Language and Gender Author Cohort Analysis of E-mail for Computer Forensics

Olivier de Vel[1], Malcolm Corney[2], Alison Anderson[2], and George Mohay[2]

[1] Defence Science and Technology Organisation,
P.O. Box 1500, Edinburgh SA 5111, Australia
[2] Faculty of Information Technology,
Queensland University of Technology, Brisbane Q4001, Australia

Abstract. We describe an investigation of authorship gender and language background cohort attribution mining from e-mail text documents. We used an extended set of predominantly topic content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features together with a Support Vector Machine learning algorithm. Experiments using a corpus of e-mail documents generated by a large number of authors of both genders gave promising results for both author gender and language background cohort categorisation.

1 Introduction

Computer forensics investigations have to increasingly deal with e-mail as this is becoming an important form of communication for many computer users, for both legitimate and illegitimate activities. E-mail is used in many legitimate activities such as message and document exchange. Unfortunately, it can also be misused, for example, in 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.

Some researchers have stated that e-mail is much like spoken communication. However, there are some important differences. For example, e-mail is more rarefied than normal spoken communication. With e-mail, participants cannot see each other's faces, hear each other's voices, or identify gestures or other visual cues. The information content in an e-mail can include simple text as well as mark-up text to convey additional information. Some senders of e-mail use only natural language text to formulate the content of the transmitted information, other users have developed an electronic "para-language" to mark-up their message and convey affective and socio-emotional information. Such informal language codes, called "emotext," include intentional misspelling (e.g., "u r ssoooo kooool"), lexical surrogates for vocalisations (e.g., "hmm"), grammatical markers (e.g., excessive use of upper-case letters, repeated question marks), and visual arrangements of text characters into "emoticons" (short combinations of normal and rotated characters to resemble facial expressions of joy, sadness etc.).

In this paper we are particularly interested in determining two characteristics of the author of an e-mail, viz. the gender and language background of the author. Gender characteristics are based on the gender-preferential (or gender-specific) language used by the author. Language background characteristics are based on the author's expression of language (for example, English as a first language, EFL, or as a second language, ESL). The paper is organised as follows. Firstly, we outline the current status of work in the area of author attribution in Section 2. We then focus our discussion on gender-preferential and language background-specific e-mail mediated communication in Section 3. Sections 4 and 5 briefly outline the Support Vector Machine learning algorithm used in our experiments, describe the e-mail corpus used, and present the methodology employed in the experiments. Validation of the method is then undertaken by presenting results of gender- and language-specific e-mail categorisation performance in Section 6. Finally, we conclude with some general observations and present future directions for the work in Section 7.

2 Background to Author Cohort Attribution

The principal objectives of author cohort (here, gender and language background cohorts) attribution are to classify an ensemble of e-mails as belonging to a particular author cohort and, if possible, obtain a set of characteristics or features that remain relatively constant for a large number of e-mails written by that particular cohort of authors. The question then arises; can characteristics such as language, structure, layout etc. of an e-mail be used, with a high degree of confidence, as a kind of author cohort phrenology and thus link the e-mail document with its author cohort? Also, can we expect the writing characteristics or style of an author cohort 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 author cohort categorisation.

Related, but separate, areas of author cohort attribution are text categorisation and authorship attribution. The former attempts to categorise a set of text documents based on its contents or topic whilst the latter attempts to identify the author of the e-mail. Many methods have been proposed for text categorisation. Most of these techniques employ the "bag–of–words" or word vector space

feature representation and use a learning algorithm such as decision trees [1], Bayesian probabilistic approaches [2], or support vector machines [3] to classify the text document. Work in e-mail text classification has also been undertaken by some researchers in the context of automated e-mail document filtering and filing (see, for example. [4]). Authorship attribution studies are also extensive and often controversial (for example, the authorship of the Federalist papers [5] and Shakespeare's works [6]). Almost all of these studies employ stylometric features ("style markers") for discriminating authors and all use large, formal texts as the source of documents. Over 1,000 stylometric features have been proposed [7], including word- or character-based stylometric features, function words, profanities, punctuation etc. Also, there exists a number of different techniques for performing the discrimination. These include statistical approaches (e.g., cusum [8], neural networks [9] and so on. Unfortunately, there does not exist a consensus on the existence of a set of uniquely discriminatory stylometric features, nor on a correct methodology as many of the mentioned techniques suffer from problems such as questionable analysis, inconsistencies for the same set of authors, failed replication etc.

