# E-mail Authorship Verification for Forensic Investigation

Farkhund Iqbal
Computer Security Laboratory
CIISE, Concordia University
Montreal, Quebec, Canada
iqbal_f@ciise.concordia.ca

Liaquat A. Khan
Computer Security Laboratory
CIISE, Concordia University
Montreal, Quebec, Canada
lakhan@ciise.concordia.ca

Benjamin C. M. Fung
Computer Security Laboratory
CIISE, Concordia University
Montreal, Quebec, Canada
fung@ciise.concordia.ca

Mourad Debbabi
Computer Security Laboratory
CIISE, Concordia University
Montreal, Quebec, Canada
debbabi@ciise.concordia.ca

## ABSTRACT

The Internet provides a convenient platform for cyber criminals to anonymously conduct their illegitimate activities, such as phishing and spamming. As a result, in recent years, authorship analysis of anonymous e-mails has received some attention in the cyber forensic and data mining communities. In this paper, we study the problem of authorship verification: given a set of e-mails written by a suspect along with an e-mail dataset collected from the sample population, we want to determine whether or not an anonymous e-mail is written by the suspect. To address the problem of authorship verification of textual documents and employ detection measures that are more suited in the context of forensic investigation, we borrow the NIST's speaker recognition evaluation (SRE) framework. Our experimental results on real world e-mail dataset suggest that the employed framework addresses the e-mail authorship verification problem with a matching success as in case of speaker verification. The proposed framework produces an average equal error rate of 15-20% and minDCF equal to 0.0671 (with 10-fold cross validation technique) in correctly verifying the author of a malicious e-mail.

## Keywords

cyber crimes, e-mail forensics, classification, stylometric features, regression

## 1. INTRODUCTION

Authorship analysis for resolving disputes over old literature has a long history in academic research [4][24]. However, during the last two decades, authorship analysis of computer-mediated communication (CMC) or online documents (such as e-mails, VoIP segments and instant messages, etc.) for prosecuting terrorists, pedophiles, and scammers in the court of law, has received great attention in different studies [6][1][12]. Authorship analysis includes authorship attribution, characterization/profiling and verification or similarity detection. In this paper, we focus on the authorship verification problem in which the goal of an investigator is to confirm whether or not a given suspect is the true author of a disputed textual document. In most plagiarism disputes, an investigator (or an expert witness) needs to decide whether the given two objects are produced by the same entity or not. The object in question can be a piece of code, a theory, a textual document or an online message. More importantly, the conclusion drawn needs not only be precise (up to a possible extent) but needs to be supported by a strong evidence as well.

Authorship analysis of CMC documents is different from the authorship analysis of traditional works in two ways. First, most online documents are relatively short in length, containing few lines or paragraphs, and are often poorly structured (containing spelling and grammatical errors) usually written in para language as compared to literary and poetic works, which are large in size and are well structured following definite syntactic rules. Second, the number of potential candidates in online documents ownership disputes are much more than the traditional authorship disputes. However, additional information including header, subject, attachment, timestamp (for instance, in case of e-mail documents) contribute towards a more deeper analysis of online documents.

Most previous studies on authorship verification focus on general text documents. Studies on CMC or online documents are limited. Similarly, features of online documents, such as the structural features of e-mails, are different than the traditional textual works. Most of the discussion in [13] focus on authorship analysis of books while studies of Halteren [21] focus on analyzing students' essays. The experimental results of Koppel et al. [13] indicate that the proposed method is suitable in situations where the document in question is at least 5000 words long for achieving trustable results. This is nearly impossible in case of online documents, such as e-mails.

In this paper, we formally define the problem of author-

*SAC'10* March 22-26, 2010, Sierre, Switzerland.

ship verification and propose an authorship verification framework for e-mails, a typical online document. Our method is primarily based on the speaker recognition evaluation (SRE) framework developed by National Institute of Standards and Technology (NIST) [16], which has proven very successful in the speech processing community. The SRE framework evaluates the performance of detection systems in terms of minDCF, false positive and false negative alarms represented by employing detection error trade-off (DET) curve, a deviant of receiver operating characteristic (ROC) curve (see details in Section 3).

The overview of the proposed approach is shown in Fig. 1. The two e-mail datasets, one collected from a very large sample population denoted by $U$ and the other is confiscated from the potential suspect $S$. After the necessary preprocessing step (cleaning, tokenization, stemming, and normalization, etc.), each e-mail is converted into a vector of stylistics or stylometric features (discussed in Section 2.2). We apply classification and regression techniques on both the datasets. In each thread of techniques, the datasets are further divided into two subsets, the training and the testing sets. Two different models, one each for suspect $S$ called hypothesized author and the alternate hypothesis, are trained and validated.

