Analysing E-mail Text Authorship for Forensic Purposes

by

Malcolm Walter Corney

B.App.Sc (App.Chem.), QIT (1981)
Grad.Dip.Comp.Sci., QUT (1992)

Submitted to the School of Software Engineering and Data Communications
in partial fulfilment of the requirements for the degree of

Master of Information Technology

at the

QUEENSLAND UNIVERSITY OF TECHNOLOGY

March 2003

Keywords

e-mail; computer forensics; authorship attribution; authorship characterisation; stylistics; support vector machine

Analysing E-mail Text Authorship for Forensic Purposes

by

Malcolm Walter Corney

Abstract

E-mail has become the most popular Internet application and with its rise in use has come an inevitable increase in the use of e-mail for criminal purposes. It is possible for an e-mail message to be sent anonymously or through spoofed servers. Computer forensics analysts need a tool that can be used to identify the author of such e-mail messages.

This thesis describes the development of such a tool using techniques from the fields of stylometry and machine learning. An author's style can be reduced to a pattern by making measurements of various stylometric features from the text. E-mail messages also contain macro-structural features that can be measured. These features together can be used with the Support Vector Machine learning algorithm to classify or attribute authorship of e-mail messages to an author providing a suitable sample of messages is available for comparison.

In an investigation, the set of authors may need to be reduced from an initial large list of possible suspects. This research has trialled authorship characterisation based on sociolinguistic cohorts, such as gender and language background, as a technique for profiling the anonymous message so that the suspect list can be reduced.

Publications Resulting from the Research

The following publications have resulted from the body of work carried out in this thesis.

Principal Author

Refereed Journal Paper

M. Corney, A. Anderson, G. Mohay and O. de Vel, "*Identifying the Authors of Suspect E-mail*", submitted for publication in Computers and Security Journal, 2002.

Refereed Conference Paper

M. Corney, O. de Vel, A. Anderson and G. Mohay, "*Gender-Preferential Text Mining of E-mail Discourse for Computer Forensics*", presented at the 18^{th} Annual Computer Security Applications Conference (ACSAC 2002), Las Vegas, NV, USA, 2002.

Other Author

Book Chapter

O. de Vel, A. Anderson, M. Corney and G. Mohay, "*E-mail Authorship Attribution for Computer Forensics*" in "Applications of Data Mining in Computer Security" edited by Daniel Barbara and Sushil Jajodia, Kluwer Academic Publishers, Boston, MA, USA, 2002.

Refereed Journal Paper

O. de Vel, A. Anderson, M. Corney and G. Mohay, "*Mining E-mail Content for Author Identification Forensics*", SIGMOD Record Web Edition, 30(4), 2001.

Workshop Papers

O. de Vel, A. Anderson, M. Corney and G. Mohay, "*Multi-Topic E-mail Authorship Attribution Forensics*", ACM Conference on Computer Security - Workshop on Data Mining for Security Applications, November 8 2001, Philadelphia, PA, USA.

O. de Vel, M. Corney, A. Anderson and G.Mohay, "*Language and Gender Author Cohort Analysis of E-mail for Computer Forensics*", Digital Forensic Research Workshop, August 7 Ű 9, 2002, Syracuse, NY, USA.

Contents

List of Figures

List of Tables

Abbreviations Used in this Thesis

The following abbreviations are used throughout this thesis.

Acronyms

$\overline{E}^{(M)}$	Weighted Macro-averaged Error Rate
CMC	Computer Mediated Communication
ENL	English as a Native Language
ESL	English as a Second Language
F_1	F_1 Combined Measure
$\overline{F}_1^{(M)}$	Weighted Macro-averaged F_1 Combined Measure
HTML	Hyper-Text Markup Language
SVM	Support Vector Machine
UA	User Agent

Feature Set Names

C	Character based feature set
D	Document based feature set
E	E-mail structural feature set
F	Function word feature set
G	Gender preferential feature set
H	HTML Tag feature set
L	Word length frequency distribution feature set
W	Word based feature set

Variables Used in Feature Calculations

C	Total number of characters in a document
H	Total number of HTML tags in a document
N	Total number of words (tokens) in a document
V	Total number of types of words in a document

Statement of Original Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed . . . . . .

Date . . . . . .

Acknowledgments

I would like to thank the following people, without whom this work would not have been possible.

Firstly, thanks to my supervisors for this project. My principal supervisor, Dr. Alison Anderson, gave much support throughout the project, remained enthusiastic throughout and really helped to kick this thesis into shape. Alison commented many times that this was a ‘fun’ project and I must agree. I would also like to thank my associate supervisor, Adjunct Professor George Mohay, for his continual feedback on the project and on the thesis during its preparation. Thanks also to George for offering me this project in the first place.

I must thank Olivier de Vel from DSTO, Edinburgh, SA, for initiating this project with a research grant and also for his collaboration with the publications that resulted from the project.

Finally, I thank my children, Tomas and Nyssa for their patience on recent weekends and I must thank my wife, Diane, for her encouragement, her support and her patience, especially during the last few months of the preparation of this thesis.

Malcolm Corney
March 2003

Chapter 1

Overview of the Thesis and Research

This chapter outlines the problem attacked by this research and the approach used to solve it. Section 1.1 discusses why forensic tools are needed to identify the authorship of anonymous e-mail messages, noting the increased usage of e-mail in recent years and the consequent increase in the usage of e-mail for criminal purposes. As criminal activity increases, so must law enforcement and investigative activities, to prevent or analyse the criminal activities. Computer forensics is a field which has grown over recent years, necessitated by the increase in computer related crime (see for example Mohay et al., 2003).

A discussion of the general approach to solving the problem follows in Section 1.2. Section 1.3 outlines the structure of the thesis and the conclusions of the chapter are given in Section 1.4.

1.1 Problem Definition

1.1.1 E-mail Usage and the Internet

Many companies and institutions have come to rely on the Internet for transacting business, and as individuals have embraced the Internet for personal use, the amount of e-mail traffic has increased markedly particularly since the inception of the World

Wide Web. Lyman and Varian (2000) estimated that in the year 2000 somewhere between 500 and 600 billion e-mail messages would be sent, with a further estimate of more than 2 trillion e-mail messages to be sent per year by 2003. In the GVU's[1] 8^{th} WWW User Survey (Pitkow et al., 1997), 84% of respondents said that e-mail was indispensable.

With this increase in e-mail traffic comes an undesirable increase in the use of e-mail for illegitimate reasons. Examples of misuse include: sending spam or unsolicited commercial e-mail (UCE), which is the widespread distribution of junk e-mail; sending threats; sending hoaxes; and the distribution of computer viruses and worms. Furthermore, criminal activities such as trafficking in drugs or child pornography can easily be aided and abetted by sending simple communications in e-mail messages.

