World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:6, No:3, 2012

Detecting Email Forgery using Random Forests and Naïve Bayes Classifiers

Emad E Abdallah, A.F. Otoom, ArwaSaqer, Ola Abu-Aisheh, Diana Omari, and Ghadeer Salem

***Abstract*—**As emails communications have no consistent authentication procedure to ensure the authenticity, we present an investigation analysis approach for detecting forged emails based on Random Forests and Naïve Bays classifiers. Instead of investigating the email headers, we use the body content to extract a unique writing style for all the possible suspects. Our approach consists of four main steps: (1) The cybercrime investigator extract different effective features including structural, lexical, linguistic, and syntactic evidence from previous emails for all the possible suspects, (2) The extracted features vectors are normalized to increase the accuracy rate. (3) The normalized features are then used to train the learning engine, (4) upon receiving the anonymous email (*M*); we apply the feature extraction process to produce a feature vector. Finally, using the machine learning classifiers the email is assigned to one of the suspects' whose writing style closely matches *M*. Experimental results on real data sets show the improved performance of the proposed method and the ability of identifying the authors with a very limited number of features.

***Keywords*—**Digital investigation, cybercrimes, emails forensics, anonymous emails, writing style, and authorship analysis

Open Science Index, Computer and Systems Engineering Vol:6, No:3, 2012 publications.waset.org/3339.pdf

I. Introduction

THE increased rate of email misuse attacks in the recent years has triggered the researchers to develop efficient methodologies or algorithms that help analyzing suspects' emails in order to gather clues and evidence about the authorship. This evidence could be used in the future to determine the likelihood of a specific suspect is the author of an anonymous email by examining other work previously produced by that author [1]. One of the major difficulties facing email forensics is the large amount of emails that need to be inspected [2]. Moreover, the sender may attempt to hide his true identity in order to avoid detection [3]. This could be done easily by routing email through several anonymous servers to hide some information about the source of the email. Most of the times, email structure is the only way to identify the author.

Authorship identification is the process of examining the features of a malicious email in order to draw conclusions on its authorship form a list of suspects. Clearly, the cybercrime investigator needs to gather several convincing clues from the malicious email and compares it with suspects writing style.

Emad E Abdallah, A.F. Otoom, ArwaSaqer, Ola Abu-Aisheh, Diana Omari, and Ghadeer Salem are with Faculty of Prince Al-Hussein Bin Abdallah II forInformation Technology,Hashemite University,P.O. Box 150459, Zarqa 13115, Jordan. E-mail:Emad@hu.edu.jo

Early algorithms on the problem of authorship in the context of email forensics introduced in [4] where a toolkit called IDENTIFIED is developed to assists with the automatic extraction of a wide variety of metrics. These metrics are used for software forensics and authorship analysis. In [3], different email features such as structural characteristics and linguistic pattern were derived; the Support Vector Machine (SVM) is employed as a learning engine. The main limitation of the above algorithms is that there is no steady categorization performance for all suspects. To overcome this limitation, a combination of features such as relative function word frequencies should be considered.

Latest research in the area of authorship attribution is done by [2-9]. The write-print of every suspect is collected as combinations of features that occurred frequently. After extracting the frequent pattern from each suspect, the common frequent patterns are filtered to have a unique pattern to every suspect. The unknown email is converted into a feature vector and compared with the set of the unique patterns to discover which vector is the most closely matched. In [9], all the training emails are clustered by set of stylometric features. Then, a unique writing style is extracted from each cluster. This technique is useful when no suspects list or training examples are known to the cybercrime investigator. Stylometry clustering is applied to categorize the main groups of stylistics belonging to different suspects. The frequent pattern is extracted from each category to provide a unique writing style to of each cluster. The problem with the stylometry clustering is that the accuracy rate is decreased when the number of the candidate suspect is increased.

Motivated by the need for a better categorization performance with an enhanced accuracy rate, we propose new features that demonstrate great improvement to the authorship verification problem. The comprehensive set of the extracted features include lexical, syntactic, structural characteristics, content specific, and the author positive/negative emotions. Our approach employed the Random Forest and the Naïve Bayes as learning engines.

The remainder of this paper is organized as follows. In Section II, we briefly review some background material and describe the authorship verification problem, Random Forests and Naïve Bayes classifiers. In Section III, we introduce the proposed approach and describe in detail the features

Open Science Index, Computer and Systems Engineering Vol:6, No:3, 2012 publications.waset.org/3339.pdf

extraction process. In Section IV, we present some experimental results. Finally, we conclude and point out future directions in Section V.

