Mining E-mail Content for Author Identification Forensics

O. de Vel
Information Technology Division
Defence Science and Technology Organisation
P.O. Box 1500
Salisbury 5108, Australia
olivier.devel@dsto.defence.gov.au

A. Anderson, M. Corney and G. Mohay
School of Information Systems
Faculty of Information Technology
Queensland University of Technology
Brisbane Q4001, Australia
a.anderson|m.corney|g.mohay@qut.edu.au

ABSTRACT

We describe an investigation into e-mail content mining for author identification, or authorship attribution, for the purpose of forensic investigation. We focus our discussion on the ability to discriminate between authors for the case of both aggregated e-mail topics as well as across different e-mail topics. An extended set of e-mail document features including structural characteristics and linguistic patterns were derived and, together with a Support Vector Machine learning algorithm, were used for mining the e-mail content. Experiments using a number of e-mail documents generated by different authors on a set of topics gave promising results for both aggregated and multi-topic author categorisation.

1. INTRODUCTION

Computer forensics undertakes the post-mortem reconstruction of the (causal) sequence of events arising from an intrusion perpetrated by an external agent or as a result of unauthorised activities generated by an authorised user. The field of computer forensics covers a broad set of applications, uses a variety of evidence and is supported by a number of different techniques. Application areas include forensic accounting, law enforcement, commodity flow analysis, threat analysis etc... Evidence (data) made available to computer forensics investigators is varied and can be sourced from, for example, storage devices (disks, discs etc...), networks (e.g., packet data, routing tables, logs), telecommunication traffic. Computer forensics investigations can involve a wide variety of techniques including information hiding analysis, data mining, link and causal analysis, timeline correlation and so on.

Of particular interest in this paper is e-mail data that are now becoming the dominant form of inter- and intra-organisational written communication for many companies and government departments. E-mail is used in many legitimate activities such as message and document exchange. Unfortunately, it can also be misused for the distribution of unsolicited and/or inappropriate messages and documents. Examples of misuse include the distribution of unsolicited junk mail, unauthorised conveyancing of sensitive information, mailing of offensive or threatening material. E-mail evidence can be central in cases of sexual harassment or racial vilification, threats, bullying and so on. In some misuse cases the sender will attempt to hide his/her true identity in order to avoid detection. For example, the sender's address can be spoofed or anonymised by routing the e-mail through an anonymous mail server, or the e-mail's contents and header information may have been modified in an attempt to hide the true identity of the sender. In other cases the sender may wish to masquerade as another user. The ability to provide empirical evidence and identify the original author of e-mail misuse is an important, though not necessarily unique, factor in the successful prosecution of an offending user.

In the context of computer forensics, the mining of e-mail authorship has a couple of unique characteristics. Firstly, the identification of an author is usually attempted from a small set of known candidates, rather than from a large set of potentially unknown authors. Secondly, the text body of the e-mail is not the only source of authorship attribution. Other evidence in the form of e-mail headers, e-mail trace route, e-mail attachments, file time stamps, etc. can, and should, be used in conjunction with the analysis of the e-mail text body. In this paper we focus uniquely on the data mining phase of the computer forensics procedure.

As a result of this growing e-mail misuse problem, efficient automated methods for analysing the content of e-mail messages and identifying or categorising the authors of these messages are becoming imperative. The principal objectives are to classify an ensemble of e-mails as belonging to a particular author and, if possible, obtain a set of characteristics that remain relatively constant for a large number of e-mails written by the author. The question then arises; can characteristics such as language, layout etc.. of an e-mail be used, with a high degree of confidence, as a kind of author phrenology and thus link the e-mail document with its author? Also, can we expect the writing characteristics or style of an author to evolve in time and change in different contexts? For example, the composition of formal e-mails will differ from informal ones (changes in vocabulary etc.). Even in the context of informal e-mails there could be several composition styles (e.g., one style for personal relations and one for work relations). However, humans are creatures of habit

and have certain personal traits which tend to persist. All humans have unique (or near-unique) patterns of behaviour, biometric attributes, and so on. We therefore conjecture that certain characteristics pertaining to language, composition and writing, such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage (e.g., converting the letter "f" to "ph", or the excessive use of digits and/or upper-case letters), stylistic and sub-stylistic features will remain relatively constant. The identification and learning of these characteristics with a sufficiently high accuracy are the principal challenges in authorship categorisation.

Authorship categorisation or attribution can be effected using various approaches. Firstly, the simplest method is to use domain experts to identify new e-mail documents and allocate them to well-defined author categories. This can be time-consuming and expensive and, perhaps most limiting, provides no continuous measure of the degree of confidence with which the allocation was made. Secondly, the domain expert can establish a set of fixed rules which can be used to classify new e-mail documents. Unfortunately, in many cases, the rule-set can be large and unwieldy, typically difficult to update, and unable to adapt to changes in document content or author characteristics. Finally, categorisation can be undertaken automatically by inductively learning the classifiers from training example documents. This approach should, hopefully, generalise well to new, unseen e-mail documents and has the advantage that it should be able to adapt to a measure of drift in the characteristics of authors and create a more accurate profile of each author.