There is a large amount of work carried out on the prevention and avoidance of spam e-mail by organisations such as the Coalition Against Unsolicited Commercial E-mail (CAUCE), who are lobbying for a legislative solution to the problem of spam e-mail. E-mail by its nature is very easy to send and this is where the problem lies. Someone with a large list of e-mail addresses can send an e-mail message to the list. It is not the sender who pays for the distribution of the message. The Internet Service Providers whose mail servers process the distribution list pay with CPU time and bandwidth usage and the recipients of the spam messages pay for the right to receive these unwanted messages. Spammers typically forge the 'From' address header field, so it is difficult to determine who the real author of a spam e-mail message is.

Threats and hoaxes can also be easily sent using an e-mail message. As with spam messages, the 'From' address header field can be easily forged. In the United States

[1]GVU is the Graphic, Visualisation and Usability Center, College of Computing, Georgia Institute of Technology, Atlanta, GA.

of America, convictions leading to prison sentences have been achieved against people who sent e-mail death threats (e.g. Masters, 1998). An example of an e-mail hoax is sending a false computer virus warning with the request to send the warning on to all people known to the recipient, thus wasting mail server time and bandwidth.

Computer viruses or worms are now commonly distributed by e-mail, by making use of loose security features in some e-mail programs. These worms copy themselves to all of the addresses in the recipient's address book. Examples of worms causing problems recently include Code Red (CERT, 2001a), Nimda (CERT, 2001c), Sircam (CERT, 2001b), and ILOVEYOU (CERT, 2000).

The common thread running through these criminal activities is that not all e-mail messages arrive at their destination with the real identity of the author of the message even though each message carries with it a wrapper or envelope containing the sender's details and the path along which the message has travelled. These details can be easily forged or anonymised and the original messages can be routed through anonymous e-mail servers thereby hiding the identity of the original sender.

This means that only the message text and the structure of the e-mail message may be available for analysis and subsequent identification of authorship. The metadata available from the e-mail header, however, should not be totally disregarded in any investigation into the identification of the author of an e-mail message. The technical format of e-mail as a text messaging format is discussed in Crocker (1982).

Along with the increase in illegitimate e-mail usage, there has been a parallel increase in the use of the computer for criminal activities. Distributed Denial of Service Attacks, viruses and worms are just a few of the different attacks generated by computers using electronic networks. This increase in computer related crime has seen the development of computer forensics techniques to detect and protect evidence

in such cases. Such techniques discussed in the next section, are generally used after attacks have taken place.

1.1.2 Computer Forensics

Computer forensics can be thought of as investigation of computer based evidence of criminal activity, using scientifically developed methods that attempt to discover and reconstruct event sequences from such activity. The practice of computer forensics also includes storage of such evidence in a way that preserves its chain of custody, and the development and presentation of prosecutorial cases against the perpetrators of computer based crimes. Yasinsac and Manzano (2001) suggest that any enterprise that uses computers and networks should have concern for both security and forensic capabilities. They suggest that forensic tools should be developed to scan continually computers and networks within an enterprise for illegal activities. When misuse is detected, these tools should record sequences of events and store relevant data for further investigation.

It would be useful, therefore, to have a computer forensics technique that can be used to identify the source of illegitimate e-mail that has been anonymised. Such a technique would be of benefit to both computer forensics professionals and law enforcement agencies.

The technique should be able to predict with some level of certainty the authorship of a suspicious or anonymous e-mail message from a list of suspected authors, which has been generated by some other means e.g. by the conduct of a criminal investigation. If the list of suspects is large, it would also be useful to have a technique to create hypotheses concerning certain profiling attributes about the author, such as his or her gender, age, level of education and whether or not English was the author's native

language. This profiling technique could then reduce the size of the list of possible suspects so that the author of the e-mail message could be more easily identified. Figure 1-1 shows a schema of how the suggested techniques could work.

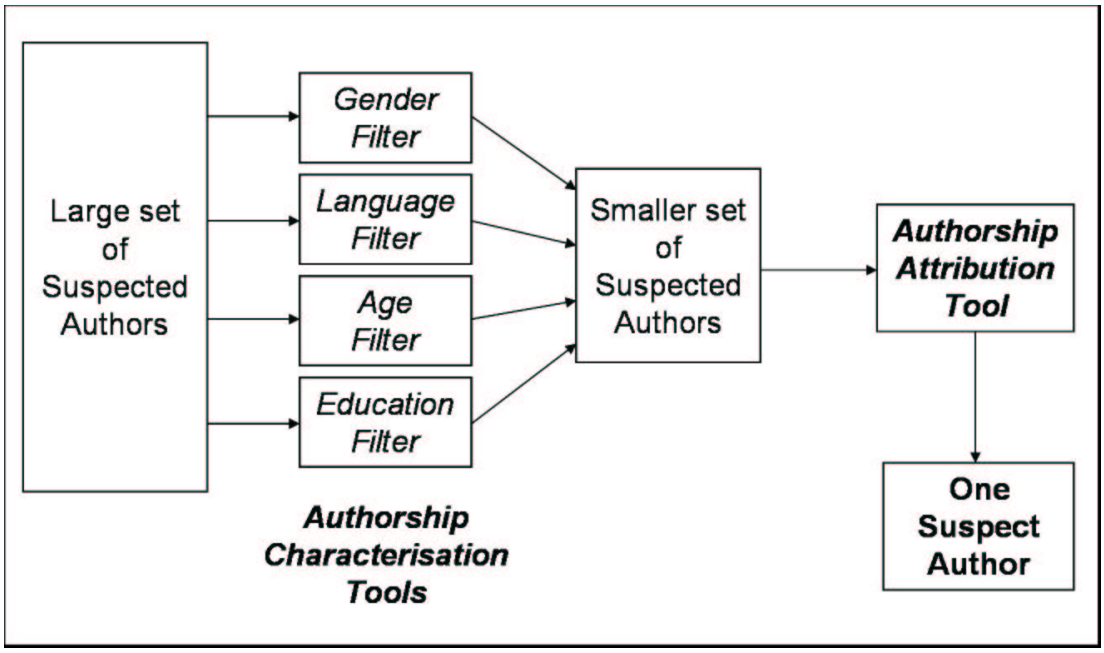


Figure 1-1: Schema Showing How a Large List of Suspect Authors Could be Reduced to One Suspect Author

1.2 Overview of the Project

1.2.1 Aims of the Research

This research set out to determine if the authorship of e-mail messages could be determined from the text and structural features contained within the messages themselves

rather than relying on the metadata contained in the messages. The reason for attempting this was to establish tools for computer forensics investigations where anonymous e-mail messages form part of the evidence.