II. Related work

In this section, we formally define the Authorship verification problem and describe the Random Forest and the Naïve Bayes classifier.

A. Authorship verification problem statement

The cyber forensics investigator received a case from the court with unknown sender of a malicious email. The investigator needs to identify a particular author of the email from a set of suspects where each suspect has a set of previous emails to be used in the training stage. The investigator then, captures the writing style of every suspect by extracting the frequent features of his/her writings (See section 3.1). The extracted features are then used to generate a classification model (See section 3.2). The investigator finally, extracts all the possible features from the malicious email and feed the classifier model to identify the suspect whose writing style closely matches the received malicious email. The authorship verification problem is summarized in Figure 1.

B. Random Forests

Random Forest has been one of the most popular machine learning methods for classification [10]. It produces various set of decision classification trees. To classify an object from a feature vector, we run down the feature vector with every tree in the forest to provide a classification result called vote. The forest chooses the one receiving the most votes to determine the class of the object. Let $D_n = \{(Xi, Yi): i = 1, .., n\}$, where, $X_i = \left(X_i^{(1)}, \dots, X_i^{(n)}\right) \in R^d$, $Y_i \in R$, be the training data set. Then, the Random Forest is constructed by creating K independent samples B_k from D_n at each $B_k, k = 1, \dots, K$. At each node the best feature variable need to be selected as a split point.

Random Forest is computationally fast classifier and runs efficiently on a large data set. Moreover, it can handle thousands of variables and it has an effective method for calculating the missing data. Random Forest also keeps the correctness when a large amount of the data is missing. The computational, efficiency and the accuracy advantages of the Random Forests make it a perfect machine learning engine for email forensics.

C. Naïve Bayes

Naïve Bayes classifiers are statistical classifiers developed to predict the probability of a given object belongs to a particular class. It is used when the dimensionality of the inputs is high. Naive Bayesian classifier [11,12] is based on the Bayes theorem and assumes that the effect of a feature on a given class is independent of the other features (class conditional independence). Despite its simplicity, Naïve Bayes classifier is computationally fast and in this sense, is it suitable for email forensics. Moreover, it often outperforms more sophisticated classification methods. Let T are the training samples, each with their class labels. There are K classes, $\{C_1, C_2, \dots, C_k\}$, each sample is represented by an η-dimensinal vector, $X = (x_1, \dots, x_\eta)$, depicting η measured values of the η attributes. Given a sample X, the classifier will predict that X belongs to the class having the highest probability, conditioned on X. That is X is predicted to belong to the class C_i if and only if $P(C_i|\mathrm{X}) > P\left(C_j|\mathrm{X}\right)$. Thus we find the class that maximizes $P(C_i|\mathrm{X})$.

III. Proposed approach

Authorship analysis includes authorship attribution, verification, profiling, and/or similarity detection. In the proposed approach we considered different type of features and several machine learning techniques for identification. Based on the previous studies there is no predefined special feature set that can be used to determine the writing style [2]. The writing characteristics contain 1- Lexical features, 2- Syntactic features, 3- Content-specific features, 4- Structural features, and 5- Idiosyncratic Features.

A. Feature Extraction

Individuals have a unique or near to a unique writing style, in this section we describe the extraction process of some certain features that we found have the most impact on email content mining for author identification. Word usage, selection of special characters, composition of sentences and paragraphs and the organization of sentences into paragraphs are used in [9] for defining the writing style.

The total number of stylometric features used in [6] exceeded 1000 features and in [9] used 419 features. In order to reduce training and improve the classification rate, our experimental results show that only 16 features (See table 1) need to be considered. The author emotions that he/she expressed in their emails, their distinct way of writing some phrases are the most important features for recognizing the author of a given email. A summary of the selected features are shown in Table 1.