A closely related, but clearly separate, area of authorship categorisation is text categorisation, which attempts to categorise a set of text documents based on its contents or topic. Text categorisation provides support for a wide variety of activities in information mining and information management. It has found applications in document filtering and can be used to support document retrieval by generating the categories required in document retrieval. Many methods that automatically learn rules have been proposed for text categorisation. Most of these techniques employ the "bag–of–words" or word vector space feature representation [30] where each word in the text document corresponds to a single feature. A learning algorithm such as decision trees [3], neural networks [25], Bayesian probabilistic approaches [23][42], or support vector machines [18] is then used to classify the text document. de Vel [9] studied the comparative performance of text document categorisation algorithms using the Naive Bayes, Support Vector Machines, multi-layer Perceptron and k–NN classifiers. Work in e-mail text classification has also been undertaken by some researchers in the context of automated e-mail document filtering and filing. Cohen [7] learned rule sets based on a small number of keywords in the e-mail. Sahami *et al* [28] focused on the more specific problem of filtering junk e-mail using a Naive Bayesian classifier and incorporating domain knowledge using manually constructed domain-specific attributes such as phrasal features and various non-textual features.

In this paper we investigate methods for the multi-topic machine learning of an authorship attribution classifier using e-mail documents as the data set. We focus on the problem of authorship attribution of e-mails and not e-mail document categorisation, i.e. not the classification of e-mail messages for topic categorisation etc. . . We incorporate various document features such as structural characteristics and linguistic evidence in the learning algorithm. We study the effect of both aggregated and multiple e-mail topics on the discrimination performance of authorship attribution. For example, can an author be identified in the context of different e-mail topics? That is, we wish to investigate the degree of orthogonality existing between e-mail authorship and e-mail topic content. We first introduce the field of authorship categorisation in Section 2 and, more specifically, e-mail authorship categorisation in Section 3. We then briefly outline the Support Vector Machines learning algorithm in Section 4 and present the database of e-mail documents used in the experiments together with the experimental methodology in Sections 5 and 6, respectively. Validation of the method is then undertaken by presenting results of categorisation performance in Section 7. Finally, we conclude with some general observations and present future directions for the work in Section 8.

2. AUTHORSHIP CATEGORISATION

Formally, authorship categorisation is the task of determining the author of a piece of work. In particular, we are interested in categorising textual work given other text samples produced by the same author. We assume that only one author is responsible for producing the text – contributions by, or text modified by, multiple authors are not considered here (though, as we describe later on, e-mails with labeled text from other authors are included in our analysis).

Authorship categorisation is a subset of the more general problem called "authorship analysis" [16]. Authorship analysis includes other distinct fields such as *author characterisation* and *similarity detection*. Authorship characterisation determines the author profile or characteristics of the author that produced a piece of work. Example characteristics include gender, educational and cultural backgrounds, language familiarity etc. [34]. Similarity detection calculates the degree of similarity between two or more pieces of work without necessarily identifying the authors. Similarity is used extensively in the context of plagiarism detection which involves the complete or partial replication of a piece of work with or without permission of the original author. We note, however, that authorship categorisation and author characterisation are different from plagiarism detection. Plagiarism detection attempts to detect the similarity between two substantially different pieces of work but is unable to determine if they were produced by the same author.

Authorship analysis has been used in a small but diverse number of application areas. Examples include identifying authors in literature, in program code, and in forensic analysis for criminal cases. We briefly outline the work undertaken in each one of these areas.

Perhaps the most extensive and comprehensive application of authorship analysis is in literature and in published articles. Well-known authorship analysis studies include the disputed Federalist papers (e.g., [24] [4]) and Shakespeare's works, the latter dating back over many years (see, for example, where attempts were made to show that Shakespeare