A small number of studies in e-mail authorship attribution have been undertaken. Corney *et al* [10] used a set of stylometric and e-mail structural features and also studied the effect of text size and the number of e-mail documents per author on the author categorisation performance. They observed a relatively constant categorisation performance for text chunk sizes greater than approximately 100 words with, however, a significant drop-off for text sizes less than this. Also, they observed that as few as 20 documents may be sufficient for satisfactory categorisation performance. de Vel *et al* achieved satisfactory results with multi-topic and multi-author categorisation using a set of predominantly content-free e-mail document features such as structural characteristics and linguistic patterns [11].

3 Gender- and Language Background-Preferential E-mail Mediated Communication

Although computer-mediated communication (CMC) does inhibit some cues such as personal identity or individuating details (e.g., dress, location, demeanour, expressiveness), there is no evidence to suggest that all other cues are also inhibited. With e-mail mediated communication, some information about social categories or social identity, such as gender, or educational or second language (ESL) background cues are likely to be inferred in the relative absence of interpersonal context cues [12].

In the case of gender-based communications, men and women use language and converse differently even though they technically speak the same language. Empirical evidence suggests that there exist gender differences in written communication, face-to-face interaction and in computer-mediated communication. It is thought that gender-preferential language is conveyed in all of these forms of communication due, in part, to the use of intersecting or generalised gender-

preferential language attributes. Many studies have been undertaken on the issue of gender and language use (for example, see the bibliography at [13]). It has been suggested by various researchers that women's language makes more frequent use of emotionally intensive adverbs and adjectives such as "so", "terribly", "awfully", "dreadful" and "quite" and that their language is more punctuated with attenuated assertions, apologies, questions, personal orientation and support". On the other hand, male conversational patterns express "independence" and assertions of vertically hierarchical power. Men are more "proactive" by directing speech at solving problems while women are more "reactive" to the contributions of others, agreeing, understanding and supporting. Some features of men's language are "strong assertions, aggressive, self-promotion, rhetorical questions, authoritative orientation, challenges and humor". In brief, men's on-line conversation resemble "report talk", rather than "rapport talk" which women tend to favour.

Many gender-preferential CMC studies have been undertaken in recent years. However, very few studies in the area of e-mail CMC have been performed (for example, Thomson *et al* [12]) and no studies, to the authors' knowledge, in automated e-mail gender-preferential author cohort attribution have been undertaken to date.

In the case of communications based on language background differences, people with ESL and EFL generally communicate differently. For example, it is often difficult to translate slang, colloquial, or idiomatic expressions from one language to another. Authors with a poor command of their second language will often translate phrases or sentences literally, make spelling mistakes and grammatical errors, and generate incorrect text. For the case of an idiomatic expression, the unusual syntactic pattern that conveys the semantics of the expression is generally different than the sum of its parts. For example, the idiomatic expression "No me tome el pelo" in Spanish is correctly (semantically) translated into English as "Don't pull my leg" or more precisely "Don't tease me", but is translated literally as "Don't pull my hair". Authors with ESL or EFL language backgrounds will often have a different mix of vocabulary and function words. We hypothesise that these authors will have different stylometric profiles and should therefore be able to be discriminated based on a set of stylometric attributes.

In our study we use a combination of stylometric, structural and gender-preferential features, together with a Support Vector Machine classifier as the learning algorithm for cohort analysis.

4 Support Vector Machine Classifier

The Support Vector Machine's (SVM) concept is based on the idea of structural risk minimisation which minimises the generalisation error (i.e. true error on unseen examples). This true error 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. SVMs belong to the class of the

more general basis expansion and regularisation problem to which methods such as smoothing splines, multidimensional splines (eg, MARS, wavelet smoothing) belong. One advantage of SVMs is that they do not require a reduction in the number of features in order to avoid the problem of over-fitting, which is useful when dealing with large dimensions as encountered in the area of text mining. See [14] for more background information on SVMs.

Some researchers have applied SVMs to the problem of text document categorisation and author attribution concluding that, in most cases, SVMs outperform conventional classifiers (see, for example, [3]). SVMs have been used for automatic filing of e-mails as well as for classifying e-mail text as spam or non-spam [15][16].