The aim was to use techniques from the fields of authorship attribution and stylometry to determine a pattern of authorship for each individual suspect author in an investigation. A message under investigation could then be compared to a group of authorship patterns using a machine learning technique.

Stylometric studies have used many features of linguistic style and comparison techniques over the many years that these studies have been undertaken. Because these studies used only some of the many available features at any one time, and the comparison techniques used were unable to take into account many features, an optimal solution has not been found. The number of words investigated for each author in these studies were quite large when compared to the typical length of an e-mail message. Most studies (see Chapter 2) suggested that a minimum of 1000 words is required to determine such a pattern. A further aim of this research was to determine if authorship analysis could be undertaken with e-mail messages containing 100 to 200 words or less.

In a forensic investigation it is quite possible that there may not be a large number of e-mail messages that can be unquestionably attributed to a suspect in the investigation. Any tool that was to be developed would need to be able to extract the authorship pattern from only a small number of example messages. This of course could lead to problems with the ability of the machine learning technique being used to predict the authorship of a questioned e-mail message. The research, therefore, also had to answer the question of how many example e-mail messages are required to form the pattern of authorship.

A further aim was to determine a method to reduce the number of possible suspected authors so that the best matching suspected author could be found using the tool mentioned above.

This research has attempted to:

- determine if there are objective differences between e-mail messages originating from different authors, based only on the text contained in the message and on the structure of the message
- determine if an author's style is consistent within their own texts
- determine some method to automate the process of authorship identification
- determine if there is some inherent difference between the way people with similar social attributes, such as gender, age, level of education or language background, construct e-mail messages.

By applying techniques from the fields of computational linguistics, stylistics and machine learning, this body of research has attempted to create authorship analysis tools for computer forensics investigations.

1.2.2 Methodology

After reviewing the related literature, a range of stylometric features was compiled. These features included character based features, word based features including measures of lexical richness, function word frequencies, the word length frequency distribution of a document, the use of letter 2-grams, and collocation frequencies.

The Support Vector Machine (SVM) was selected as the machine learning algorithm most likely to classify authorship successfully based on a large number of features. The reason for selecting SVM was due to its performance in the area of text

classification, where many text based features are used as the basis for classifying documents based on content (Joachims, 1998).

Baseline experiments were undertaken with plain text chunks of equal size sourced from fiction books and PhD theses. Investigations were carried out to identify the best sets of stylometric features and to determine the minimum number of words in each document or data point and also the minimum number of data points for reliable classification of authorship of e-mail messages. The basic parameters of the SVM implementation used, i.e. SVM^{light} (Joachims, 1999), were also investigated and their performance was tuned.

The findings from the baseline experiments were used as initial parameters when e-mail messages were first tested. Further features specific to e-mail messages were added to the stylometric feature sets previously used. Stepwise improvements were made to maximise the classification performance of the technique. The effect of topic was investigated to ensure that the topic of e-mail messages being investigated did not positively bias the classification performance.

To produce a means of reducing the list of possible authors, sociolinguistic models of authorship were constructed. Two sociolinguistic facets were investigated, the gender of the authors and their language background i.e. English as a native language and English as a second language. The number of e-mail messages and the number of words in each message were investigated as parameters that had an effect on the production of the models.

This research was not aimed at advancing the field of machine learning, but it did use machine learning techniques so that the forensic technique developed for the attribution of authorship could be automated by generating predictive models of authorship. These models were used to distinguish between the styles of various

authors. Once a suite of machine learning models was produced, unseen data could be classified by analysing that data with the models.

1.2.3 Summary of the Results

- The Support Vector Machine learning algorithm was found to be suitable for classification of authorship of both plain text and e-mail message text.

- The approach taken to group features into sets and to determine each feature set's impact on the classification of authorship was successful. Character based features, word based features, document based features, function word frequencies, word length frequency distributions, e-mail structural features and HTML tag features proved useful and each feature set contributed to the discrimination between authorship classes. Bi-gram features, while successful with plain text classification were thought to be detecting the topic or content of the text rather than authorship. The frequencies of collocations of words were not successful discriminators, possibly due to being too noisy due to the short text length of the data when these features were tested.

- Baseline testing with plain text chunks sourced from fiction books and PhD theses indicated that approximately 20 data points (e-mail messages) containing 100 to 200 words per e-mail message were required for each author in order to generate satisfactory authorship classification results.

- When the authorship of e-mail messages was investigated, the topic of the e-mail messages was found not to have an impact on classification of authorship.

- Sociolinguistic filters were developed for cohorts of gender and language background i.e. English as a native language versus English as a second language.

1.3 Overview of the Following Chapters

Chapter 1 has described why forensic tools for the identification of the authorship of e-mail messages are required, and presented an overview of the work. Chapter 2 describes the background to the problem of authorship attribution of e-mail messages and the strategies that have been used to date.

The details of the way that the experiments for this body of research were conducted are discussed in Chapter 3. This includes a description of why machine learning is helpful in this instance and which machine learning techniques were used. The sources of the data used for experimental work are also described.

The results of the experimental work are presented in Chapters 4 and 5. Chapter 4 presents the results of a set of baseline tests that were used in Chapter 5 to determine if stylistics could be applied to e-mail messages for attribution of authorship. This chapter determined some of the basic parameters for the research. Chapter 5 shows the results of the experimental work carried out on e-mail messages and also includes the results of authorship characterisation experiments where some sociolinguistic characteristics are determined about the authors of e-mail messages.

Chapter 6 contains a discussion of the major outcomes from this body of research and outlines the impact this work may have on future work in the area. Finally a glossary of terms, a set of appendices and a bibliography are included.

1.4 Chapter Summary

This chapter has discussed how e-mail is being abused more frequently for activities such as sending spam e-mail messages, sending e-mail hoaxes and e-mail threats and distributing computer viruses or worms via e-mail messages. These e-mail messages

can be easily forged or routed through anonymous e-mail servers, highlighting the need for techniques to determine the authorship of any text within them.

A discussion of the major concepts of the thesis and the approach taken to enable the identification of the authorship of anonymous e-mail messages has been outlined.

The next chapter discusses research that has been carried out into authorship attribution of literary texts through the use of stylistics and the classification techniques that have been used for such attributions. The implications of the related work on this body of research are discussed.

Chapter 2

Review of Related Research

Chapter 1 outlined the need in the field of computer forensics for tools to assist with the identification of the authorship of e-mail messages that have been sent anonymously or deliberately forged.

This chapter draws upon the results of research carried out in the fields of computational linguistics, stylistics and non-traditional authorship attribution[1] to develop a possible framework for the attribution of e-mail text authorship. Other research fields such as text classification, software forensics, forensic linguistics, sociolinguistics and machine learning also impact on the current study. Although much work has been done on text classification and authorship attribution related to prose, little work has been conducted in the specific area of authorship attribution of e-mail messages for forensic purposes.