was a hoax and that the real author was Edward de Vere, the Earl of Oxford [13]). In these studies, specific author features such as unusual diction, frequency of certain words, choice of rhymes, and habits of hyphenation have been used as tests for author attribution. These authorial features are examples of *stylistic evidence* which is thought to be useful in establishing the authorship of a text document. It is conjectured that a given author's style is comprised of a number of distinctive features or attributes sufficient to uniquely identify the author. Stylometric features ("style markers") used in early authorship attribution studies were character or word based, such as vocabulary richness metrics (e.g., Zipf's word frequency distribution and its variants), word length etc.. However, some of these stylometric features could be generated under the conscious control of the author and, consequently, may be content-dependent and are a function of the document topic, genre, epoch etc.. Rather than using content-dependent features, we employ features derived from words and/or syntactic patterns since such features are more likely to be content-independent and thus potentially more useful in discriminating authors in different contexts. It is thought that syntactic structure is generated dynamically and sub-consciously when language is created, similar to the case of the generation of utterances during speech composition and production [8]. That is, language patterns or syntactic features are generated beyond an author's conscious control. An example of such features is short, all-purpose words (referred to as *function words*) such as "the", "if", "to" etc. whose frequency or relative frequency of usage is unaffected by the subject matter. Another example syntactic feature is punctuation which is thought to be the graphical correlate of intonation which is the phonetic correlate of syntactic structure [5]. As punctuation is not guided by any strict placement rules (e.g., comment placement), punctuation will vary from author to author. Chaski [6] has shown that punctuation can be useful in discriminating authors. Therefore, a combination of syntactic features may be sufficient to uniquely identify an author. According to Rudman, over 1,000 stylometric features have been proposed [27]. Tweedie *et al* also list a variety of different stylometric features [35]. However, no set of significant style markers have been identified as uniquely discriminatory. Furthermore, some proposed features may not be valid discriminators as, for example, prescriptive grammar errors, profanities etc. which are not generally considered to be idiosyncratic. Just as there is a range of available stylometric features, there are many different techniques using these features for author identification. These include statistical approaches (e.g., cusum [14], Thisted and Efron test [33]), neural networks (e.g., radial basis functions [22], feedforward neural networks [36], cascade correlation [39]), genetic algorithms (e.g., [17]), Markov chains (e.g., [19]). However, there does not seem to exist a consensus on a correct methodology, with many of these techniques suffering from problems such as questionable analysis, inconsistencies for the same set of authors, failed replication etc.

Program code authorship has been researched by some workers in the context of software theft and plagiarism, software author tracking and intrusion detection. For example, software author tracking enables the identification of the author of a particular code fragment from a large set of programmers working on a software project. This can be useful for identifying authors for the purpose of effecting upgrades to software and software maintenance. The authorship of a computer virus or trojan horse can be identified in a similar manner [31]. By examining peculiar characteristics or metrics of programming style it is possible to identify the author of a section of program code [26], in a similar way that linguistic evidence can be used for categorising the authors of free text. Program metrics such as typographical characteristics (e.g., use of lower and upper case characters, multiplicity of program statements per line, etc.), stylistic metrics (e.g., length of variable names, preference for `while` or `for` loops, etc.), programming structure metrics (e.g., placement of comments, use of debugging symbols, etc.) have been employed [21][20][29].

The forensic analysis of text attempts to match text to authors for the purpose of a criminal investigation. The forensic analysis of text generally includes techniques derived from linguistics or behavioural profiling. Linguistic techniques usually employ common knowledge features such as grammatical errors, spelling, and stylistic deviations. These techniques, contrary to popular belief, do not quantify linguistic patterns and fail to discriminate between authors with a high degree of precision. However, the use of language-based author attribution testimony as admissible evidence in legal proceedings has been identified in many cases [5]. The textual analysis of the Unabomber manifesto is a well-known example of the use of forensic linguistics. In this case, the manifesto and the suspect bomber used a set of similar characteristics, such as a distinctive vocabulary, irregular hyphenations etc. [8][15]. Techniques based on scientific evidence of language have not, to the authors' knowledge, been used in court proceedings. Profiling is based on the behavioural characteristics contained within an author's text. For example, educated guesses on the type of personality of an author based on particular sequences of words are employed in profiling studies.

E-mail documents have several characteristics which make authorship categorisation challenging compared with longer, formal text documents such as literary works or published articles (such as the Federalist Papers). Firstly, e-mails are generally short in length indicating that certain language-based metrics may not be appropriate (e.g., vocabulary richness). Secondly, the composition style used in formulating an e-mail document is often different from normal text documents written by the same author. That is, an author profile derived from normal text documents (e.g., publications) may not necessarily be the same as that obtained from an e-mail document. For example, e-mail documents are generally brief and to the point, can involve a dialogue between two or more authors, can be punctuated with a larger number of grammatical errors etc. Also, e-mail interaction between authors can be frequent and rapid, similar to speech interactivity and rather dissimilar to normal text document interchange patterns. Indeed, the authoring composition style and interactivity characteristics attributed to e-mails shares some elements of both formal writing and speech. Thirdly, the author's composition style used in e-mails can vary depending upon the intended recipient and can evolve quite rapidly over time. Fourthly, the vocabulary used by authors in e-mails is not stable, facilitating imitation. Thus the possibility of being able to disguise authorship of an e-mail

through imitation is potentially high. Furthermore, similar vocabulary subsets (e.g., technology-based words) may be used within author communities. Finally, e-mail documents have generally few sentences/paragraphs, thus making contents profiling based on traditional text document analysis techniques, such as the "bag–of–words" representation (e.g., when using the Naive Bayes approach), more difficult. However, as stated previously, certain characteristics such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage, stylistic and substylistic features will remain relatively constant for a given e-mail author. This provides the major motivation for the particular choice of attributes/features for the authorship categorisation of e-mails, as we shall discuss in Section 6.