5 E-mail Corpus and Methodology

We describe the process of generating the e-mail corpus and the selection of attributes for both the gender- and language-specific author categorisation experiments. We also briefly describe the sampling methodology used and calculation of the categorisation performance.

5.1 E-mail Corpus Generation

The generation of a suitable corpus of e-mails for the study was complicated by various factors. Firstly, the process of generating any e-mail corpus is constrained by privacy issues and ethical considerations. It is not possible to use e-mails from other people's inboxes without their consent. Unfortunately, obtaining a person's consent is an almost impossible exercise. Secondly, even though it is possible to use publicly available e-mail corpuses such as newsgroups, mailing lists etc., it is not always easy to validate the gender of the sender of each e-mail in the corpus. For example, it is not sufficient to use the sender's name as this could be an alias, indeterminate, spoofed etc.. Thirdly, it is generally 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 e-mails. Finally, it is important not to generate an e-mail corpus that is biased towards, for example, a different cohort type or e-mail topic as these may affect the categorisation results of the author cohort attribution experiment. A judicious, and time-consuming, selection of e-mails for model building is therefore paramount.

The corpus of e-mail documents used in the experimental evaluation of the gender author categorisation study was sourced from two inboxes (Pine and Netscape e-mail clients) of a member of a large (greater than 15,000 users) academic organisation[1]. The senders of the e-mail messages were selected based on the fact that they belonged to the organisation and their gender and, to a lesser extent, language background checked. All other senders (external) were

[1] In order to preserve anonymity, all third parties (such as any member of the DSTO) that were involved in the experiment were only presented with the summary statistics of the experiment and not with the contents of the e-mails in the corpus.

not considered as it was not possible to confirm their gender and/or language background reliably. Any cross-postings, re-quoted spammed e-mails (e.g., jokes, stories), general notification or broadcast e-mails relating to the organisation etc. were purged from the corpus. An initial total of 8820 e-mail documents sourced from 342 authors (approx. equally distributed between the two genders) were selected. The gender (M/F) and language background (EFL/ESL) of each author was confirmed for all e-mail documents. This document set was subsequently pared down to two subsets (not mutually exclusive), one for each cohort type, namely 4369 e-mail messages (for 325 authors) and 4932 e-mail messages (for 522 authors) for the gender and language background cohorts, respectively, to ensure only email messages with a minimum number of words equal to 50 are used (see [10] for suggested guidelines on the choice of e-mail document size). The body of each e-mail document was then parsed using an e-mail grammar, 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 were retained as inputs to the categoriser (see below). Attachments were excluded, though the e-mail body itself was used.

In order to study the impact of the number of words in an e-mail on the categorisation performance (see later), the e-mail corpus was further divided into multiple subsets. The subsets were generated by first creating a root-level subset with a minimum number of 50 words per e-mail, and then recursively generating lower-level subsets from their parent subsets with a minimum of 100, 150, 200 etc. words per e-mail. A summary of the e-mail document corpus statistics measured in terms of the number of authors in each gender and language background cohort and the number of e-mails as a function of the minimum number of words per e-mail, is shown in Tables 1 and 2.

The e-mails in the e-mail database were also sampled in sets of 50, 100, 200, 300 etc. e-mails per author cohort type, for different minimum word counts in each cohort (as shown in Tables 1 and 2). The resulting sampling gave rise to the author count for each gender and language background cohort, as shown in Tables 3 and 4, respectively.

5.2 Attribute Selection

The attributes/features selected for the experiment were members of two sets namely, a baseline stylometric- and structural-specific set, and a gender-specific set. The total number of attributes used in the experiment was 222.

The baseline stylometric set of attributes/features chosen was selected from the set identified in previous authorship attribution experiments (see [10][17][11]) for e-mail authorship discrimination. These attributes, which included both a mix of character- and word-based style markers as well as structural features, were extracted from each e-mail body document. A total of 211 baseline attributes, comprising 183 style marker attributes and 28 structural attributes, were employed in the experiment (see Table 5). Note that M = total number of *tokens* (i.e., words), V = total number of *types* (i.e., distinct words), C =

Table 1. Summary statistics of the e-mail corpus used in the experiment for the gender ([M|F]) author cohort.