In the authorship attribution literature it is thought that there are three kinds of evidence that can be used to establish authorship: external, linguistic and interpretive (Crain, 1998). External evidence includes the author's handwriting or a signed manuscript. Interpretive evidence is the study of what the author meant when a document was written and how that can be compared to other works by the same author.

[1]Non-traditional authorship attribution employs computational linguistic techniques rather than relying on external evidence, such as handwriting and signatures, obtained from original manuscripts.

Linguistic evidence is centred on the actual words and the patterns of words that are used in a document. The main focus of this research will be on linguistic evidence and stylistics, as this approach lends itself to the automated analysis of computer mediated forms of communication such as e-mail.

Most work to date in this latter area has used chunks of text that have significantly more words than most e-mail messages (Johnson, 1997, Craig, 1999). A question that this research must answer is, therefore: how can linguistics and stylistics be adapted to identify the authorship of e-mail messages? There are sub-problems to be investigated, such as how long an e-mail must be and how the similarities between a particular author's e-mail messages are to be measured.

2.1 Stylometry and Authorship Attribution

The field of stylometry is a development of literary stylistics and can be defined as the statistical analysis of literary style (Holmes, 1998). It makes the basic assumption that an author has distinctive writing habits that are displayed in features such as the author's core vocabulary usage, sentence complexity and the phraseology that is used. A further assumption is that these habits are unconscious and deeply ingrained, meaning that even if one were to make a conscious effort to disguise one's style this would be difficult to achieve. Stylometry attempts to define the features of an author's style and to determine statistical methods to measure these features so that the similarity between two or more pieces of text can be analysed. These assumptions are accepted as core tenets for the research conducted in this thesis.

Authorship analysis can be broken into a number of more specific yet distinct problems such as authorship attribution, authorship characterisation and plagiarism detection. The relationship between these problems is shown in Figure 2-1.

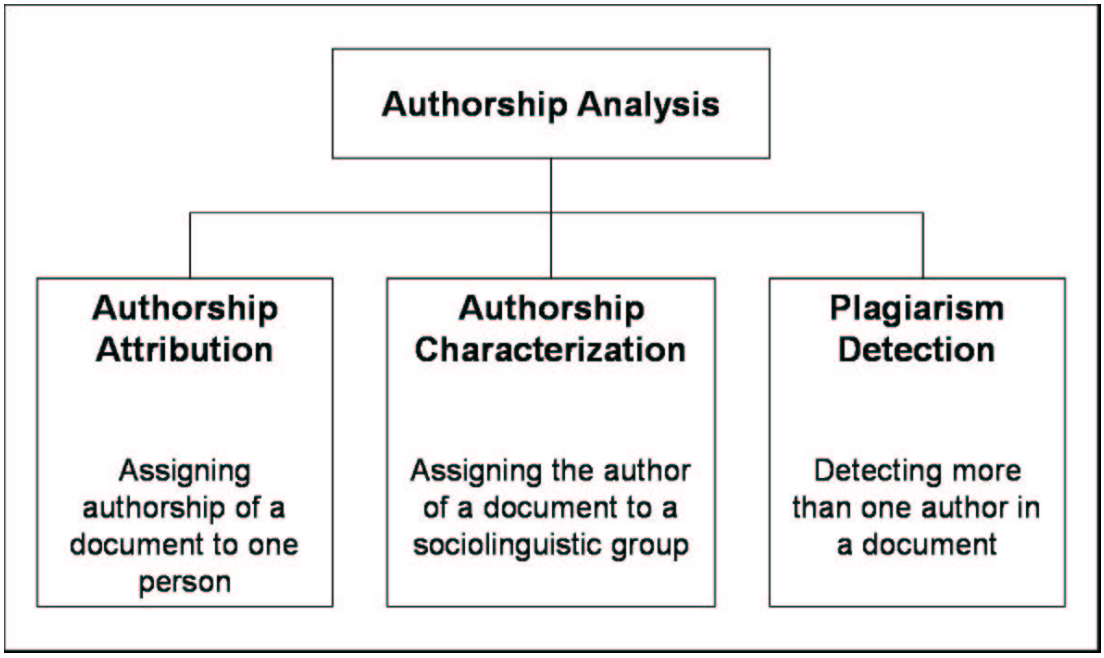


Figure 2-1: Subproblems in the Field of Authorship Analysis

Authorship attribution can be defined as the task of determining the author of a piece of text. It relies on some sort of evidence to prove that a piece of text was written by that author. Such evidence would be other text samples produced by the same author.

Authorship characterisation attempts to determine the sociolinguistic profile of the characteristics of the author who wrote a piece of text. Examples of characteristics that define a sociolinguistic profile include gender, educational and cultural background and language familiarity (Thomson and Murachver, 2001).

Plagiarism detection is used to calculate the degree of similarity between two or more pieces of text, without necessarily determining the authors, for the purposes of determining if a piece of text has been plagiarised. Authorship attribution and authorship characterisation are quite distinct problems from plagiarism detection.

Authorship analysis has been used in a number of application areas such as identifying authors in literature, in program code and in forensic analysis for criminal cases. The most widely studied application of authorship analysis is in attributing authorship of works of literature and of published articles. Well known studies include the attribution of disputed Shakespeare works e.g. Efron and Thisted (1976), Elliott and Valenza (1991a), Lowe and Matthews (1995), Merriam (1996) and the attribution of the Federalist papers (Mosteller and Wallace, 1964, Holmes and Forsyth, 1995, Tweedie et al., 1996).

2.1.1 A Brief History

The earliest studies into authorship attribution include those by Mendenhall (1887), Yule (1938, 1944) and Zipf (1932).

Mendenhall (1887) studied the authorship of Bacon, Marlowe and Shakespeare by comparing word spectra or characteristic curves, which were graphic representations of the arrangement of their word length and the relative frequency of their occurrence. He suggested that if the curves remained constant and were particular to the author, this would be a good method for authorship discrimination.

Zipf (1932) focussed his work on the frequencies of the different words in an author's documents. He determined that there was a logarithmic relationship, which became known as Zipf's Law, between the number of words appearing exactly r times in a text, where $(r = 1, 2, 3 \ldots)$ and r itself.

Yule (1938) initially used sentence length as a method for differentiating authors but concluded that this was not completely reliable. He later created a measure using Zipf's findings based on word frequencies, which has become known as Yule's

characteristic K. He found that a word's use is probabilistic and can be approximated with the Poisson distribution.