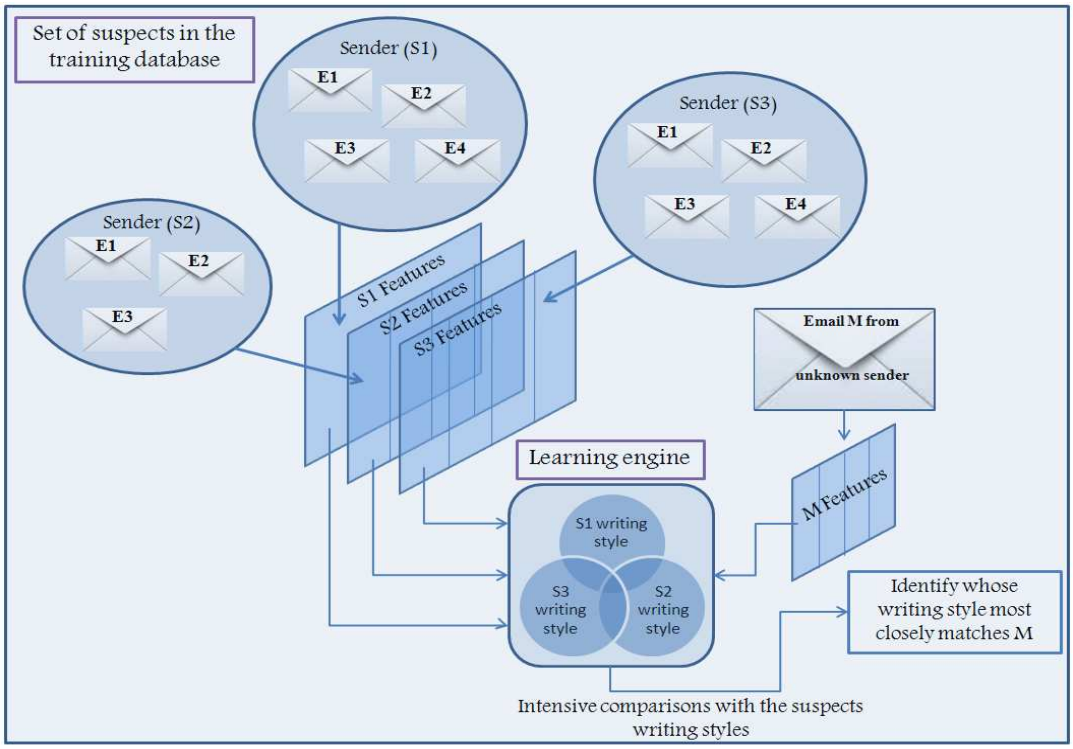

Fig. 1 Authorship verification process

We used the Enron corpus email data set [13] that is mostly recognized for the research of email forensics. The Enron corpus is a set of real email collected and prepared by MIT. It contains data from about 150 users. Figure 2 shows one email written by one of the authors of Enron email corpus. As the features have different types and weight, we applied a normalization formula to balance all the extracted features to numerical values. The experimental results show that the complexity, positive emotions, negative emotions, minimum sentence length, and two word phrases have a significant impact in uniquely identifying the author writing style.

B. Email Classification

Emails in the training and testing data sets are labeled with known authors. The extracted feature of each email is represented as a features vector. The classification model is built using the set of the training features of already classified instances. Random Forests and Naïve Bayes are used as learning engines (See section 2). We utilize the Weka data mining toolkit [14] for simulating the two machine learning algorithms.

The feature vectors of the testing emails are used to evaluate the accuracy of the proposed algorithm. Each email is assigned to one of the suspects authors using the machine learning classifiers.

TABLE I
SUMMARY OF THE 16 EXTRACTED FEATURES

Self-reference	Average sentence length
Social words	Minimum sentence length
Positive emotions	Ratio of dots
Negative emotions	Ratio of commas
Overall-cognitive words	Ratio special characters
Articles (a,an,the)	The Complexity of the text
Big words	Two word phrases frequency
Maximum sentence length	The Readability of the text

The correctly and incorrectly classified emails are reported to calculate the classification accuracy rate and the false acceptance rate:

$$Classification\ accuracy = \frac{Number\ of\ correctly\ classified\ identities}{Total\ number\ of\ identities}$$

Open Science Index, Computer and Systems Engineering Vol:6, No:3, 2012 publications.waset.org/3339.pdf

```
The winner of the 9/15 Dave Matthews Tickets accidently bid on both
pairs of tickets I had for sale (9/15 and 9/16) and only wanted one
pair.  He is buying the 9/16 tickets (11th row) for $345.  If you are
interested in the 9/15 tickets (8th row) which are even better seats,
I can sell them to you for $305.  Let me know asap because I am
leaving town around 2pm for the weekend.  You can reach me at work or
anytime this weekend on my cell phone.

If you are interested please call me at 713-853-5933 (work) or 713-
516-8820 (cell) and we will work it out.  FYI - I am also notifying
the person who bid $305 and promised to give them first look but
otherwise the tickets are yours if you want.

Thanks,

Martin
```

Fig. 2 An email example for the author Martin (Class X1) in the Enron corpus

Open Science Index, Computer and Systems Engineering Vol:6, No:3, 2012 publications.waset.org/3339.pdf

IV. Experimental Evaluation

To evaluate the performance of our proposed algorithm, we performed several experiments on the Enron email data set. A subset of four senders from the original dataset is randomly selected. We show our experiment on 76 emails from the four different senders where no restrictions on the recipients have considered. Although most of the research in the literature is manually filtered the emails to have a common format, we tried to reduce the manual filtering to minimum. We constructed six different groups of training and testing sets. Each experiment is repeated five times, in order to increase the reliability of the presented results.