Next, the given anonymous e-mail is evaluated using the two models in both regression as well as in classification thread. Unlike the usual classification where the decision is made solely on the basis of matching probability, here the decision to verify the author is based on the threshold defined for the hypothesis testing. The threshold is calculated by varying the relative number of false positives and false negatives, depending upon the nature of the perceived application of the system. The accuracy of the system is judged in terms of EER, represented by the DET curve, and the minDCF, as using only EER can be misleading [13].

The *DCF*, defined in Equation 3 of Section 3, is the weighted sum of miss and false alarm probabilities [16]. The minDCF means the minimum value of Equation 3. The DET curve is used to represent the number of false positives versus false negatives. The point on the DET curve where the number of both the false alarms become equal is called EER. The closer the DET curve to the origin, the minimum EER is and thus the better the system is. We used different classification and regression methods and were able to achieve an equal error rate of 17 percent and minDCF equal to 0.0671 with the SVM-RBF (support vector machine-radial basis function).

1. *Adopting NIST speaker recognition framework:* we are the first to have successfully adopted the NIST's SRE framework for addressing the issue of authorship verification of textual contents including e-mail dataset.

2. *Employing regression for binary classification:* regression functions, which are normally used for predicting numeric attributes (class labels), is employed for taking binary decision about whether a suspect is or is not the author of a disputed anonymous document. It is evident from the experimental results that SVM with RBF kernel produced the best verification accuracy with the lowest minDCF value as compared to the classifiers used.

3. *Error detection measures:* to measure the performance of most detection tasks, traditionally ROC curve where false alarms are plotted against the correct detection rate, is used. In this approach it is hard to determine the relative ratio of both types of errors, which is crucial in criminal investigation. The DET curve employed in this paper can better analyze the exact contribution of both the false positive and false negative values. The use of EER is augmented with minDCF in gauging the framework accuracy.

4. *Generic application:* our experiments on the real-life data, the Enron e-mail corpus, suggest that the proposed approach produces more trusted results.

The rest of the paper is organized as follow: Section 2 studies the state-of-the-art of stylometric features and the existing authorship analytical techniques. Section 3 defines the problem statement and different evaluation metrics. Section 4 presents our proposed method. Section 5 shows the experimental results on real-life e-mail data. Section 6 concludes the paper with future directions.

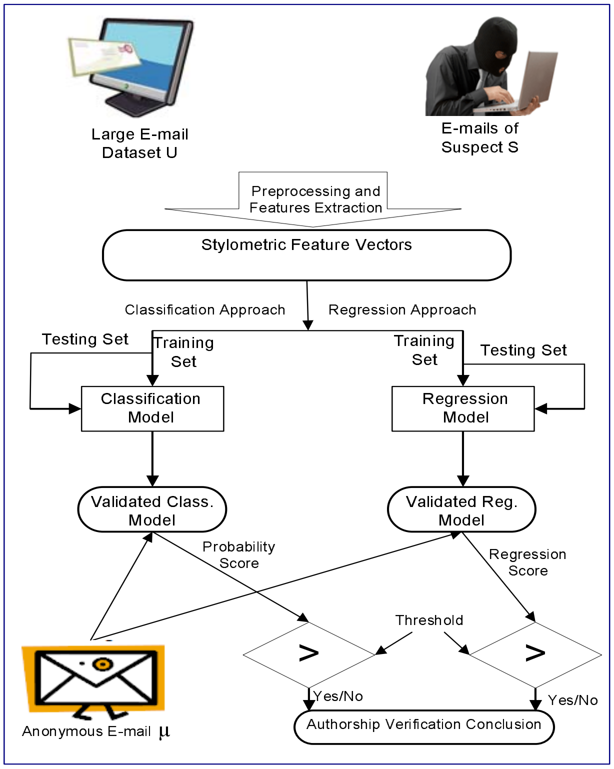

**Figure 1: E-mail Author Verification Approach**

## 2. RELATED WORK

Authorship analysis, in general, is the statistical study of linguistic and computational characteristics of written examples of individuals [7]. The writing styles or stylometric features are extracted from the sample data of potential suspects and are used to build a representative model for identifying the true author of anonymous documents or differentiating one author from another. Human stylistics, which is believed to be relatively consistent throughout the writings of an individual, can be employed to discriminate one individual from another.

