusively the presence or absence of *n-grams* and their relationship with the training dataset. This allows us to reduce the number of *n-grams* features to one, and address the above mentioned challenges.

B. Classification Model

Our model consists of a collection of profiles generated separately for individual users. The model involves two modes of operations, namely, training and verification, where the users profiles are built and then checked, respectively. The training phase involves two steps. During the first step, the user profile is derived by extracting *n-grams* from sample documents. During the second step, a user specific threshold is computed and used later in the verification phase.

Given a user U, we divide her training data into two subsets, denoted T_1^U and T_2^U. Let $N(T_1^U)$ denote the set of all unique *n-grams* occurring in T_1^U. We divide T_2^U into p blocks of characters of equal size: $b_1^U, ..., b_p^U$.

Given a block b_i^U, let $N(b_i^U)$ denote the set of all unique *n-grams* occurring in b_i^U.

Given two users U and I, let $r_U(b_i^I)$ denote the percentage of unique *n-grams* shared by block b_i^I (of user I) and (training set) T_1^U, giving:

$r_U(b_i^I) = \frac{|N(b_i^I) \cap N(T_1^U)|}{|N(b_i^I)|}$ where $|X|$ denotes the cardinality of set X.

Given a user 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.

A block b is said to be a genuine sample of user U if and only if $|r_U(b)| \geq (\epsilon_U + \gamma)$, where ϵ_U is a specific threshold for user U, and γ is a predefined constant.

We derive the value of ϵ_U for user U using a supervised learning technique outlined by $Algorithm$ 1. 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 $\gamma = 0$.

$Algorithm$ 2 returns the FAR and FRR for a user U given some training data, a user-specific threshold value, and some constant value assigned to γ.

IV. Experimental Evaluation

We present in this section the experimental evaluation of our proposed approach by describing our dataset and data preprocessing technique, and then outlining our evaluation method and results.

```
/* $U$ a user for whom the threshold is
   being calculated                          */
/* $I_1, ..., I_m$: a set of other users
   ($I_k \neq U$)                            */
/* $\epsilon_U$: threshold computed for user $U$
   */
Input: Training data for $U, I_1, ..., I_m$
Output: $\epsilon_U$
begin
    $up \leftarrow$ false;
    $down \leftarrow false$;
    $\delta \leftarrow 1$;
    $\epsilon_U \leftarrow \mu_U - (\sigma_U/2)$;
    $\gamma \leftarrow 0$;
    while $\delta > 0.0001$ do
        /* Calculating FAR and FRR for
           user U                            */
        $FRR_U, FAR_U =$
            $calculate(U, I_1, ..., I_m, \epsilon_U, \gamma)$;
            /* Minimizing the difference
            between FAR and FRR              */
        if $(FRR_U - FAR_U) > 0$ then
            $down \leftarrow true$;
            $\epsilon_U \leftarrow \epsilon_U - \delta$;
        end
        if $(FAR_U - FRR_U) > 0$ then
            $up \leftarrow true$;
            $\epsilon_U \leftarrow \epsilon_U + \delta$;
        end
        if (up & down) then
            $up \leftarrow false$;
            $down \leftarrow false$;
            $\delta \leftarrow \delta/10$;
        end
    end
    return $\epsilon_U$;
end
```

Algorithm 1: Threshold calculation for a given user.

A. Dataset and Data Preprocessing

In order to validate our system, we performed experiments on a real-life dataset from Enron e-mail corpus[1]. Enron was an energy company (located in Houston, Texas) that was bankrupt in 2001 due to white collar raud. 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.

While traditional documents are very well structured and large in size providing several stylometric features, an e-mail typically consists of a few paragraphs, wrote quickly and often with syntactic and grammatical errors. In our approach, we grouped all the sample e-mails used to build a given author profile into a single document that could be subsequently divided into small blocks.