3. E-MAIL AUTHORSHIP CATEGORISATION

Only a small number of studies in e-mail authorship analysis have been undertaken to date. de Vel [10] has investigated e-mail authorship categorisation using a basic subset of structural and stylometric features on a set of authors without consideration of the author characteristics (gender, language, etc.) nor of the e-mail topic and size. Anderson *et al* [1] have used a larger set of stylometric features and studied the effect of a number of parameters such as, the type of feature sets, text size, and the number of documents per author, on the author categorisation performance for both e-mails and text documents. Some feature types such as N-graphs (where $N = 2$ was used) gave good categorisation results for different text chunk sizes but these results were thought to be due to an inherent bias of some types of N-graphs towards content rather than style alone (N-graphs are contiguous sequences of characters, including whitespaces, punctuation etc. . .). They observed almost no effect of the text chunk size on the categorisation performance, for text chunks larger than approximately 100 words. Also, they observed that as few as 20 documents may be sufficient for satisfactory categorisation performance. These results are significant in the context of e-mail authorship categorisation as they indicate that satisfactory results can still be achieved with a small text size and a small number of available e-mails. Although Anderson *et al* concluded that it is possible to categorise e-mail authors based on a small number of e-mails and small text sizes, they did not consider other author attribution characteristics such as multi-topic categorisation performance, nor author characteristics. A preliminary study on multi-topic e-mail attribution was presented by Anderson *et al* [2]. They used a sparse set of newsgroup e-mails as the e-mail corpus and obtained variable performance results, with some authors obtaining good categorisation results across different newsgroup topic categories but with other authors obtaining lower performance results. More recently, in the context of e-mail authorship characterisation, Thomson *et al* [34] have investigated the existence of gender-preferential language styles in e-mail communication. The types of styles investigated included references to emotion, provision of personal information, use of intensive adverbs, the frequency of insults and opinions (it was hypothesised that the first three of these features are characteristic of female authors whereas the last set of features are male-preferential). Using manual feature extraction and discriminant analysis, Thomson *et al* claimed that they were able to predict the gender of e-mail authors.

In this paper we extend the results of these investigations and study the author attribution performance in the context of multiple e-mail topic categories. We investigate the e-mail document feature types that enable us to discriminate between e-mail authors independent of e-mail topic. In particular, we extend the work of Anderson *et al* [2] and use a more complete e-mail corpus with controlled topic characteristics.

4. SUPPORT VECTOR MACHINE CLASSIFIER

The fundamental concepts of Support Vector Machines (SVM) were developed by Vapnik [38]. The SVMs' concept is based on the idea of structural risk minimisation which minimises the generalisation error (i.e. true error on unseen examples) which is bounded by the sum of the training set error and a term which depends on the Vapnik-Chervonenkis (VC) dimension of the classifier and on the number of training examples. The use of a structural risk minimisation performance measure is in contrast with the empirical risk minimisation approach used by conventional classifiers. Conventional classifiers attempt to minimise the training set error which does not necessarily achieve a minimum generalisation error. Therefore, SVMs have theoretically a greater ability to generalise. For further reading, see [38].

Unlike many other learning algorithms, the number of free parameters used in the SVM depends on the margin that separates the data and does not depend on the number of input features. Thus the SVM does not require a reduction in the number of features in order to avoid the problem of over-fitting (see, however, Section 7). This property is clearly an advantage in the context of high-dimensional applications, such as text document and authorship categorisation, as long as the data vectors are separable with a wide margin. Unfortunately, SVMs require the implementation of optimisation algorithms for the minimisation procedure which can be computationally expensive. A few researchers have applied SVMs to the problem of text document categorisation using approximately 10,000 features in some cases, concluding that, in most cases, SVMs outperform conventional classifiers [42][18]. Drucker *et al* used SVMs for classifying e-mail text as spam or non-spam and compared it to boosting decision trees, Ripper and Rocchio classification algorithms [12]. Bosch *et al* used a separating hyperplane based on a similar idea to that of a linearly separable SVM for determining the authorship of two authors of the formal articles published within the set of the Federalist Papers [4]. Teytaud *et al* [32] investigated different SVM kernels for author identification (principally well-known French authors) and language discrimination using N-graphs as the relevant features. Diederich *et al* evaluated the performance of SVMs with various features such as term frequencies, as well as structural features such as tagword bi-grams using the German Berliner Zeitung newspaper corpus [11]. Multi-topic author attribution experiments were also undertaken by Diederich *et al*. They obtained poor recall performance results when using function word bi-grams, in apparent disagreement with the assumption that function words minimise content information.

5. E-MAIL CORPUS

The choice of the e-mail corpus is limited by privacy and ethical considerations. Publicly available e-mail corpuses include newsgroups, mailing lists etc. However, in such public e-mail databases, it is generally quite difficult to obtain a sufficiently large and "clean" (i.e., void of cross-postings, off-the-topic spam, empty bodied e-mails with attachments etc.) corpus of both multi-author and multi-topic e-mails. The resulting author-topic matrix of multiple authors discussing the same set of topics is generally sparse and often characterised by having some interdependent topics (e.g., one topic is a specialisation of another, see [1]). Also, there is generally no control over the authors' characteristics or profile.