Research in the field continued throughout the 1900's with mainly statistical approaches being used on one or a small number of distinguishing features. In his review of the analysis of literary style, Holmes (1985), lists a number of possible sources for features and techniques for the analysis of authorship. These include:

- word-length frequency distributions
- average syllables per word and distribution of syllables per word
- average sentence length
- distribution of parts of speech
- function word frequencies
- vocabulary or lexical richness measures, such as the Type-Token ratio, Simpson's Index (D), Yule's Characteristic (K) and entropy[2]
- vocabulary distributions, including the number of *hapax legomena*[3] and *hapax dislegomena*[4]
- word frequency distributions

Many of the studies utilizing single features make use of the Chi squared statistic for discrimination between different authors. Multivariate techniques such as factor analysis, discriminant analysis and cluster analysis have also been used.

[2]These terms are defined in the Glossary
[3]*hapax legomena* are words that are used once only in any text
[4]*hapax dislegomena* are words used twice in a text

Since the early 1990's Foster (1996c, 1999) has had differences of opinion with Elliott and Valenza (1991b, 1996, 1998, 2002) on the techniques used by the latter for the attribution of Shakespearean play and poem authorship. Foster (1999) claims that the tests used by Elliott and Valenza are "deeply flawed, both in their design and execution". Elliott and Valenza (2002) have countered these claims, have corrected the small errors in their technique, and claim that after two years of intense scrutiny their methods stand up for the attribution of Shakespearean authorship. Foster (1996a) in the meantime has also claimed that a text containing a poem titled 'A Funeral Elegy' was the work of Shakespeare while studies by other researchers did not arrive at similar conclusions. Foster compared the text of this poem with the canonical works of Shakespeare[5] by studying his diction, grammatical accidence[6], syntax and use of rare words.

In other attribution studies, Shakespeare has been compared with Edward de Vere, the Earl of Oxford (Elliott and Valenza, 1991b), John Fletcher (Lowe and Matthews, 1995) and Christopher Marlowe (Merriam, 1996). Elliott and Valenza used incidences of badge words[7], fluke words[8], rare words, new words, prefixes, suffixes, contractions and a number of other tests to build a Shakespeare profile for comparison with other authors. Lowe and Matthews used frequencies of five function words and a neural network analyser, while Merriam used some function words and principal component analysis.

The Federalist papers are a series of articles written in 1787 and 1788 to persuade the citizens of New York to adopt the Constitution of the United States of America.

[5]Shakespeare's canon includes those poems and plays that fit the accepted productive time line of his life.

[6]Grammatical accidence is the study of changes in the form of words by internal modification for the expression of tense, person, case, number etc.

[7]Badge words are words that are preferred by a particular author relative to other authors.

[8]Fluke words are words that are not preferred by a particular author relative to other authors.

There are 85 articles in total, with agreement by the authors and historians that 51 were written by Alexander Hamilton and 14 were written by James Madison. Of the remaining articles, five were written by John Jay, three were jointly written by Hamilton and Madison and 12 have disputed authorship between Hamilton and Madison.

This authorship attribution problem has been visited numerous times since the original study of Mosteller and Wallace (1964), with a number of different techniques employed. Using four different techniques to compare the texts under examination, the original study compared frequencies of a set of function words selected for their ability to discriminate between two authors. The techniques used by Mosteller and Wallace included a Bayesian analysis, the use of a linear discrimination function, a hand calculated robust Bayesian analysis and a simplified word usage rate study. Mosteller and Wallace came to the conclusion that the twelve disputed papers were written by Madison.

Other studies (Tweedie et al., 1996, Holmes, 1998, Khmelev and Tweedie, 2002) on the Federalist papers have also been conducted using various techniques. Further details of these studies are given in Sections 2.1.3 and 2.1.4. In nearly all cases, these techniques came to the same conclusion as Mosteller and Wallace.

Foster (1996b) used text analysis to identify the author of the novel *Primary Colors*, a satire of the rise of President Clinton, which was originally published anonymously. He identified linguistic habits such as spelling, diction, grammatical accidence, syntax, badge words, rare words and other markers of an author's style to narrow a list of suspected authors of the book to Joe Klein, a former advisor to the President. Foster (2000) also contributed to the search for the 'Unabomber', Ted Kaczynski, by using his text analysis techniques to compare the 'Unabomb Manifesto' with other writings by

Kaczynski given to the FBI by Kaczynski's brother. Much of Foster's work, however, appears to be quite subjective, as he does not give enough detail for others to validate his technique.

In these studies mentioned above, many different features, such as frequencies of certain words, habits of hyphenation and letter 2-grams have been used to discriminate authorship. There is no consensus of opinion between the many research groups studying the problem as to which features should be used or in fact which are the best features for discrimination of authorship. According to Rudman (1998) "at least 1000 'style markers' exist in stylometric research"

There is also no consensus as to the best technique for discriminating among authors using the chosen features. This continued debate between various proponents in the literature exposes the disagreement within stylometry research over the choice of discriminatory features and over the statistical or other classification techniques used to calculate the differences between authors' style.

It would appear that combinations of style markers should be more discriminatory than single markers, but the classification techniques used to date have not been sophisticated enough to be able to employ many features. It is suggested here that an author's style can be captured from a number of distinctive features that can be measured from the author's text and that these features will form a unique pattern of authorship.

Forsyth (1997) compiled a benchmark suite of stylometry test problems known as Tbench96 to provide a broader variety of test problems than those being used by other researchers in stylometry and related fields. This suite of text includes prose and poetry for authorship problems, poems for the study of stylochronometry, and magazine and newspaper articles for analysis of content. Few researchers in the area select more than

one problem for testing their techniques. Forsyth suggests that any technique should be tested on more than one problem so that overfitting of the data can manifest itself.

2.1.1.1 Stylochronometry

While stylometry assumes that each author has their own particular style, stylochronometry further assumes that this style can change over a period of time. Stylochronometry concerns itself with the issue of assigning a chronological order to a corpus of an author's works.

Forsyth (1999) argues that stylochronometric studies should be proven on works where the dating is well documented. He studied the verse of W. B. Yeats by counting distinctive substrings. He successfully conducted a number of tests, including the assignment of ten poems absent from the training set to their correct period, and the detection of differences between two poems written when Yeats was in his twenties and revised when he was in his fifties.

Smith and Kelly (2002) used measures of lexical richness such as *hapax legomena* and vocabulary richness such as Yule's Characteristic (K) and Zipf's Law to order the works of three authors from classical antiquity chronologically.

The results of these various studies seem to indicate that an author's style can and does change over a period of time. In these cases the period of time in question was more than ten years. These results should be kept in mind for any forensic investigations, and the known writings of any particular investigated author should be sampled from a period of time which is relatively short in this context, such as one or two years.