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

```
Input: $\epsilon_U, \gamma$, Training data for $U, I_1, ..., I_m$
Output: $(FAR_U, FRR_U)$
begin
    /* Calculating FRR for user U */
    for $i \rightarrow 1$ to $p$ do
        $FR \leftarrow 0$;
        if $r_U(b_i^U)) < (\epsilon_U + \gamma)$ then
            $FR \leftarrow FR + 1$;
        end
    end
    $FRR_U \leftarrow \frac{FR}{p}$;
    /* Calculating FAR for user U */
    for $k \rightarrow 1$ to $m$ do
        for $j \rightarrow 1$ to $n$ do
            $FA \leftarrow 0$;
            if $r_U(b_j^{I_k}) \geq (\epsilon_U + \gamma)$ then
                $FA \leftarrow FA + 1$;
            end
        end
    end
    $FAR_U \leftarrow \frac{FA}{p \times m}$; return $(FAR_U, FRR_U)$;
end
```

Algorithm 2: FAR and FRR calculation for a given user

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

- 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;
- Since different texts must be logically equivalent, (i.e., must have the same canonical form), the following filters were applied:
    - Strip e-mail replay;
    - Replace e-mail address for double @ character (i.e. @@);
    - Replace htpd address for a single http word;
    - Replace currency for $XX;
    - Replace percentage for XX%;
    - Replace numbers for the digit 0;
    - Normalize the document to printable ASCII;
    - Convert the document to lowercase characters;
    - Strip white space;
    - Strip any punctuation from the document.
- All messages, per author, were grouped creating a long text or stream of characters that was divided into blocks.

B. *Evaluation Method*

After the preprocessing phase, the dataset was reduced from 150 authors to sets of 107, 92 and 87 authors to ensure that only streams of text with 12,500, 18,750 and 25,000 characters were used in our analysis, respectively.

We assess experimentally the effectiveness of our approach through 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. The 90% training data allocated to a given user U was further divided as follows: 2/3 of the training data allocated to subset T_1^U and 1/3 of the data for subset T_2^U, respectively. The 10% test data for user U was divided in p blocks of equal size s. We tested two different block sizes, $s = 250$ and $s = 500$ characters, respectively. The number of blocks per user p varied from 25 to 100. In addition, we investigated separately *n-grams* of sizes (n=) 3, 4, and 5, for each of these analyses yielding in total 18 different experiments. Table I shows the configuration of our experiments.

TABLE I. CONFIGURATION OF EXPERIMENTS

Experiment configuration #	Number of Users (m)	Number of blocks per author (p)	Block size (s)
1	107	50	250
2	92	75	
3	87	100	
4	107	25	500
5	92	37	
6	87	50	

For each user U, we computed a corresponding profile by using their training data and training data from other users considered as impostors. This allows computing the acceptance threshold ϵ_U for user U as explained before. A given block b is considered to belong to an hypothesized genuine user U when the ratio $|r_U(b)|$ is greater than $\epsilon_U + \gamma$, where γ is a predefined constant and ϵ_U is the user specific threshold.

We compute the FRR for user U by comparing each of the blocks from her test data against her profile. A false rejection (FR) is counted when the system rejects one of these blocks. The FAR is computed by comparing each of the test blocks 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 blocks 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.

C. *Evaluation Results*

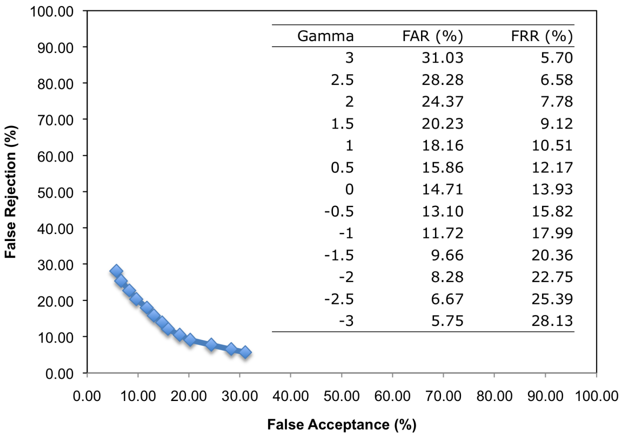
Table II shows the overall FRR and FAR for the 18 experiments, where the constant $\gamma = 0$. It can be noted that the accuracy decreases not only when the number of authors increases, but also when the number of blocks per user p and the block size s decreases.