Authorship is studied from three aspects. First, *authorship identification* is used to determine the likelihood that a particular suspect $S_i \in \{S_1, \ldots, S_n\}$ is the most perceivable author of a disputed document. Here, the true author is assumed to be among the potential suspects which is not true in most real world scenarios. Second, *authorship verification* is applied to confirm whether a suspect $S$ is or is not the author of a document in question. Third, *authorship characterization or profiling* is used to collect clues (such as gender information, language background and age, etc.) about the author of an anonymous document. Profiling is employed in situations where no training set of the potential suspects is available for analysis. Our focus in this paper is on authorship verification of anonymous e-mails.

Our objective in this section is to discuss existing studies both in terms of (1) stylometric features, and (2) analytical techniques proposed in the area of e-mail authorship verification.

Prior to elaborating on the existing contributions we prefer to discuss the characteristics of e-mail dataset as compared to the traditional documents.

## 2.1 E-mail Characteristics

Traditional documents are large in size, usually several hundreds of pages, well-structured in composition and usually written in a more formal fashion. They follow well-defined syntactic and grammatical rules. Moreover, the availability of numerous number of natural language processing tools and techniques make it easy to improve the quality of these documents by removing spelling and idiosyncratic mistakes. Therefore, the known written works are rich sources to learn about the writing styles of its' writers. The study of stylometric features have been very successful in resolving ownership disputes over literary and conventional writings since very long [17].

E-mail dataset like other CMC documents (such as chat logs, online messages, forums and newsgroups, etc.) pose special challenges due to their special characteristics of size and composition, as compared to literary works [7]. E-mails are short in size varying from a few words to a few paragraphs and often do not follow definite syntactic and/or grammatical rules. Therefore, it is hard to learn about the writing habits of people from their e-mail documents. Ledger and Merriam [14], for instance, established that authorship analysis results would not be significant for texts containing less than 500 words. Moreover, e-mails are more interactive and informal in style. People may not pay attention to their spelling and grammatical mistakes. Therefore, the analytical techniques that are successful in addressing the authorship issues over literary and historic works may not produce trustable results in the context of e-mail document analysis.

E-mail datasets do have certain properties that help researchers in comparing the writing style of individuals. One can find more e-mail documents for analysis as every e-mail user on average writes (say) 6-10 e-mails per day. Similarly, additional information contained in the header, subject and/or attachment(s), and the relative response time of a user, are very helpful in learning about the writing styles of a user. Moreover, e-mails are rich in structural features (such as greetings, general layout, and the contact information about the sender), which are powerful discriminators of writing styles [7].

## 2.2 Stylometric Features

Writing styles are defined in terms of stylometric features. Though, there is no such features set that is optimized and is applicable equally in all domains. However, there are more than 1000 stylometric features comprising of lexical, syntactic, structural, content-specific, and idiosyncratic characteristics that are evaluated and compared in various studies [27][1][11] in authorship studies. A brief description of the relative discriminating capabilities of each five different types of stylometric features are given below.

*Token-based features:* are collected either in terms of characters or words. In terms of characters, for instance, frequency of letters, frequency of capital letters, total number of characters per token and character count per sentence are the most relevant metrics. These indicate the preference of an individual for certain special characters or symbols or the preferred choice of using certain units. Word-based lexical features may include word length distribution, average number of words per sentence, and vocabulary richness. Initially, researchers thought that vocabulary richness [24][25] and word usage [10] are the kind of features that can discriminate the writing patterns of different people.

*Syntactic features:* Baayen et al. [2] were the first who discovered that punctuation and function words are context-independent and thus can be applied to identify writers based on their written work. Furthered, the list of function words such as 'upon', 'who', and 'above' has been extended to more than 300 by Tweedie et al. [20].

*Structural features:* are used to measure the over all appearance and layout of the documents. For instance, average paragraph length, number of paragraphs per document, presence of greetings and their position within an e-mail, are the common structural features. Moreover, the presence of a sender signature including his contact information, is one of the special structural features of e-mail documents.

*Content-specific features:* are collections of certain keywords commonly found in a specific domain and may vary from context to context even for the same author. Zheng et al. [27][26] used around 11 keywords (such as 'obo' and 'sexy', etc.) from the cyber crime taxonomy in authorship analysis experimentations.

*Idiosyncratic features:* include common spelling mistakes such as transcribing 'f' instead of 'ph' (such as in the word phishing) and grammatical mistakes such as sentences containing incorrect form of verbs. The list of such characteristics varies from person to person and is difficult to control.

## 2.3 Computational Methods

The analytical authorship techniques employed so far include univariate and multivariate statistics [4][8], machine learning processes such as support vector machine and decision trees [6][26] and frequent pattern mining [11]. However, there is still a long way to develop consensus about the features set and the techniques that can be trusted to the degree to present it in the court of law for fixing responsibility in authorship attribution disputes.