Minimum Number of Words	*Male Author Cohort (Number of Authors)*	*Female Author Cohort (Number of Authors)*	*Total Number of Authors*
50	117	208	325
100	104	176	280
150	91	135	226
200	83	99	182

Minimum Number of Words	*Male Author Cohort (Number of E-mails)*	*Female Author Cohort (Number of E-mails)*	*Total Number of E-mails*
50	2071	2298	4369
100	1257	1072	2329
150	842	585	1427
200	564	384	948

Table 2. Summary statistics of the e-mail corpus used in the experiment for the language background ([EFL|ESL]) author cohort.

Minimum Number of Words	*EFL Author Cohort (Number of Authors)*	*ESL Author Cohort (Number of Authors)*	*Total Number of Authors*
50	296	226	522
100	256	136	392
150	205	92	297
200	169	62	231

Minimum Number of Words	*EFL Author Cohort (Number of E-mails)*	*ESL Author Cohort (Number of E-mails)*	*Total Number of E-mails*
50	3926	706	4932
100	2128	357	2485
150	1311	231	1542
200	878	161	1039

Table 3. Summary statistics of the author count for e-mails with different numbers of e-mail per gender cohort class and minimum word count per e-mail. Values indicated by "–" correspond to insufficient e-mail document size/word count population.

Number of E-mails per	*Minimum Word Count*							
Gender Cohort	50		100		150		200	
Class	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
50	34	40	29	38	27	31	30	34
100	49	64	49	64	43	50	42	48
200	70	97	65	96	59	79	64	68
300	84	122	76	114	73	99	73	90
400	90	141	87	128	78	115	–	–
500	100	155	91	138	82	128	–	–
1000	111	185	102	173	–	–	–	–
2000	116	206	–	–	–	–	–	–

Table 4. Summary statistics of the author count for e-mails with different numbers of e-mail per language background cohort class and minimum word count per e-mail. Values indicated by "–" correspond to insufficient e-mail document size/word count population.

Number of E-mails per	*Minimum Word Count*							
Language Cohort	50		100		150		200	
Class	*EFL*	*ESL*	*EFL*	*ESL*	*EFL*	*ESL*	*EFL*	*ESL*
50	40	39	38	38	32	29	37	35
100	69	66	67	60	65	51	60	48
200	113	109	100	95	92	86	–	–
300	145	140	131	122	–	–	–	–
400	171	171	–	–	–	–	–	–
500	187	191	–	–	–	–	–	–
600	202	206	–	–	–	–	–	–
700	214	224	–	–	–	–	–	–

total number of characters, and H = total number of HTML tags in the e-mail body. Also, attribute A_{21} is the total number of characters in words, including apostrophes and hyphens, divided by C. The *hapax legomena* count is defined as the number of types that occur only once in the e-mail text. Attributes A_8 to A_{20} are defined in Tweedie *et al* [7]. For example, Rubet's K value is computed as $\log(V)/\log(M)$.

We briefly clarify how we derive some of the attributes shown in Table 5. 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 total 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", "about", "after", "all", "also", ..., "yet", "you", "your", "yours") and its frequency distribution is obtained and also used as a sub-vector attribute. The number of function words used is 122. Finally, a word length frequency distribution consisting of 30 features (up to a maximum word length of 30 characters) is employed.

The re-quoted 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 re-quoted text allowed for 6 different possibilities (e-mail body text interspersed with the re-quoted text, e-mail body text preceded by re-quoted 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 28 structural attributes.

The set of basic gender-specific language attributes were selected from the literature presented in Section 3. These are listed in Table 6 (attributes A_{211} to A_{221}). The selected attributes attempt to measure the frequency of use of adjectives, adverbs (mainly through the presence of suffixes) and apologies. This attribute set is a small subset of possible gender-preferential language attributes listed in the literature. Note that these attributes are not unique in the ability to discriminate between genders. Indeed some of the attributes listed in Table 5 are also capable of contributing to effective gender discrimination, though which ones is an open problem at this stage.

Though our choice of attributes is specifically biased towards features that have been shown to be able to effectively discriminate between authors and, hopefully, between author gender and language background, 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* as defined in attributes A_6 and A_7 in Table 5 (see [18]). Attributes, such as N-graphs, have not been included due to their strong topic bias, even though they may be useful as language background-specific attributes (N-graphs are contiguous sequences of characters, including whitespaces, punctuation etc...).