Experiments using 5-*grams* achieve better results than those using 3 and 4-*grams* for large number of blocks per user and block size. Experiments using 4-*grams* yield better results when the number of blocks per user decreases. Overall, the best result is achieved in experiment 6, with 87 authors, 50 blocks per user, and a block size of 500 characters (FRR=14.71%, FAR=13.93%).

Figure 1 shows the receiver operating characteristic (ROC) curve for experiment configuration # 6 (from Table I) using 5-*gram*. The curve illustrates the relationship between the FRR and FAR for different values of γ. The equal error rate (ERR) was estimated as 14.35% and achieved when $\gamma = -0.25$.

TABLE II. PERFORMANCE RESULTS FOR THE DIFFERENT EXPERIMENTS ($\gamma = 0$)

No.	*3-gram*		*4-gram*		*5-gram*	
	FRR	FAR	FRR	FAR	FRR	FAR
1	24.85	28.61	22.05	24.09	24.11	20.50
2	26.76	26.82	23.64	21.68	25.13	19.39
3	24.82	28.15	23.56	21.15	17.24	20.39
4	26.47	22.70	23.67	17.81	23.98	16.29
5	23.36	21.81	18.75	18.01	18.20	15.40
6	22.29	22.21	19.77	16.11	**14.71**	**13.93**



Gamma	FAR (%)	FRR (%)
3	31.03	5.70
2.5	28.28	6.58
2	24.37	7.78
1.5	20.23	9.12
1	18.16	10.51
0.5	15.86	12.17
0	14.71	13.93
-0.5	13.10	15.82
-1	11.72	17.99
-1.5	9.66	20.36
-2	8.28	22.75
-2.5	6.67	25.39
-3	5.75	28.13

Fig. 1. Receiver Operating Characteristic curve for experiment configuration #6 using 5-*gram* and sample performance values for different values of γ.

V. DISCUSSIONS

The Enron dataset has previously been used not only in authorship verification [5], but also in authorship identification [2], [10], [13], [14], [15] and authorship characterization [6], [7], [14]. These previous experiments used a number of users ranging from 3 to 114, and achieved in the best cases EER varying from 17% to 30%. In the present study, the best configuration was achieved with block size of 500 characters, achieving EER below 15% which is better compared to the accuracy obtained using similar techniques in the literature. Table III summarizes the performances, block sizes, and population size of previous stylometry studies.

Despite our encouraging results, more works must be done to improve the accuracy to an acceptable level for authorship verification in forensics investigation. We believe that our proposed scheme is a good step toward achieving that goal. It is important to notice that these results were obtained using only one type of features out of hundreds of potential stylometric features. We believe that we can reduce significantly our error rates by incorporating other types of features in our framework.

We investigated in this work block sizes of 250 and 500 characters, respectively, which represent significantly shorter messages compared to the messages used so far in the literature for identity verification. To our knowledge, one of the few works that have investigated comparable message sizes includes the work by Sanderson and Guenter, who split a long text in chunks of 500 characters [23]. We still need to investigate even shorter messages (e.g. 10 to 50 characters) to be able to cover (beyond emails) a broader range of online messages such as twitter feeds and text messages. However, attempting to reduce at the same time the block size and 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. We intend to tackle such challenge in the future.

Another important limitation of many previous stylometry studies is that the performance metrics computed during their evaluations cover only one side of the story, and this is clearly emphasized by Table III. Accuracy is traditionally measured using the following two different types of errors:

1) Type I error, which corresponds to the FRR, also referred to as False Non-Match Rate (FNMR) or False Positive Rate (FPR);
2) Type II error, which corresponds to the FAR, also referred to as False Match Rate (FMR) or False Negative Rate.

However, most previous studies calculate the so-called (classification) accuracy (see Table III) which actually corresponds to the true match rate and allows deriving only one type of error, namely, Type II error: $FAR = 1 - Accuracy$. Nothing is said about Type I error in these studies, which makes it difficult to judge their real stremngth in terms of accuracy. As shown by Table III, only few studies have provided both types of errors, among which our work can be considered as one of the most strongest in terms of sample population size, block size, and accuracy.