Unlike authorship attribution and authorship characterization where the problem is clearly defined, there is no consensus on how to precisely define the problem in the authorship verification studies. Some researchers consider it as a 'similarity detection' task, which states that given two pieces of text, the problem is to determine whether they are produced by the same entity or not, without knowing the actual author. Vel et al. [6], and Abbasi and Chen [1] have applied SVM and KL transformation techniques, respectively for authorship attribution and similarity detection. Following the same notion of verification, Halteren [21] has proposed a relatively different approach called linguistic profiling. He has proposed some distance and scoring functions for creating profiles for a group of example data. The average feature counts for each author is compared with a general stylistic profile, built from the training samples of widely selected authors. The study focus on students' essays, and not investigate into the CMC documents dataset.

There is another group of researchers including Larry et al. [15] and Koppel et al. [12] who look at the authorship verification one-class and two-class text classification problem. For instance, Larry et al. [15] investigated the problem as: given a disputed document $d$ together with the known training examples $\{t_1, \ldots, t_n\}$ of a suspect $S$, the task is to verify whether document $d$ is written by suspect $S$ or not. Documents written by other authors are labeled as 'outlier' in that study. A slightly modified version of one-class approach called 'imposter' is the two-class problem is proposed by Koppel et al. [12]. According to this study, the known works of the potential suspect $S$ are labeled as 'S' and that of other authors as 'imposter'. A classification model is developed one each for 'S' and 'imposter' documents. The anonymous document $d$ is divided into different chunks and each chunk is given to the built model to find its class. However, the method fails to discriminate if the imposter documents are closely similar to that of the suspect documents.

A fairly opposite approach would be to train one model for $S$ and not-S and determine the degree of distinctness of the two by employing a 10-fold cross validation approach [13]. If the validation accuracy is high it is concluded that $S$ did not write $d$ otherwise the model fails to assign plausible class label.

A relatively new approach called 'unmasking', proposed in [13], is the extension of the 'imposter' method. In this study the authors attempt to quantify the dissimilarity between the documents of the suspect and that of the 'imposter'. The experimental results reported by Koppel et al. [13] indicate that the method is suitable in situations where the document in question is at least 5000 words long for achieving trustable results. This is nearly impossible in case of e-mail documents.

## 3. PROBLEM STATEMENT AND EVALUATION METRICS

Given a set of example e-mails of a potential suspect $S$ and an e-mail dataset $U$ collected from a very large population of authors, the task of an expert witness or an investigator is to verify whether or not the disputed anonymous e-mail $\mu$ is written by the suspect $S$.

Formally, the problem can be defined as: given an anonymous e-mail $\mu$, and a hypothesized author or suspect $S$, authorship verification can be termed as a basic hypothesis test between

$H_0$: $\mu$ is written by the hypothesized author $S$
and
$H_1$: $\mu$ is not written by the hypothesized author $S$.

The optimum test to decide between these two hypotheses is a likelihood ratio test given by

$$\frac{p(\mu|H_0)}{p(\mu|H_1)} \geq \theta \tag{1}$$

accept $H_0$, otherwise reject $H_0$ (accept $H_1$) where $p(\mu|H_i)$, $i = 0, 1$ is the probability density function for the hypothesis $H_i$ evaluated for the observed e-mail $\mu$ and $\theta$ is the decision threshold for accepting or rejecting $H_0$. The basic goal is to find techniques for calculating the two likelihood functions $p(\mu|H_0)$ and $p(\mu|H_1)$.

The author-specific model $H_0$ is well-defined and is built using e-mails written by the hypothesized author while the model $H_1$ is not well-defined as (potentially) it must represent the entire space of the possible alternatives to the hypothesized author.

In order to define $H_1$ model, we borrow the techniques used in the speaker verification literature. Two main approaches have been in use for the alternative hypothesis modeling in the speaker recognition research. The first approach is to use a set of *other-author* models to cover the space of the alternative hypothesis. This set of authors is called the cohort or the background authors. Given a set of N background author models $\lambda_1, \lambda_2, \cdots \lambda_N$, the alternative hypothesis model is represented by

$$p(\mu|H_1) = f(p(\mu|\lambda_1), p(\mu|\lambda_2), \cdots, p(\mu|\lambda_N)) \tag{2}$$

where $f(.)$ is some function, such as average or maximum, of the likelihood values from the background author set. The selection, size and combination of the background authors can be the subject of further research.