2.1.1.2 Literary Fraud and Stylometry

The *Prix Goncourt* is France's highest literary award and as such is only allowed to be awarded to an individual author once. Romain Gary, however, won the award a second time by writing under the pseudonym Émile Ajar (Tirvengadum, 1998). Gary admitted this in a book published after his suicide. Tirvengadum used vocabulary distributions as style discriminants, particularly high frequency words and synonyms to study the works of Gary and Ajar. Student's *t* test, the Pearson correlation and the Chi squared tests were used as the statistical methods for discrimination of the books. The Gary books correlated well, as did most of the Ajar books. Correlations between the Gary and Ajar books also were high. However, the second *Prix Goncourt* winner, written under the Ajar pseudonym, was significantly different from the others and from the Gary books. Tirvengadum concluded that "Gary consciously changed his style so as to avoid detection as Ajar."

If style can be disguised, it remains to be seen whether or not disguise can be implemented in short documents as well as long ones. For the lay persons who may be unaware of what style encompasses, it may well be beyond their skill level to disguise that style.

2.1.2 Probabilistic and Statistical Approaches

The number of words used once, twice etc. in the Shakespearean canon was analysed probabilistically in a study performed by Efron and Thisted (1976). They concluded that if a new volume of Shakespeare were discovered containing a certain number of words, it would contain a certain quantity of words that had never been used by Shakespeare in any of his previous works. This approach was based on a method used by statistician Sir Ronald Fisher in 1943, to predict how many new species

of butterfly might be discovered if butterfly hunters were to return to Malaysia to re-establish a trapping programme. A new poem that begins *'Shall I die'*, thought to have been written by Shakespeare, was found about ten years after the original study. The predicted numbers of words never used, used once, twice etc. given the number of words in the poem, fit the profile quite well (Thisted and Efron, 1987). For example, if the poem was written by Shakespeare, they calculated it should contain seven previously unseen words. When checked, the new poem contained nine words that had not been used previously by Shakespeare.

Smith (1983) used average word-length and average sentence-length, collocations[9] and measures of words in certain positions in sentences with the Chi squared statistic for detection of differences between Shakespeare and Marlowe. He concluded that word-length often produces incorrect results, as does sentence-length. He also suggested that the Chi squared statistic has been subject to misinterpretation and was misused by some proponents in the field.

The Cusum technique is described in detail by Farringdon et al. (1996). It is based on a technique from statistical quality control and relies on the assumption that the usage of short words i.e. 2 or 3 letter words, and words beginning with a vowel, are habitual and can discriminate between authors. The technique plots the cumulative sum of the differences between observed small word counts in a sentence and the average of all small word counts in the entire document. It is supposed to be able to detect multiple authorship in a document. The appeal of this technique is the claim that a text sample as small as five sentences can be tested against material of known authorship. Furthermore, the Cusum technique has been put forward as forensic linguistic evidence in court on more than one occasion, both in the United Kingdom

[9] A collocation is a combination of two words together or separated by a certain number of words. Examples of collocations include 'as the', 'in the', 'of the', 'to the' etc.

and in Australia (e.g. Lohrey, 1991, Storey, 1993). There is further discussion of the use of stylistics for forensic purposes in Section 2.1.5.

The work of Farringdon et al. has been criticised by De-Haan (1998) in a review outlining the shortcomings of the Cusum technique. De-Haan reports tests demonstrating its unreliability. Hardcastle (1993, 1997) also questions the validity of the technique, presenting examples of its failures. He summarises the findings of other researchers and concludes that Cusum results should not be accepted as reliable evidence of authorship.

Supporters of the Cusum technique (Canter and Chester, 1997) suggested an improvement by weighting the calculations of the Cusum values. When this was tested it was found that the technique could not reliably discriminate between documents that had been authored by a single author and those from multiple authors.

Sanford et al. (1994) investigated some of the basic assumptions that the Cusum technique is based upon, such as the assumption that individuals demonstrate "habit" indicators and that the habits are the same whether the material is written or spoken. The authors concluded that the technique is based on assumptions that are of limited reliability and are most likely false.

2.1.3 Computational Approaches

The development of stylistics was ongoing during the period of the Cusum debate. The research carried out attempted to define the 'best' features and to apply more sensitive classification techniques than simple count statistics. Some of the leaders in the field of stylistics during this period were Burrows, Baayen and co-authors, and Holmes and co-authors. A discussion of some of their work follows.

Burrows (1992) carried out authorship attribution by analysing the frequency patterns of the words that appeared most often in the texts being examined, correlating each word with all others using the Pearson product-moment method. He then used principal component analysis to transform the original variables to a set of new uncorrelated variables, which were arranged in descending order of importance. Typically, the new data was plotted as a graph of the first component against the second, displaying the values for each data point so that a visual separation could be effected. This essentially reduced the dimensionality of the multivariate problem to two or three dimensions. This is a good technique for visualisation of the differences between authorship and as such, it remains a qualitative tool.

Baayen et al. (1996) conducted experiments with a syntactically annotated corpus. The syntactic annotation was in the form of 'rewrite rules' generated by parsing the text of the corpus. Each 'rewrite rule' contained part of speech and phrasal information and for the purposes of the experiments, each 'rewrite rule' was considered to be a pseudo-word. Baseline tests were conducted using measures of lexical richness at the word level and by identifying the fifty most frequent words in 2000 word document chunks. The attributions produced resulted in some errors. Similar tests were conducted on the pseudo-words, which resulted in an improvement in the classification efficiency.

Holmes et al. (2001) used traditional and non-traditional methods of authorship attribution to identify seventeen previously unknown articles published in the *New York Tribune* between 1889 and 1892 as the work of Stephen Crane[10]. 3000 word samples of text were analysed for frequencies of 50 common words proposed by Burrows (1992). Principal component analysis was used as the method of discrimination.

[10]Stephen Crane was a nineteenth century American writer, and is best known for *The Red Badge of Courage*. He also worked as a journalist for the *New York Tribune*.

Khmelev and Tweedie (2002) used Markov chains of two letters, also known as 2-grams, with a probabilistic approach to analyse authorship of 387 texts from 45 authors collected from the Project Gutenberg (n.d.) archives with the classification being successful 73% of the time. They also used this technique for analysis of the Federalist papers successfully.

Benedetto et al. (2002a) applied compression techniques and entropy measures to characterise the language being used and also for classification of authorship. In this approach, a document to be classified is concatenated to a known document and the relative compression levels of the two documents gives an indication of the source of the text. If the same author wrote both pieces of text, the compression rates for the original and concatenated documents should be similar. This document was criticized by Goodman (2002) for not being related to physics yet being published in a physics journal, for not being novel work, and also for having results which were slower to produce and less accurate than those produced by a baseline technique, the Naïve Bayes classifier. Benedetto et al. (2002b) correctly point out in their response that Goodman conducted his comparison experiments by classifying document topic rather than language and authorship, which was the focus of the original paper.