Each attribute A_i is also scaled as follows:

$$A_i^{(\text{scaled})} = (A_i - A_{i,\min})SF_{A_i} + LB_{A_i}$$

Table 5. E-mail document body style marker and structural attributes.

Attribute Type, A_i *(*$i = 0, \ldots, 210$*)*	
Document-based:	
A_0:	Number of blank lines/total number of lines
A_1:	Average sentence length (number of words)
Word-based:	
A_2:	Average word length
A_3:	Vocabulary richness i.e., V/M
A_4:	Number of function words/M
A_5:	Number of short words/M (word length ≤ 3)
A_6:	Count of hapax legomena/M
A_7:	Count of hapax legomena/V
A_8:	Guirad's R
A_9:	Herdan's C
A_{10}:	Herdan's V
A_{11}:	Rubet's K
A_{12}:	Maas' A
A_{13}:	Dugast's U
A_{14}:	Lukjanenkov and Neistoj's measure
A_{15}:	Brunet's W
A_{16}:	Honore's H
A_{17}:	Sichel's S
A_{18}:	Yule's K
A_{19}:	Simpson's D
A_{20}:	Entropy measure
Character-based:	
A_{21}:	Number of characters in words/C (see text)
A_{22}:	Number of alphabetic characters/C
A_{23}:	Number of upper-case characters in words/C
A_{24}:	Number of digit characters in words/C
A_{25}:	Number of white-space characters/C
A_{26}:	Number of spaces/C
A_{27}:	Number of spaces/Number white-space characters
A_{28}:	Number of tab spaces/C
A_{29}:	Number of tab spaces/Number white-space characters
A_{30}:	Number of punctuation characters/C
Function Words:	
A_{31} to A_{152}:	Function word frequency distribution (122 features)
Other:	
A_{153} to A_{182}:	Word length frequency distribution/M (30 features)
Structural:	
A_{183}:	Reply status
A_{184}:	Has a greeting acknowledgement
A_{185}:	Uses a farewell acknowledgement
A_{186}:	Contains signature text
A_{187}:	Number of attachments
A_{188}:	Position of re-quoted text within e-mail body
A_{189} to A_{210}:	HTML tag frequency distribution/H (22 features)

Table 6. E-mail document gender-preferential language attributes.

Attribute Type, A_i *(*$i = 211, \ldots, 221$*)*	
Gender-Preferential:	
A_{211}:	`Number of words ending with` *able* `/`M
A_{212}:	`Number of words ending with` *al* `/`M
A_{213}:	`Number of words ending with` *ful* `/`M
A_{214}:	`Number of words ending with` *ible* `/`M
A_{215}:	`Number of words ending with` *ic* `/`M
A_{216}:	`Number of words ending with` *ive* `/`M
A_{217}:	`Number of words ending with` *less* `/`M
A_{218}:	`Number of words ending with` *ly* `/`M
A_{219}:	`Number of words ending with` *ous* `/`M
A_{220}:	`Number of` *sorry* `words /`M
A_{221}:	`Number of words starting with` *apolog* `/`M

so as to ensure all attributes are treated equally in the classification process. The scaling factor, SF_{A_i}, is computed as:

$$SF_{A_i} = \frac{UB_{A_i} - LB_{A_i}}{A_{i,\max} - A_{i,\min}}$$

with $A_{i,\min}$ and $A_{i,\max}$ being the minimum and maximum values of the attribute A_i, respectively. Also, LB_{A_i} and UB_{A_i} are the defined lower and upper bounds of the scaled attribute, respectively (we have used $LB_{A_i} = 0.0$ and $UB_{A_i} = 1.0$).

5.3 Performance Evaluation Methodology

The SVM^{light} Support Vector Machine classifier developed by T. Joachims from the University of Dortmund [19] was used in the experiments. SVM^{light} is an implementation of Vapnik's Support Vector Machine [14], as described in Section 4 . It (SVM^{light}) 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.

The Support Vector Machine computes two-way categorisation. Therefore, in our experiments on author gender categorisation, only a single two-way classification model with a two-way confusion matrix needed to be generated. The training-testing sampling methodology used was a 10-fold cross-validation of the entire e-mail document set.