Another approach is the alternative hypothesis modeling in which a model is developed on sample documents are collected from a very large number of individuals. The model developed in this way is called the universal background model (UBM) in the speech processing community. We adopted the same approach for online textual documents. Given a collection of e-mail samples from a very large number of authors, a single model is trained to represent the alternative hypothesis. The main advantage of this approach is that a single author-independent model can be trained once for a particular task and then used for all hypothesized authors in that task.

Two types of errors can occur in the author verification system namely false rejection (rejecting a valid author) and false acceptance (accepting an invalid author). The probability of these errors called as miss probability or false rejection probability $P_{fr}$ and false alarm probability $P_{fa}$. Both types of error depend on the value of user defined threshold $\theta$. It is therefore possible to represent the performance of the system by plotting $P_{fa}$ versus $P_{fr}$, the curve generally known as $DET$ curve in the speech processing community.

In order to judge the performance of the author verification systems different performance measures can be used. We borrow the two main measures namely Equal Error Rate ($EER$) and Detection Cost Function ($DCF$) from the speech processing community. The $EER$ corresponds to the point on the DET curve where $P_{fa} = P_{fr}$. Since using only EER can be misleading [13], we used the DCF in conjunction with EER to judge the performance of author verification system. The $DCF$ is defined in the SRE framework [16] as the weighted sum of miss and false alarm probability, as shown in the following equation.

$$DCF = C_{fr} \times P_{fr} \times P_{target} + C_{fa} \times P_{fa} \times (1 - P_{target}) \quad (3)$$

The parameters of the cost function are the relative costs of detection errors, $C_{fr}$ and $C_{fa}$ and the *a priori* probability of the specified target author, $P_{target}$. In our experiments, we used the parameter values as specified in the NIST's SRE framework. These values are $C_{fr} = 10$, $C_{fa} = 1$ and $P_{target} = 0.01$.

The minimum cost detection function (mDCF) is redefined as the minimum value of '$0.1\times$*false rejection rate* + $0.99\times$*false acceptance rate*'. Since it is primarily dependent on the false acceptance rate and false rejection rate and has nothing to do specifically with the speech, it can be used for the authorship verification as well. It is in conformance with the forensic analysis and strictly punishes the false acceptance rate as it would implicate an innocent person as the perpetrator.

# 4. OUR METHOD

In this paper, we have addressed the authorship verification as a two-class classification problem by building two models one from e-mails of the potential suspect and the other from a very large e-mail dataset belonging to different individuals called universal background model. How to train and validate the two representative models, we borrowed the techniques from the SRE framework [16]. The framework is initiated by the National Institute of Standards and Technology. The purpose of the SRE framework is not only to develop state-of-the-art frameworks for addressing the issues of speaker identification and verification but to standardize and specify a common evaluation platform for judging the performance of these systems as well.

Next, the evaluation measures such as DCF, minDCF, and EER that are used in the SRE framework are more tailored to forensic analysis as compared to simple ROC and classification accuracies, etc.

Another reason for borrowing ideas from the speaker recognition community is that this area has a long and rich scientific basis with more than 30 years of research, development and evaluation [16]. The objective of both authorship and speaker verification is the same i.e. to find whether a particular unknown object is produced by a particular subject or not. The object in our case is the anonymous e-mail whereas in case of speaker verification it is the speech segment. The subject is the speaker in their case whereas it is the author in our case.

As depicted in Fig. 1, the proposed method starts with the features extraction followed by model development and matching the disputed e-mail with the models to verify its true author.

## 4.1 Features Extraction

As described in Section 2.2, there are more than 1000 stylometric features used so far in different studies [5][1]. As listed in the Appendix, we have carefully selected 292 features in our study. In general, there are three types of features. The first type is a numerical value, e.g., the frequencies of some individual characters, punctuation and special characters. To avoid the situation where very large values overweigh other features, we applied normalization to scale down all the numerical values to $[0, 1]$.

The second type is a boolean value, e.g., to check whether an e-mail contains a reply message or not or does it has an attachment or not? The third type of features are computed by taking as input some other lexical functions such as vocabulary richness, indexed at 94-105 in the Appendix. Most of these features are computed in terms of vocabulary size V(N) and text length N [20]. Once feature extraction is done, each e-mail is represented as a vector of feature values. In this study we focused more on using structural features as they play a significant role in distinguishing writing styles.

Short words usually comprising 1-3 characters (such as 'is', 'are', 'or', 'and', etc.) are mostly context-independent and are counted together. Frequencies of words of various lengths 1-20 characters (indexed at 59-88) are counted separately. Hepax Legomena and Hapax dislegomena are the terms used for once-occurring and twice-occurring words. As mentioned earlier, we have used more than 150 function words (115-264). We also check whether an e-mail has welcoming and/or farewell greetings. Paragraph separator can be a blank line or just a tab/indentation or there may be no separator between paragraphs.