One approach that avoids the problems of e-mails obtained from newsgroups etc. is to generate a controlled set of e-mails for each author and topic. The resulting author-topic matrix is non-sparse with maximum independence between topics and minimal bias towards particular author characteristics. This approach was used in our experiment.

The corpus of e-mail documents used in the experimental evaluation of author-topic categorisation contained a total of 156 documents sourced from three native language (English) authors, with each author contributing e-mails on three topics (approx. 12,000 words per author for all topics). The topics chosen were *movies*, *food* and *travel*. The relatively small number of e-mail documents per topic category was not thought to be critical as it was observed in [1] that as few as a total of 10 or 20 documents for each author should be sufficient for satisfactory categorisation performance. The body of each e-mail document was parsed, based on an e-mail grammar that we designed, and the relevant e-mail body features were extracted. The body was pre-processed to remove (if present) any salutations, reply text and signatures. However, the existence, position within the e-mail body and type of some of these are retained as inputs to the categoriser (see below). Attachments are excluded, though the e-mail body itself is used. A summary of the global e-mail document corpus statistics is shown in Table 1.

6. EXPERIMENTAL METHODOLOGY

A number of attributes/features identified in baseline authorship attribution experiments undertaken on constrained topics (see [1] and [10]) as most useful for e-mail authorship discrimination were extracted from each e-mail body document. These attributes included both style markers as well as structural features. A total of 170 style marker attributes and 21 structural attributes were employed in the experiment. These are listed in Tables 2 and 3, respectively. Note that M = total number of *tokens* (i.e., words), V = total number of *types* (i.e., distinct words), and C = total number of characters in e-mail body. Also, the hapax legomena count is defined as the number of types that occur only once in the e-mail text.

We briefly clarify how we derive some of the attributes shown in Table 2. Firstly, the set of short words in each e-mail document consists of all words of length less than or equal to 3 characters (e.g., "all", "at", "his" etc.). Only the count of short words is used as a feature. The short word frequency distribution may be biased towards e-mail content and was therefore not used in our experiments. Secondly, the set of all-purpose function words ("a", "all", "also", ..., "to", "with") and its frequency distribution is obtained and also used as a sub-vector attribute. The total number of function words used as features was 122. Finally, a word length frequency distribution consisting of 30 features (up to a maximum word length of 30 characters) is employed.

Though our choice of attributes is specifically biased towards features that have been shown to be able to effectively discriminate between authors, rather than discriminating between topics, some of the style marker attributes may have a combination of author and content bias as, for example, hapax legomena [5].

The requoted text position refers to the reply status of e-mail. A reply text can generally be placed in any position in the e-mail document and each line is usually prefixed with a special character (e.g., ">"). In our experiment, the position of requoted text allowed for 6 different possibilities (e-mail body text interspersed with the requoted text, e-mail body text preceded by requoted text etc.). Due to some e-mailers using HTML formatting, we include the set of HTML tags as a structural metric. The frequency distribution of HTML tags was included as one of the 21 structural attributes.

The classifier used in the experiments was the Support Vector Machines classifier, SVM^{light} [37], developed by T. Joachims from the University of Dortmund. SVM^{light} is an implementation of Vapnik's Support Vector Machines. It scales well to a large number of sparse instance vectors as well as efficiently handling a large number of support vectors. In our experiments we explored a number of different kernel functions for the SVM classifier namely, the linear, polynomial, radial basis and sigmoid *tanh* functions. We obtained maximal F_1 classification results (see below for the definition of F_1) on our data set with a polynomial kernel of degree 3. The "LOQO" optimiser was used for maximising the margin.

As Support Vector Machines only compute two-way categorisation, Q two-way classification models were generated, where Q is the number of author categories ($Q = 3$ for our e-mail document corpus), and each SVM categorisation was applied Q times. This produced Q two-way confusion matrices.

To evaluate the categorisation performance on the e-mail document corpus, we calculate the accuracy, recall (R), precision (P) and combined F_1 performance measures commonly employed in the information retrieval and text categorisation literature (for a discussion of these measures see, for example, [40]), where:

$$F_1 = \frac{2RP}{(R+P)}$$

To obtain an overall performance figure over all binary categorisation tasks, a macro-averaged F_1 statistic is calculated [41]. Here, N_{AC} per-author-category confusion matrices (where N_{AC} is the total number of author categories, $N_{AC} = 3$ in our experiment) are computed and then av-

Topic Category	*Author Category* AC_i $(i = 1, 2, 3)$			*Topic Total*
	Author AC_1	*Author* AC_2	*Author* AC_3	
`Movie`	15	21	21	59
`Food`	12	21	25	58
`Travel`	3	21	15	39
Author Total	30	63	63	156

Table 1: Summary statistics of the e-mail topic and author corpus used in the experiment.