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 mining literature (for a discussion of these measures see, for example, [20]), where:

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

6 Results and Discussion

We present both our author gender-preferential cohort and language background-specific cohort attribution results and report the F_1 statistic using the Support Vector Machine (SVM) classifier. The results are given for different e-mail document sizes (measured as the minimum word count) and for different e-mail author cohort sizes (number of e-mail documents per female and male author cohort for the gender cohort, and number of e-mail documents per EFL and ESL author cohort for the language background cohort). The attribution performance results for the two experiments i) gender author cohort, and ii) language background author cohort are shown in Tables 7 and 8, respectively.

Table 7. Gender-specific cohort F_1 categorisation performance results (in %) for different e-mail document sizes and for different e-mail cohort sizes. Combined attribute/feature set used (Tables 5 and 6). See text for explanation. Values indicated by "-" correspond to insufficient e-mail document size/word count population.

Number of E-mails per Gender Cohort Class	*Minimum Word Count*			
	50	100	150	200
50	64.4	62.2	57.1	59.8
100	68.4	64.0	56.8	65.0
200	64.8	61.5	62.2	63.8
300	66.4	67.6	66.6	67.3
400	67.5	68.7	70.2	–
1000	69.4	71.1	–	–

As observed in Table 7, the gender-specific cohort F_1 categorisation performance results indicate that, in general, the SVM classifier combined with the style markers, structural attributes, and gender-preferential language attributes is able to satisfactorily discriminate between the author gender cohorts. As expected, there is a general improvement, though not dramatic, in performance as the the number of e-mails in each gender cohort class increases. However, the improvement in performance as a function of the minimum word count is not as consistent as the e-mail count performance results. A noticeable improvement is only achieved when the number of e-mails in each gender cohort class is not too small (> 300). An increased minimum word count does not seem to have a large

impact on the performance results. These results indicate that a small number of e-mails per author cohort class is generally sufficient for satisfactory gender classification. This result compares favourably with similar observations made in authorship attribution studies [10].

Table 8. Language background-specific cohort F_1 categorisation performance results (in %) for different e-mail document sizes and for different e-mail cohort sizes. Combined attribute/feature set used (see Tables 5 and 6). See text for explanation. Values indicated by "-" correspond to insufficient e-mail document size/word count population.

Number of E-mails per Language Cohort Class	*Minimum Word Count*			
	50	100	150	200
50	59.7	63.3	74.0	71.0
100	65.6	70.4	76.0	77.3
200	70.6	70.3	74.7	80.8
300	72.5	71.2	–	–
400	70.9	73.0	–	–
500	72.5	–	–	–
600	73.3	–	–	–
700	74.6	–	–	–

Table 8 present slightly better F_1 categorisation performance results for the case of the language background-specific cohort compared with the gender-specific cohort results. Again, there is a general improvement, though not dramatic, in performance as the the number of e-mails in each language cohort class increases. The improvement in performance as a function of the minimum word count is, however, more significant than for the case of the gender-specific author cohort. A noticeable improvement is achieved when the number of e-mails in each language cohort class is > 100 and when, in most cases, the minimum word count increases.

Some preliminary analysis of the impact of the different types of attributes (stylistic, structural, gender-preferential) on the author gender cohort categorisation performance was also undertaken. Each type of attribute set was removed from the feature set and the performance results calculated. These are shown in Table 9.

Though preliminary at this stage, the results in Table 9 show that the full combination of attributes gives the best author gender categorisation. Removal of any of the attributes gives rise to a reduced performance value, though some more importantly than others. In particular, the set of function words (attributes A_{31} to A_{152}) are seen to be an important gender discriminator. This is as expected since function words has been shown to be a good author discriminator [11] as well as containing words that could belong to gender-preferential

Table 9. Effect of the attribute type on the F_1 categorisation performance results.

Feature Set Type	*Operation*	F_1 (%)
`Character-based attributes`	`Removed`	70.0
`Word-based attributes`	`Removed`	69.6
`Word length distribution`	`Removed`	67.4
`Structural attributes`	`Removed`	68.1
`Function words`	`Removed`	64.0
`All baseline attributes`	`-`	70.1
`All attributes (baseline + gender-based)`	`-`	70.2

language (such as "so", "very" etc.). However, we also note that the gender-preferential attributes used in the experiment only give a marginal improvement in the categorisation performance. This indicates that the current set of gender-based attributes are insufficient and a more selective and/or more extensive set of gender-preferential attributes will need to be used to achieve better categorisation performance.