Thirteen content-specific terms (280-292) were selected from the Enron e-mail corpus [1] by applying content-based clustering.

## 4.2 Modeling and Classification

The decision whether a given anonymous e-mail $\mu$ belongs to the hypothesized author (or suspect $S$) or not, is taken on the basis of the scores produced by e-mail $\mu$ during the classification process and the threshold $\theta$. The threshold is defined by the user and is employed for taking binary decision. As described in the following paragraphs, we used two approaches for binary classification of e-mails.

### 4.2.1 Verification by Classification

In this approach the e-mails in the training set corresponding to the hypothesized author, and that belonging to the sample population, are nominally labeled. During the testing phase, a score is assigned to each e-mail on the basis of the probability assigned to the e-mail by the classifier. The scores calculated for the true author and the 'imposters' are evaluated for the false acceptance and false rejection rates through a DET plot.

We used three different classification techniques, namely Adaboost.M1 [23], Discriminative Multinomial Naive Bayes (DMNB) [19] and Bayesian Network [9] classifiers. Most of the commonly used classification techniques including the one employed in the current study are implemented in the Weka toolkit [22].

### 4.2.2 Verification by Regression

Though, authorship verification is conceptually a classification problem but in our case we need to take a binary decision of whether the e-mail under test belongs to the potential suspect or not. Similarly, as the decision is taken on the basis of the similarity score assigned to the e-mail under test, we employ regression functions to calculate the score. We used three different regression techniques including linear regression [22], SVM with Sequential Minimum Optimization (SMO) [18], and SVM with RBF kernel [3]. We used regression scores of the true authors and the impostors

[1] `http://www-2.cs.cmu.edu/~enron/`

for calculating equal error rate and minimum detection cost function.

We assign scores to e-mails of the true author and those belong to the 'imposters'. In this case we assign two numerical values to each e-mail in the training set. For instance, +10 is assigned to the hypothesized author's e-mails and −10 to e-mails of the target population. When applied, the regression function assigns a value generally between +10 and −10 to the disputed anonymous e-mail. The decision whether it belongs to the hypothesized author or not, is based on the resultant score and the user defined threshold $\theta$.

Setting the threshold too low will increase the false alarm probability whereas setting it too high will have high miss probability (false rejection rate). In order to decide about the optimum value of the threshold and to judge the performance of our verification system, we plot the variation of the false alarm rate with the false rejection rate. The curve is generally known as the detection error trade off curve, which is drawn on a deviate scale [16]. The closer the curve to the origin, the better the verification system is. The point on the curve where the false alarm rate equals the false rejection rate is called the equal error rate.

## 5. EXPERIMENTAL EVALUATION

To evaluate our implementation, we performed experiments on the Enron e-mail corpus made available by MIT. First, we created a universal background model from the entire Enron e-mail corpus. This is an author-independent model and is used as the basis for taking the decision whether or not the e-mail in question belongs to the suspected author. A separate model is created for each author. For this, we used 200 e-mails per author.

The decision whether an e-mail under test belongs to the hypothesized author or not is based on the difference of similarity of the e-mail to the author-independent model and that to the hypothesized author model. Based on this similarity metric, a score is assigned to the disputed e-mail. For evaluation of our classification methods, we employed the widely used 10-fold cross validation approach by reserving 90% for training and 10% for testing. The reason is to avoid any biaseness in during the evaluation process and to judge the classification method over the entire database.

One of the performance measure used in the SRE framework is to calculate the equal error rate [16]. The EER is calculated by taking two types of scores as input namely the true author score and the false author score, which in turn are calculated by the classification methods applied over the test dataset.

### *Verification by Classification.*

As depicted in Fig. 2, a DET plot of one author, randomly selected from our database, using the classification techniques. Usually, the closer the DET curve to the origin, the minimum the EER is and thus the better the system is. The point on the DET plot which gives the minimum cost detection function is marked as a small circle on each curve.

The DET curve plotted for Bayesian Network (Bayes Net) is more consistent and indicates better results both in terms of equal error rate and minimum cost detection function with less complexity. The value of minDCF for both DMNB and AdaBoost is comparable, however, performance of DMNB in terms of EER is closed to Bayes Net. The performance gap between the two classifiers is consistent throughout the experimentation results.