In our experiments, we split the emails into 60% for training and 40% for the purpose of testing. In the training phase, we feed the training emails to the machine learning algorithm to depict a unique writing style of each sender. In the testing phase, we feed the testing emails randomly to the classification model to identify the sender of each email in testing set. In the first experiment, we applied our feature extraction tool on 10 emails randomly selected from two suspects. Figures (3-6) depict the extracted results of four different features. Figure 3, shows clearly that positive emotion appears on all the samples written by suspect 1. However, it appears only twice in suspects 2 emails. Hence, we set the positive emotions feature for suspect1 and it becomes part of his writing style. Similarly from Figure 4, negative emotion feature is set to suspect 2. Note that it is possible to set the same feature for several suspects.

The Second experiment is used for analyzing the performance of the proposed email authorship identification system using Random Forests and Naïve Bayes classifiers over the Enron email data set. The obtained results are provided in Table II.

As can be observed from the experimental results, classification accuracy obtained using the Random Forests comparatively better than the recognition accuracy achieved using the Naïve Bays classifier.

V. Conclusion

In this paper, we presented a simple and computationally inexpensive investigation analysis tool for authorship identification of anonymous emails. The key idea is to extract selected effective features from the suspects' previous writings to be used in the learning process. Two different learning algorithms are used to build classification models. To evaluate the effectiveness of the proposed analysis tool, we conducted several experiments on a real emails data set. The results clearly showed the ability of identifying the authors with very limited number of features. For future work, we plan to analyze the relationship between the number features used in the extraction process, optimal two words phrases, and the best learning engine to further improve the classification performance in the context of email forensics.

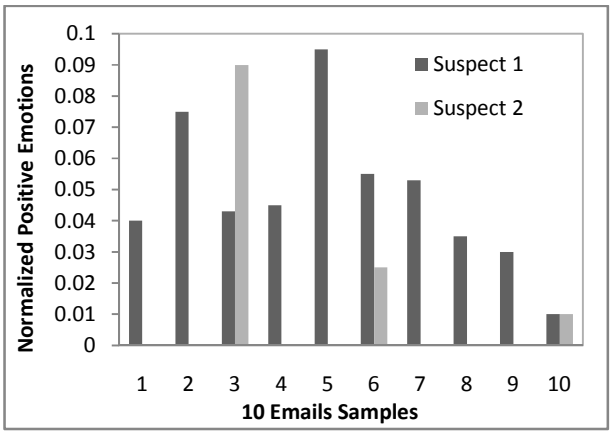

Fig. 3 Positive emotions feature vs. 10 random emails for two suspects from Enron email dataset

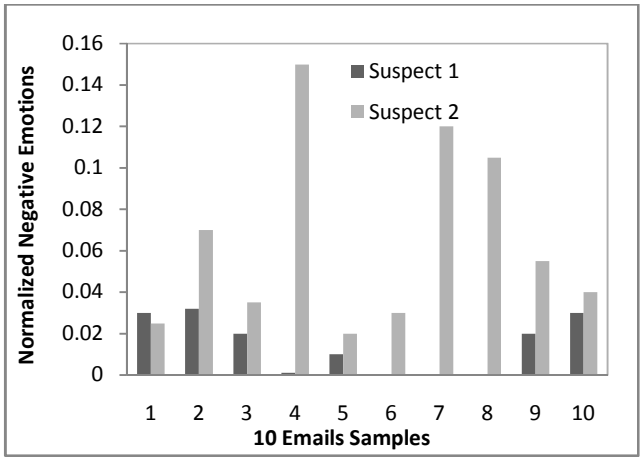

Fig. 4 Negative emotions feature vs. 10 random emails for two suspects from Enron email dataset