7 Conclusions

In this paper, we have investigated the learning of the author gender and language background cohort categories from e-mail documents. We used an extended set of predominantly content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features together with a Support Vector Machine learning algorithm. Experiments on a number of e-mail documents generated by over 800 authors of both genders (M/F) and language background (EFL/ESL) gave promising results for both author gender and language background cohort categorisation. Author language background cohort categorisation results were observed to be better than the author gender cohort results. We observed an improvement in categorisation performance with increasing number of e-mails and minimum word count in both gender and language background cohorts.

The current approach has several limitations. Firstly, as mentioned in Section 6, a larger set of gender-preferential language attributes needs to be used to improve the gender cohort categorisation performance results further. Secondly, more studies on the usefulness of specific style markers for author gender and language background cohort identification should be investigated as it is conjectured that, for example, certain bi-graphs incorporating punctuation could be effective discriminators [21]. Thirdly, experiments to determine the best subset of attributes need to be undertaken (e.g., forward feature selection). Finally, the diversity in author characteristics in the author cohort e-mail database is currently quite small owing to the type of organisation where the e-mails were sourced. Though it is not easy to obtain a sufficiently large set of e-mails from

authors with varying cohort characteristics (educational level, language background etc.), we hope to be able to build up a suitable forensic database and further test our approach.

References

1. Apte, C., Damerau, F., Weiss, S.: Text mining with decision rules and decision trees. Workshop on Learning from text and the Web, Conference on Automated Learning and Discovery (1998)
2. Yang, Y., Liu, X.: A re-examination of text categorisation methods. Proc. 22nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR99) (1999) 67–73
3. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. Proc. European Conf. Machine Learning (ECML'98) (1998) 137–142
4. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. Learning for Text Categorization Workshop: 15th National Conf. on AI. AAAI Technical Report (WS-98-05) (1998) 55–62
5. Mosteller, F., Wallace, D.: Inference and Disputed Authorship: The Federalist. Addison-Wesley, Reading, Mass. (1964)
6. Elliot, E., Valenza, R.: Was the Earl of Oxford the true Shakespeare? Notes and Queries, **38** (1991) 501–506
7. Tweedie, F., Baayen, R.: How variable may a constant be? Measure of lexical richness in perspective. Computers and the Humanities, **32** (1998) 323–352
8. Farringdon, J.: Analysing for Authorship: A Guide to the Cusum Technique. University of Wales Press, Cardiff (1996)
9. Tweedie, F., Singh, S., Holmes, D.: Neural network applications in stylometry: The Federalist papers. Computers and the Humanities (1996) 1–10
10. Corney, M., Anderson, A., Mohay, G., de Vel, O.: Identifying the Authors of Suspect E-mail. Computers and Security (submitted) (2001)
11. de Vel, O., Anderson, A., Corney, M., Mohay, G.: E-mail Authorship Attribution for Computer Forensics. Data Mining for Security Applications, Kluwer Publishers (2002)
12. Thomson, R., Murachver, T.: Predicting gender from electronic discourse. British Journal of Social Psychology **40** (2001) 193–208
13. Schiffman, H.: Bibliography of gender and language. http://ccat.sas.upenn.edu/ haroldfs/popcult/bibliogs/gender/genbib.htm (2002)
14. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)
15. Druker, H., Wu, D., Vapnik, V.: Support vector machines for spam categorisation. IEEE Trans. on Neural Networks, **10** (1999) 1048–1054
16. Diederich, J. Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with Support Vector Machines. Applied Intelligence (2000) (submitted)
17. de Vel, O.: Mining e-mail authorship. Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000) (2000)
18. Chaski, C.: A Daubert-inspired assessment of current techniques for language-based author identification. US National Institute of Justice (1998) http://www.ncjrs.org

19. Support Vector Machine, SVMLight: University of Dortmund (2001) http://www-ai.cs.uni-dortmund.de/FORSCHUNG/VERFAHREN/SVM_LIGHT/svm_light.eng.html
20. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)
21. Chaski, C.: Empirical evaluations of language-based author identification techniques. Forensic Linguistics, **8** (2001)