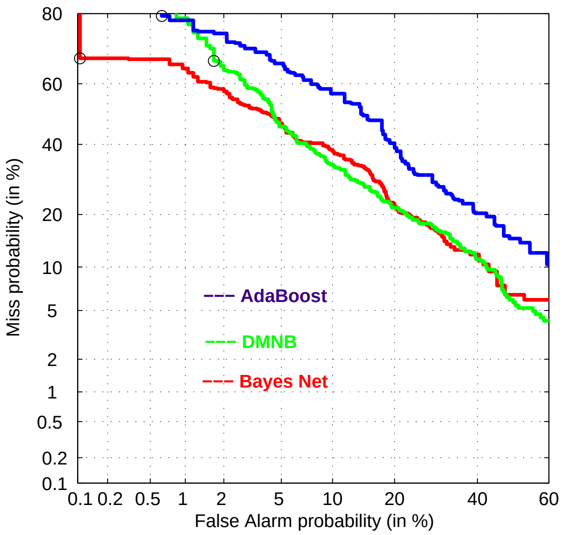

**Figure 2: A typical DET curve for author verification with different classification techniques**

### *Verification by Regression.*

Fig. 3 shows the typical DET plot of one of the randomly selected author from our database, constructed by using the scores obtained from the three regression techniques as described above. The DET curve indicates that the regression approach usually produce better results in terms of EER and minDCF as compared to the classification approach. The regression approach via SVM with RBF kernel with EER 17.1% outperformed linear regression (with EER = 19.3%) and SVM-SMO (with EER = 22.3%). The same tendency of performance can be seen in minDCF vales as well (see the last row of Table 1). DET curves for linear regression and SVM-SMO are running neck to neck starting with a highest value of false negative.

The bottom line is that SVM with RBF kernel produced the best verification accuracy with the lowest minDCF value. These results suggest that regression techniques are more suitable in addressing verification problem than classifiers which perform better in attribution issues. However, the same assumption may not be always true depending on the dataset as well as features set used.

Table 1 shows the mean EER and minimum DCF obtained when using the above discussed classification and regression methods.

## 6. CONCLUSION

We studied the problem of e-mail authorship verification and presented a solution by adopting the NIST speaker verification framework and the accuracy measuring methods. The problem is addressed as a two-class classification problem by building two models one from e-mails of the potential suspect and the other from a very large e-mail dataset belonging to different individuals called universal background model. Experiments on a real-life dataset produces an equal

**Table 1: Author Verification Accuracies of Various Classification and Regression Methods**

| | *Classification* | | | *Regression* | | |
|---|---|---|---|---|---|---|
| **Verification** | A.Boost | DMNB | Bayes | SVM-SMO | Lin. Reg | SVM-RBF |
| EER(%) | 22.4 | 20.1 | 19.4 | 22.3 | 19.3 | 17.1 |
| minDCF | 0.0836 | 0.0858 | 0.0693 | 0.0921 | 0.0840 | 0.0671 |

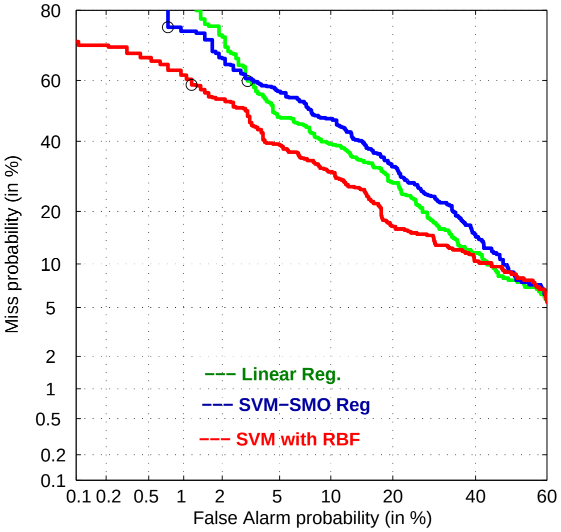

**Figure 3: A typical DET curve for author verification with different regression techniques**

error rate of 17% by employing support vector machines with RBF kernel, a regression function. The results are comparable with other state-of-the-art verification methods. Building a true 'universal' background model is not an easy task due to the non-availability of insufficient sample e-mails. The style variation of the same suspect with the changing state of mind and the context in which he writes may affect his representative model. The framework originally designed for a different kind of data (speech examples) need to be further tuned for to achieve better accuracy for textual online documents.

## 7. REFERENCES

[1] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2), March 2008.

[2] R.H. Baayen, H. Van Halteren, and F.J. Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 2:110–120, 1996.

[3] C.J.C Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[4] J.F. Burrows. Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2:61–67, 1987.