Style Marker Attribute Type
`Number of blank lines/total number of lines`
`Average sentence length`
`Average word length (number of characters)`
`Vocabulary richness i.e.,` V/M
`Total number of function words/`M
`Function word frequency distribution` (122 features)
`Total number of short words/`M
`Count of hapax legomena/`M
`Count of hapax legomena/`V
`Total number of characters in words/`C
`Total number of alphabetic characters in words/`C
`Total number of upper-case characters in words/`C
`Total number of digit characters in words/`C
`Total number of white-space characters/`C
`Total number of space characters/`C
`Total number of space characters/number white-space characters`
`Total number of tab spaces/`C
`Total number of tab spaces/number white-space characters`
`Total number of punctuations/`C
`Word length frequency distribution/`M (30 features)

Table 2: E-mail document body style marker attributes. Total of 170 features are used in the experiment. See text for clarification.

Structural Attribute Type
`Has a greeting acknowledgment`
`Uses a farewell acknowledgment`
`Contains signature text`
`Number of attachments`
`Position of requoted text within e-mail body`
`HTML tag frequency distribution/total number of HTML tags` (16 features)

Table 3: E-mail document body structural attributes. Total of 21 features are used in the experiment. See text for clarification.

eraged over all categories to produce the macro-averaged statistic, $F_1^{(M)}$:

$$F_1^{(M)} = \frac{\sum_{i=1}^{N_{AC}} F_{1,AC_i}}{N_{AC}}$$

where F_{1,AC_i} is the per-author-category F_1 statistic for author category AC_i $(i = 1, 2, , \ldots N_{AC})$:

$$F_{1,AC_i} = \frac{2R_{AC_i}P_{AC_i}}{(R_{AC_i} + P_{AC_i})}$$

7. RESULTS AND DISCUSSION

We report our results presenting the per-author-category macro-averaged F_1 statistic for the Support Vector Machines (SVM) classifier. The classification results are first presented for the case of aggregated topic categories, followed by the results for multi-topic classification.

7.1 Aggregated Topic Classes

We first present the authorship attribution results for the aggregated topics. That is, the e-mails in all of the topic classes are combined into a single topic class. This classification experiment demonstrates the ability of the learning algorithm to discriminate between authors without consideration to the topic. The results are displayed in Table 4. N.B.: Owing to the small number of data points, a stratified 10-fold cross-validation procedure was used as the sampling technique. For each author class, the resulting ten CV-sampled per-author-category confusion matrices were then averaged to give the cross-validated per-author-category confusion matrix (for that author). The per-author-category P_{AC_i}, R_{AC_i} and F_{1,AC_i} statistics are then calculated.

Table 4 indicates that the SVM classifier combined with the style markers and structural attributes is able to discriminate between the authors for the case of aggregated topics. Both the recall and precision measures for authors AC_2 and AC_3 are quite good. However, a reduced F_1 performance is noted for author AC_1, due to a low recall value.

The experiment is repeated with the classifier using only the style markers as the features compared with both style markers and structural features in the previous experiment. The results are shown in Table 5. Comparing the results in Tables 4 and 5 we observe that the style markers are the dominant features which contribute to the classification performance. The structural features only contribute a maximum of a few percentage points to the classification performance results (e.g., $F_1 = 86.2\%$ without structural features for author AC_2, compared with $F_1 = 90.5\%$ when including structural features).

7.2 Separate Topic Classes

In the second experiment, all e-mails are separated out into their individual topic categories and multi-topic author attribution is undertaken. The SVM classifier was trained on the `Movie` e-mail topic document set and tested on the remaining (unseen) e-mail topic sets (`Food` and `Travel`). Results of the second experiment are shown in Table 6.

As observed in Table 6, results indicate that, in general, the SVM classifier when used with the style markers and structural attributes is able to effectively discriminate between the authors even when multiple topic categories are involved. The results are comparable with the case of aggregated e-mail topics. One exception in these results is author AC_1's low recall (and, consequently, low F_1) results obtained with both the `Food` and `Travel` topic categories. This was also observed, though not with such low values, for the case of aggregated e-mail topics (Section 7.1). The results for author AC_1 are not conclusive since this particular author has only a small number of data points (Table 1). It may also be that this author shares some stylometric traits with the other authors.

The topic categories used in our experiments had little, if any, topic content interdependency. That is, there is minimal content correlation at the word level (though, some would say, there is a "cultural dependency" between topics, which is beyond the scope of this paper). If there were a content overlap between any two or more of the topic classes, these results would tend to deteriorate as some of the selected attributes have a content-based bias, such as hapax legomena, possibly biasing the categorisation towards the e-mail document topic content rather than its author (as observed in [2]). In such cases, this problem can be obviated by removing such attributes from the set of features used.

7.3 Function Word Type and Dimensionality

In a separate experiment, we also investigated the categorisation performance as a function of word collocation and the type and dimensionality of the function word vector attributes. In this experiment, the number of function words was increased to 320 (from 122) and the set of these were split into two categories, namely parts-of-speech (POS) words and others. Example POS word types included adverbs, auxiliary verbs, determiners, prepositions etc... The "others" category include numbers and ordinal numbers ("first" etc.). In the experiment it was observed that word collocation, increased function word distribution dimensionality, and the use of POS words did not improve the author categorisation performance across the different newsgroups. In fact, the categorisation performance deteriorated with increasing function word distribution size, which seems to be at odds with the belief that SVMs are robust in high dimensional settings.