The sheer variety of features and analytic methods described above indicate that there is some success but no consensus as to the best features and methods for authorship attribution. Perhaps some method may be the best for authorship attribution among these, if the variables can be controlled.

2.1.4 Machine Learning Approaches

A number of machine learning approaches have been applied in the field of stylometry in recent times. Matthews and Merriam (1993) and Merriam and Matthews (1994)

were the first to employ neural network classifiers for stylometry. These initial studies compared Shakespeare to both Fletcher and Marlowe. Kjell (1994b) used relative frequencies of the 50 or 100 most significant *n*-grams[11] as his feature set for attribution of the Federalist papers. He varied *n* from 2 to 5 and found that the best value for *n* was 2. He also compared the use of a neural network with a Naïve Bayesian classifier (Kjell, 1994a) and with a Nearest Neighbour classifier (Kjell et al., 1995).

Hoorn et al. (1999) used neural network, k-nearest neighbour and Naïve Bayes classifiers with 2-grams and 3-grams as the feature set for authorship analysis of three Dutch poets. The classification accuracy in these experiments ranges from 56.1% to 83.6%. The neural network classifier gave the best results out of the three classifiers used, while the Naïve Bayes classifier had the poorest results. The results indicate that the stylometric techniques discussed here are applicable to not only English but also to other languages.

Lowe and Matthews (1995) describe the use of a Radial Basis Function (RBF) neural network for the classification of plays written by Shakespeare and Fletcher. The features used here were five function words[12] ('are', 'in', 'no', 'of' and 'the'). Fifty samples of text from each author were used for this analysis. They report that the RBF correctly classifies 99 of the 100 training examples, and when used on disputed works, it produced results in general agreement with contemporary opinions generated by other means. The RBF performed more accurately than the benchmark methods also reported in their paper.

[11] An *n*-gram is a sequence of *n* letters from a piece of text. In the word *hello*, the 2-grams *he*, *el*, *ll* and *lo* can be formed. A significant *n*-gram was one that was found to produce the best discrimination between the two authors in question.

[12] Function words are words unrelated to the content of a piece of text e.g. conjunctions, prepositions, pronouns, determiners, contractions and some verbs and adverbs.

The Federalist papers were also studied by Tweedie et al. (1996) using neural network classification, with eleven function words as features. The results from this study concurred with those generated by Mosteller and Wallace (1964). Holmes and Forsyth (1995) used a rule based learner for their study of the Federalist papers, once again agreeing with the results of the previous researchers of the attribution.

Holmes and Forsyth (1995), Holmes (1998) compared the effects of vocabulary richness, and word frequency analysis with a genetic rule based learner, BEAGLE, on the problem of attributing the Federalist papers. Both approaches were successful. They concluded that these more sophisticated machine learning techniques have a place in the field of authorship attribution.

Waugh et al. (2000) used up to 100 features in their study where they attempted to produce the smallest network possible to solve the authorship problem. To aid in reaching this goal, they built their neural network with the Cascade-Correlation algorithm[13] and by restricting the number of features, by removing 25%, 50% or 75% of them in a random fashion. They found that a large number of features was not strictly necessary to achieve high levels of classification.

The approaches described here using neural networks have used more features than many of the other studies listed in previous sections, but the number of features in most cases has been constrained due to the difficulties in training a neural network with large input dimension. These difficulties arise due to the computational complexity of the training task and the possibility of overfitting the training data, leading to a loss in generalisation of the resulting network. The number of training examples for a neural network is also of importance. With traditional back propagation neural network

[13]This algorithm works by adding nodes gradually to a network to improve classification efficiency.

training, good generalisation usually occurs only if there is a larger number of training examples relative to the the training weights or the number of trainable parameters[14].

In the research work reviewed here, much of the work has involved the study of only one problem e.g. the Federalist papers or some novels chosen from Project Gutenberg. The work has in some cases compared classification techniques for analysis of the problem being investigated, and in other cases has compared the analysis of different data sets using one classification technique. There is little discussion of the results of tuning the parameters of the classifiers or of how the best features were chosen. In most cases, raw success and error frequencies have been reported.

To date, the most complex machine learning methods used in stylometry have been neural network classifiers. These networks have been produced with a limited number of features and the networks produced may have been overly complex, resulting in poor generalisation performance. As research continues in this area, a more scientific approach to the use of machine learning techniques is becoming apparent as the classifiers themselves are being examined rather than just the features required for classification.

2.1.5 Forensic Linguistics

Forensic linguistics is a field of study where the application of linguistic techniques such as stylometry and authorship attribution is used for forensic purposes, i.e. for the collection of evidence for use in a court of law. Totty et al. (1987) discuss the difficulties in applying linguistic techniques in the forensic setting often due to the small size of the text being examined. These techniques have been used for the

[14] Although, as shown by Bartlett (1997), good generalisation performance may well be achieved if the magnitude of the training weights is suitably controlled.

comparison of witness statements taken down by police officers with the original verbal utterances. In some cases the written statements were found to be in conflict.

Storey (1993) gives some examples of cases in England and in Australia where forensic analysis of text has been undertaken for both trial and appeal cases. Various aspects of sociolinguistics and psycholinguistics were studied for these cases. She suggests that while incontrovertible proof is not yet possible, forensic linguistics provides a powerful investigative tool for profiling and identification.

Chaski (1997) points out that forensic linguistics has not been accepted as a recognised technique in a court of law because up to now, it fails the so called Daubert criteria for admitting scientific and technical evidence in US courts of law. The Daubert criteria (Brodsky, 1999) now applicable in the U.S.A. require that a method must demonstrate reproducibility, empirical validation, peer review and known error rates. Chaski (2001) presents results of empirical testing of three groups of techniques for author attribution. The first group contains syntactically classified punctuation and syntactic analysis of phrase structure. The second group contains techniques such as sentential complexity, vocabulary richness, readability and content analysis; and the third group contains forensic stylistics techniques such as spelling errors, punctuation errors, word form errors and grammatical errors.

Chaski (2001) selected a group of four authors who were sociolinguistically matched on features - such as age, sex, race, dialect and education level - and employed the Chi squared statistic for comparison of one author's measures to another. The outcomes of the tests showed that only some of the employed techniques in the first group produced a suitable outcome. Only one basic measurement at a time was compared and was investigated for being a discriminatory technique or not. There was no attempt made to see the effect of the features in conjunction with one another.

Chaski claims that if the technique works for this set of authors it should work for any set of authors, whether or not they have a similar sociolinguistic basis. This last claim seems to be counterintuitive. If the technique works for a distinct sociolinguistic subset of authors, then one could not be sure that if authors from a different background were introduced that one was not measuring the sociolinguistic difference rather than the authorial difference.