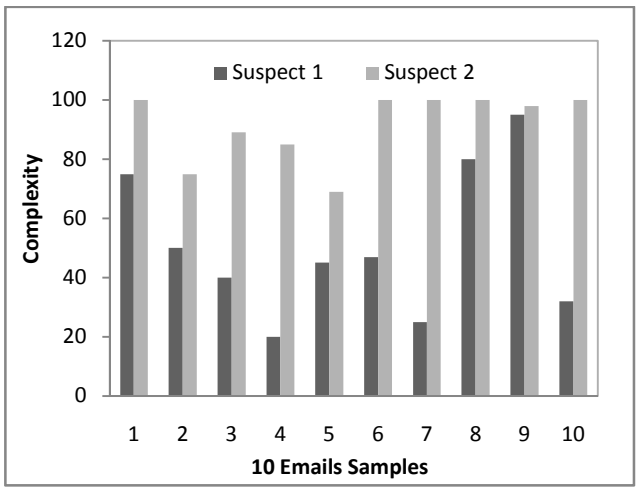

Fig. 5 Email complexity feature vs. 10 random emails for two suspects from Enron email dataset

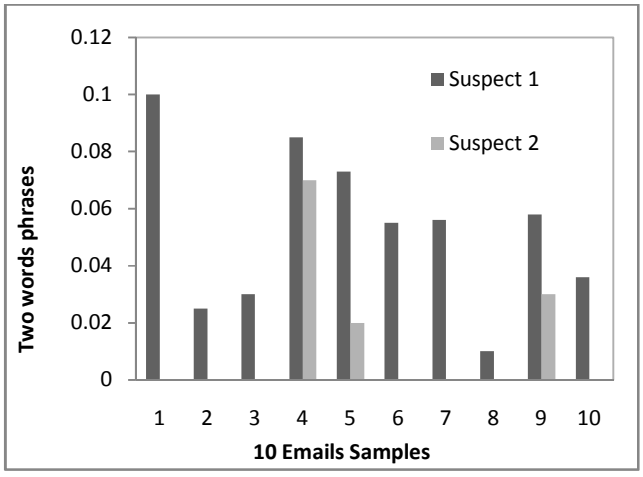

Fig. 6 Two words phrases feature vs. 10 random emails for two suspects from Enron email dataset

TABLE II
PERFORMANCE OF THE PROPOSED AUTOMATED EMAIL AUTHORSHIP IDENTIFICATION SYSTEM IN CLASSIFYING ANONYMOUS EMAILS OVER THE ENRON EMAILS DATA SET

Class	Recognition Accuracy (%) over the Enron email database	
	Naïve Bayes	Random Forests
X_1	92.307%	100%
X_2	60.00%	90.00%
X_3	65.217%	82.60%
X_4	60.00%	80.00%
Average	77.631%	86.842%

REFERENCES

[1] T. McElroy and J. J. Seta, "Framing the frame: How task goals determine the likelihood and direction of framing effects," Judgment and Decision Making, Vol. 2 (4), Aug 2007, pp. 251-256.
[2] F. Iqbal, R. Hadjidj, B.C.M. Fung, M. Debbabi, "A novel approach of mining write-prints for authorship attribution in email forensics," Digital Investigation, Vol. 5 (1), 2008, pp. 42-51.
[3] O. De Vel, A. Anderson, M. Corney, and G. Mohay, "Mining Email Content for Author Identification Forensics", SIGMOD Record, Vol. 30(4), 2001, pp. 55-64.
[4] A. Gray, P. Sallis, and S. MacDonell, "Software Forensics: Extending Authorship Analysis Techniques to Computer Programs," in the 3rd Biannual Conference International Association of Forensic Linguists, 1997.
[5] M. Koppel, S. Argamon, and A.R. Shimoni, "Automatically categorizing written texts by author gender," Literary and Linguistic Computing, Vol. 17(4), 2002, pp. 401–412.
[6] A. Abbasi, and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," ACM Transactions on Information Systems, Vol. 26(2), March 2008, pp. 1-29.
[7] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," Journal of the American Society for Information Science and Technology, Vol. 60(1), 2009, pp. 9–26.
[8] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," Journal of the American Society for Information Science and Technology, Vol. 57(3), February 2006, pp. 378–393,.
[9] F. Iqbal, H. Binsalleeh, B.C.M. Fung, and M. Debbabi, "Mining writeprints from anonymous emails for forensic investigation," Digital Investigation, 2010, pp. 1–9.
[10] L. Breiman, "Random forests," Machine Learning, 2001, pp. 5–32.
[11] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine Learning, 2001, pp. 103–137.
[12] DJ. Hand and K. Yu, "Idiot's Bayes - not so stupid after all?," International Statistical Review, Vol. 69(3), 2001, pp. 385-399.
[13] L. Kaelbling, "Enron email dataset," CALO Project, http://www.cs.cmu.edu/enron/, August 21 2009.
[14] I. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Margan Kaufmann, San Francisco, 2nd edition, 2005.

Open Science Index, Computer and Systems Engineering Vol:6, No:3, 2012 publications.waset.org/3339.pdf