8. CONCLUSIONS

We have investigated the learning of authorship categories for the case of both aggregated and multi-topic e-mail documents. We used an extended set of predominantly content-free e-mail document features such as structural characteristics and linguistic patterns. The classifier used was the Support Vector Machine learning algorithm. Experiments on a number of e-mail documents generated by different authors on a set of topics gave encouraging results for both aggregated and multi-topic author categorisation. However, one author category produced worse categorisation performance results, probably due to the reduced number of documents for that author. We also observed no improvement in classification performance when including word collocation and even a reduction in performance when the function word dimensionality was increased.

Performance Statistic	*Author Category, AC_i ($i = 1, 2, 3$)*		
	Author AC_1	*Author AC_2*	*Author AC_3*
P_{AC_i}	100.0	83.8	93.8
R_{AC_i}	63.3	98.3	89.6
F_{1,AC_i}	77.6	90.5	91.6

Table 4: Per-author-category P_{AC_i}, R_{AC_i} and F_{1,AC_i} categorisation performance results (in %) for the three different author categories ($i = 1, \ldots, 3$). All e-mail topic classes are aggregated into a single topic class. Both style markers and structural features described in Tables 2 and 3, respectively, are used by the classifier.

Performance Statistic	*Author Category, AC_i ($i = 1, 2, 3$)*		
	Author AC_1	*Author AC_2*	*Author AC_3*
P_{AC_i}	100.0	93.0	83.6
R_{AC_i}	60.0	80.3	93.3
F_{1,AC_i}	75.0	86.2	88.2

Table 5: Per-author-category P_{AC_i}, R_{AC_i} and F_{1,AC_i} categorisation performance results (in %) for the three different author categories ($i = 1, \ldots, 3$). All e-mail topic classes are aggregated into a single topic class. Only the style markers as described in Table 2 are used as features by the classifier.

Topic Class	*Author Category, AC_i ($i = 1, 2, 3$)*								
	Author AC_1			*Author AC_2*			*Author AC_3*		
	P_{AC_1}	R_{AC_1}	F_{1,AC_1}	P_{AC_2}	R_{AC_2}	F_{1,AC_2}	P_{AC_3}	R_{AC_3}	F_{1,AC_3}
`Food`	100.0	16.7	28.6	77.8	100.0	87.5	85.2	92.0	88.5
`Travel`	100.0	33.3	50.0	90.9	100.0	95.2	100.0	100.0	100.0

Table 6: Per-author-category P_{AC_i}, R_{AC_i} and F_{1,AC_i} categorisation performance results (in %) for the two topic categories and for the three different author categories ($i = 1, 2, 3$). The e-mail discussion topic `Movie` is used as the training set (see text).

There are a couple of limitations with the current approach. Firstly, the fact that one of the authors has a worse categorisation performance than other authors indicates that a) more data points are needed for this particular author and/or b) obtain other features that are better able to discriminate the author. For example, we are investigating function word vector subset selection in an attempt to identify the most useful function words for a given set of authors. Secondly, more studies on the usefulness of specific N-graphs for author identification should be investigated as it has been conjectured that, for example, certain bi-graphs incorporating punctuation are effective author discriminators [6]. Secondly, the number of author categories considered in our experiments at the moment is small, though not unusually small in the context of forensics where a reduced number of "suspect" authors are involved. The inclusion of additional authors is currently being investigated.

9. REFERENCES

[1] A. Anderson, M. Corney, O. de Vel, and G. Mohay. "Identifying the Authors of Suspect E-mail". *Communications of the ACM*, 2001. *(Submitted).*

[2] A. Anderson, M. Corney, O. de Vel, and G. Mohay. "Multi-topic E-mail authorship attribution forensics". In *Proc. Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security (CCS'2001)*, 2001.

[3] C. Apte, F. Damerau, and S. Weiss. "Text mining with decision rules and decision trees". In *Workshop on Learning from text and the Web, Conference on Automated Learning and Discovery*, 1998.

[4] R. Bosch and J. Smith. "Separating hyperplanes and the authorship of the disputed federalist papers". *American Mathematical Monthly*, 105(7):601–608, 1998.

[5] C. Chaski. "A Daubert-inspired assessment of current techniques for language-based author identification". Technical report, US National Institute of Justice, 1998. Available through *www.ncjrs.org*.

[6] C. Chaski. "Empirical evaluations of language-based author identification techniques". *Forensic Linguistics*, 2001. (to appear).

[7] W. Cohen. "Learning rules that classify e-mail". In *Proc. Machine Learning in Information Access: AAAI Spring Symposium (SS-96-05)*, pages 18–25, 1996.