VI. CONCLUSION

We have investigated in this work the possibility of using stylometry for authorship verification for short online messages. Our technique is based on a combination of supervised learning and *n-gram* analysis. Our 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. Our experimental evaluation yields an EER 14.35% for 87 users for relatively small block sizes. While the obtained results are promising, it is clear that more work must be done for the proposed scheme to be usable in real-world forensics investigation of online activities. We discussed the limitations of our approach and plan to address them in our future work. In particular, we will improve verification accuracy by expanding our feature set beyond *n-grams*. We will also improve the robustness of the scheme in handling shorter and shorter message structures.

VII. ACKNOWLEDGMENTS

This research has been enabled by the use of computing resources provided by WestGrid and Compute/Calcul Canada. M.L.B. is funded through a Vanier doctoral scholarship from Natural Sciences and Engineering Research Council of Canada (NSERC) and Bolsista do CNPq Brasil.

REFERENCES

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

TABLE III. COMPARATIVE PERFORMANCES, BLOCK SIZES AND, POPULATION SIZES FOR STYLOMETRY STUDIES.

Category	Reference	Sample Population Size	Block Size	Accuracy* (%)	EER (%)
Attribution	[2]	100 **	277 words	83.10	- -
	[4]	10	200 words	95.70	- -
	[8]	2 - 4	60,000 words	93.8 - 97.8	- -
	[10]	3 **	200 words	69 -83	- -
	[12]	87	287 words	50 - 60	- -
	[13]	3 - 10 **	200 words	80 - 90	- -
	[14]	4 - 20 **	300 words	69.75 - 88.37	- -
	[15]	6 - 10 **	200 words	77 - 80.5	- -
	[17]	1000	500 words	42.2 - 93.2	- -
	[20]	20	169 words	99.01	- -
	[21]	20	600 words	84.30	- -
	[23]	50	500 characters	- -	8.08 - 30.88
Characterization	[1]	5	76 words	90.1 - 97	- -
	[6]	108 **	50 - 200 words	73 - 82.23	- -
	[7]	114 **	50 - 200 words	80.08 - 82.20	- -
	[9]	325	50 - 200 words	70.20	- -
	[14]	4 - 20 **	300 words	39.13% - 60.44%	- -
	[19]	100	300 words	39.0 - 99.70	- -
	[22]	10 - 40	450 words	68.3 - 91.5	- -
Verification	[3]	40	1710 - 70300 characters	- -	30
	[5]	25 - 40 **	30 - 50 words	83.90 - 88.31	
	[11]	8	628 - 1342 words	- -	3
	[16]	10	500 words	95.70	- -
	[18]	29	2400 words	- -	22
	Proposed Approach	87	500 character	- -	14.35%

* The accuracy is measured by the percentage of correctly matched authors in the testing set.
* Used Enron dataset for testing.

[2] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26:7:1–7:29, 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. CSIS, Pace University, May 6 2011.

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

[5] 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*, MLDM'11, pages 375–386, Berlin, Heidelberg, 2011. Springer-Verlag.

[6] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78 – 88, 2011.

[7] 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*, pages 154 –158, 30 2009-april 2 2009.

[8] 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*, MICAI'07, pages 839–849, Berlin, Heidelberg, 2007. Springer-Verlag.

[9] 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*, pages 282 – 289, 2002.

[10] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem. Towards an integrated e-mail forensic analysis framework. *Digital Investigation*, 5(3-4):124 – 137, 2009.

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

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

[13] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1-2):56 – 64, 2010.

[14] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, (0):–, 2011.

[15] 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*, 5, Supplement(0):S42 – S51, 2008. The Proceedings of the Eighth Annual DFRWS Conference.

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

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

[18] I. Krsul and E. H. Spafford. Authorship analysis: identifying the author of a program. *Computers and Security*, 16(3):233 – 257, 1997.

[19] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing Management*, 44(4):1448 – 1466, 2008.

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

[21] 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*, pages 936 –940, july 2009.

[22] 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*, CoNLL '11, pages 78–86, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[23] 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*, EMNLP '06, pages 482–491, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.