[5] M. Corney, O. Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. In *proc. 18th Annual Computer Security Applications Conference. 2002*, pages 21–27, 2002.

[6] O. de Vel. Mining e-mail authorship. In *proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2000.

[7] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD RECORD*, 30(4):55–64, December 2001.

[8] R.S. Forsyth and D.I. Holmes. Feature finding for text classification. *Literary and Linguistic Computing*, 11(4):163–174, 1996.

[9] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, 29, 1977.

[10] D.I. Holmes. The evolution of stylometry in humanities. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[11] F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5:42–51, 2008.

[12] M. Koppel, S. Argamon, and A.R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[13] M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):9–26, 2009.

[14] G.R. Ledger and T.V.N. Merriam. Shakespeare, fletcher, and the two noble kinsmen. *Literary and Linguistic Computing*, 9:235–248, 1994.

[15] L. M. Manevitz, M. Yousef, N. Cristianini, J. Shawe-taylor, and B. Williamson. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154, 2001.

[16] A. Martin and M. Przybocki. The NIST Speaker Recognition Evaluation Series. National Institute of Standards and Technology Web site, June 2009.

[17] T.C. Mendenhall. The characteristic curves of composition. *Science*, 11(11):237–249, 1887.

[18] J.C. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. pages 185–208, 1999.

[19] J. Su, H. Zhang, C.X. Ling, and S. Matwin. Discriminative Parameter Learning for Bayesian Networks. In *International Conference on Machine Learning*, pages 1016–1023, 2008.

[20] F.J. Tweedie and R.H. Baayen. How variable may a constant be? measures of lexical richness in perspective. *Literary and Linguistic Computing*, 32:323–352, 1998.

[21] H. Van Halteren. Author verification by linguistic

profiling: An exploration of the parameter space. *ACM Trans. Speech Lang. Process.*, 4(1):1, 2007.

[22] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Margan Kaufmann, San Francisco, 2nd edition, 2005.

[23] F. Yoav and R.E. Schapire. Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[24] G.U. Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika Trust*, 30:363–390, 1939.

[25] G.U. Yule. The statistical study of literary vocabulary. *The Modern Language Review*, 39(3):291–293, 1944.

[26] 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*, 57(3):378–393, February 2006.

[27] R. Zheng, Y. Qin, Z. Huang, and H. Chen. Authorship analysis in cybercrime investigation. In *proc. 1st NSF/NIJ Symposium, ISI Springer-Verlag.*, pages 59–73, 2003.

## 8. APPENDIX

**Lexical Features: Character-based**

1 Character count excluding space characters (M)

2 Ratio of digits to M

3 Ratio of letters to M

4 Ratio of uppercase letters to M

5 Ratio of spaces to M

6 Number of white-space characters/M

7 Number of spaces/Number white-space chars

8 Ratio of tabs to M

9-34 Alphabets frequency (A-Z) (26 features)

35-54 Occurrences of special characters: < > % | { } [ ] / \ @ # ˜ + - * $ ˆ & _ ÷ (21 features)

**Lexical Features: Word-based**

55 Word count (W)

56 Average word length

57 Average sentence length in terms of characters

58 Ratio of short words (1-3 characters) to W

59-88 Ratio of word length frequency distribution to W (30 features)

89 Ratio of function words to W

90 Vocabulary richness i.e. T/W

91 Ratio of Hapax legomena to M

92 Ratio of Hapax legomena to T

93 Ratio of Hapax dislegomena to M

94 Guirad's R

95 Herdan's C

96 Herdan's V

97 Rubet's K

98 Maas' A

99 Dugast's U

100 Lukjanenkov and Neistoj's measure

101 Brunet's W

102 Honore's H

103 Sichel's S

104 Yule's K

105 Simpson's D

**Syntactic Features**

106-113 Occurrences of punctuations , . ? ! : ; ’ ” (8 features)

114 Ratio of punctuations with M

115-264 Occurrences of function words (150 features)

**Structural Features**

265 Ratio of blank lines/total number of lines within e-mail

266 Sentence count

267 Paragraph count

268 Presence/absence of greetings

269 Has tab as separators between paragraphs

270 Has blank line between paragraphs

271 Presence/absence of separator between paragraphs

272 Average paragraph length in terms of characters

273 Average paragraph length in terms of words

274 Average paragraph length in terms of sentences

275 Contains Replied message?

276 Position of replied message in the e-mail

277 Use e-mail as signature

278 Use telephone as signature

279 Use URL as signature

**Domain-specific Features**

280-292 *deal, HP, sale, payment, check, windows, software, offer, microsoft, meeting, conference, room, report* (13 features)