[8] C. Crain. "The Bard's fingerprints". *Lingua Franca*, pages 29–39, 1998.

[9] O. de Vel. "Evaluation of Text Document Categorisation Techniques for Computer Forensics". *Journal of Computer Security*, 1999. *(Submitted).*

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

[11] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. "Authorship attribution with Support Vector Machines". *Applied Intelligence*, 2000. *Submitted.*

[12] H. Druker, D. Wu, and V. Vapnik. "Support vector machines for spam categorisation". *IEEE Trans. on Neural Networks*, 10:1048–1054, 1999.

[13] W. Elliot and R. Valenza. "Was the Earl of Oxford the true Shakespeare?". *Notes and Queries*, 38:501–506, 1991.

[14] J. Farringdon. *Analysing for Authorship: A Guide to the Cusum Technique*. University of Wales Press, Cardiff, 1996.

[15] D. Foster. *Author Unknown: On the Trail of Anonymous*. Henry Holt, New York, 2000.

[16] A. Gray, P. Sallis, and S. MacDonell. "Software forensics: Extending authorship analysis techniques to computer programs". In *Proc. 3rd Biannual Conf. Int. Assoc. of Forensic Linguists (IAFL'97)*, pages 1–8, 1997.

[17] D. Holmes and R. Forsyth. "The Federalist revisited: New directions in authorship attribution". *Literary and Linguistic Computing*, pages 111–127, 1995.

[18] T. Joachims. "Text categorization with support vector machines: Learning with many relevant features". In *Proc. European Conf. Machine Learning (ECML'98)*, pages 137–142, 1998.

[19] D. Khmelev. "Disputed authorship resolution using relative entropy for Markov chain of letters in a text". In R. Baayen, editor, *Proc. 4th Conference Int. Quantitative Linguistics Association*, Prague, 2000.

[20] I. Krsul. "Authorship analysis: Identifying the author of a program". Technical report, Department of Computer Science, Purdue University, 1994. Technical Report CSD-TR-94-030.

[21] I. Krsul and E. Spafford. "Authorship analysis: Identifying the author of a program". *Computers and Security*, 16:248–259, 1997.

[22] D. Lowe and R. Matthews. "Shakespeare vs Fletcher: A stylometric analysis by radial basis functions". *Computers and the Humanities*, pages 449–461, 1995.

[23] T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

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

[25] H. Ng, W. Goh, and K. Low. "Feature selection, perceptron learning, and a usability case study for text categorization". In *Proc. 20th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR97)*, pages 67–73, 1997.

[26] P. Oman and C. Cook. "Programming style authorship analysis". In *Proc. 17th Annual ACM Computer Science Conference*, pages 320–326, 1989.

[27] J. Rudman. "The state of authorship attribution studies: Some problems and solutions". *Computers and the Humanities*, 31(4):351–365, 1997.

[28] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. "A Bayesian approach to filtering junk e-mail". In *Learning for Text Categorization Workshop: 15th National Conf. on AI. AAAI Technical Report WS-98-05*, pages 55–62, 1998.

[29] P. Sallis, S. MacDonell, G. MacLennan, A. Gray, and R. Kilgour. "Identified: Software authorship analysis with case-based reasoning". In *Proc. Addendum Session Int. Conf. Neural Info. Processing and Intelligent Info. Systems*, pages 53–56, 1997.

[30] G. Salton and M. McGill. *Introduction to Modern Information Filtering*. McGraw-Hill, New York, 1983.

[31] E. Spafford and S. Weeber. "Software forensics: tracking code to its authors". *Computers and Security*, 12:585–595, 1993.

[32] O. Teytaud and R. Jalam. "Kernel-based text categorization". In *International Joint Conference on Neural Networks (IJCNN'2001)*, 2001. Washington DC, to appear.

[33] B. Thisted and R. Efron. "Did Shakespeare write a newly discovered poem?". *Biometrika*, pages 445–455, 1987.

[34] R. Thomson and T. Murachver. "Predicting gender from electronic discourse". *British Journal of Social Psychology*, 40:193–208, 2001.

[35] F. Tweedie and R. Baayen. "How variable may a constant be? Measure of lexical richness in perspective". *Computers and the Humanities*, 32(5):323–352, 1998.

[36] F. Tweedie, S. Singh, and D. Holmes. "Neural network applications in stylometry: The Federalist papers". *Computers and the Humanities*, 30(1):1–10, 1996.

[37] University of Dortmund. *Support Vector Machine, SVMLight*. `http://www-ai.cs.uni-dortmund.de/FORSCHUNG/VERFAHREN/SVM_LIGHT/svm_light.eng.html`.

[38] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[39] S. Waugh, A. Adams, and F. Tweedie. "Computational stylistics using artificial neural networks". *Literary and Linguistic Computing*, 15(2):187–198, 2000.

[40] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[41] Y. Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2):67–88, 1999.

[42] Y. Yang and X. Liu. "A re-examination of text categorisation methods". In *Proc. 22nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR99)*, pages 67–73, 1999.