There are also concerns that Chaski has pinned authorship attribution for forensic purposes to a narrow set of unsophisticated techniques. Grant and Baker (2001) respond to Chaski (2001), raising concerns over the techniques that were used. Grant and Baker point out that Chaski has ignored probably fifty years of research into authorship attribution, and focuses on a few techniques rather than on what has been proven to be reliable in the past, e.g. function words and the work of Mosteller and Wallace on The Federalist Papers. Grant and Baker discuss the concepts of validity and reliability of authorship attribution in detail. They suggest that it is not simply enough to identify markers of authorship but that a theoretical understanding of why a marker discriminates authorship should be sought, a point reinforced by McMenamin's (2001 p. 94) further comments that "CC says *junk science* must be eliminated from the courtroom (p. 2), then proceeds to demonstrate, in her own way, that all approaches other than her own 'violate theoretical principles of modern linguistics' . . . The specter of the straw man is so omnipresent in all this pernicious junk science as to make us shudder."

Forensic linguistics is thus a long way from becoming a science. However, for the purposes of this study, it is encouraging that this research sets out to use verifiable, repeatable, empirical methods with known error rates, in the spirit of the Daubert criteria.

2.2 E-mail and Related Media

2.2.1 E-mail as a Form of Communication

As a genre of computer mediated communication, e-mail has been sited between spoken word and letter writing. It has some properties of both genres (Baron, 1998) and it seems plausible that different messaging media affect the way that messages are constructed. The most recent example of this is SMS messaging from mobile phones, where the limitations of the keypad of the devices has affected the way words are written e.g. ‘u’ replaces ‘you’ and ‘r’ replaces ‘are’. Messages can be constructed in much less time if these replacements are made.

Gains (1999) attempted to determine some of the features of e-mail messages and how they were written by comparing academic and commercial authors and the purpose for which their e-mails were written. He investigated the distribution mode, i.e. to a single recipient or multiply distributed; and the function, i.e. as information, as a request or as a directive with further classification into whether the message was an initial message or a response. With only a small number of e-mail messages studied $(n = 116)$, he concluded that e-mail written for commercial purposes appeared to be an established form of everyday internal communication, offering benefits of being instantaneous and easily distributable. Academic e-mail on the other hand appeared to be more diverse in its usage, being used for dissemination of information, requests for information, to make and maintain contact with others and to chat.

Sallis and Kassabova (2000) suggested that e-mail authors tend to neglect grammar, spelling and good vocabulary, instead writing with shortened words and incomplete sentences. These tendencies can also lead to the depersonalisation of communication as there is no face-to-face conversation, which has its own social cues. Misunderstandings can also occur as there is no vocal intonation in the transmission of the message.

Emoticons[15] can only go so far to alleviate these problems. Sallis and Kassabova used messages from Usenet newsgroups to study various stylometric characteristics, including sentence and paragraph lengths, number of unique words and readability scores, their results indicating that e-mail is an informal means of communication.

E-mail communication in the workplace, however, is not necessarily private (Sipior and Ward, 1995). While many employees believe that e-mail messages may be regarded as confidential between the sender and receiver, some employers view the monitoring of e-mail messages as a right if not a necessity to prevent abuse of company resources.

Most authorship attribution studies to date have focussed on determining the authorship of written anonymous texts such as poems, essays and books. The research carried out for this thesis makes the assumption that e-mail text is more similar in style to written text than to spoken word, although elements of both genres will appear in e-mail messages as they are a genre in their own right.

2.2.2 E-mail Classification

The classification problem of whether or not an e-mail message is spam has been investigated extensively (e.g. Cohen, 1996, Sahami et al., 1998, Drucker et al., 1999, Androutsopolous et al., 2000a). Machine learning tools such as Naïve Bayesian classifiers (Androutsopolous et al., 2000b) and Support Vector Machines (Drucker et al., 1999) have been used for this task.

Another e-mail classification problem has been that of learning to filter e-mail messages into relevant subject folders based on the content of the message. The Ripper learning algorithm (Cohen, 1996) has been used for this task to induce rules

[15] Combinations of punctuation characters to show some form of the tone of the message to the reader e.g. :-) for happy or :-(for sad.

for classification. Crawford et al. (2001) described the "i-ems" project which aimed to automatically induce appropriate categories or subjects for sorting e-mail messages.

None of the above work, however, has attempted to classify authorship of the e-mail messages in question.

2.2.3 E-mail Authorship Attribution

De Vel (2000) attempted to identify and attribute authorship of e-mail messages using Support Vector Machines (discussed in Section 2.4.1) by using a collection of typical stylistic features such as the frequencies of short words and function words. Promising results were achieved, but the approach used was limited and far from optimised. The main features used were presented as raw word frequencies which had not been normalised to take into account the differing lengths of the e-mail messages in the corpus being investigated. de Vel introduced the idea of some structural features that can be exploited, including the 'reply status' of the e-mail message, i.e. an original message or a reply.

Tsuboi (2002) studied authorship attribution of e-mail messages and World Wide Web documents written in Japanese. He used the 'bag of words' approach, popular in text classification, for his feature set and the SVM as the machine learning classifier. He also used sequential word patterns or word n-grams, with $n = 2$ and 3, from each sentence in the documents. As word order is relatively free in Japanese, word segments could be an important indicator in that language. The reported accuracy of classification for e-mail documents is greater than 95%. Although this study contributes to authorship attribution in Japanese, the technique may not be transferable to English.

2.2.4 Software Forensics

Research in the field of software forensics has been carried out to identify the author of a computer program (Spafford and Weeber, 1993, Krsul and Spafford, 1997). Various objective metrics, such as the proportion of blank lines, the proportion of comments, the average length of identifiers and statistics based on those metrics have been proposed to characterise the author of a program with some success. Kilgour et al. (1997) proposed the use of other variable measures such as the presence or absence of a feature in an attempt to identify better the authorship of a program. In these cases the variables were assigned a discrete value, for example 0 for absent and 1 for present, instead of using a value calculated from some frequency distribution.

Some structural features may be of use in the classification of e-mail message authorship too. E-mail messages have macro-structural features, such as the presence or absence of greetings, farewells, signatures and attachments, that can be analysed along with the micro-features of the text contained within them although these are readily falsified.

2.2.5 Text Classification

Text classification has become a widely researched problem with the advent of the Internet and search engines for information retrieval from the World Wide Web. Newly crawled web pages have to be assigned a subject so that when a search is conducted, relevant documents can be retrieved. Text classification attempts to categorise a set of text documents based on its contents or topic. Many methods have been proposed for text classification, and most of these use the “bag of words” approach as it has been found that the ordering of the words in a document is of minor importance for determining its content or subject. In this approach, each word in each document of a

corpus is assigned to be a distinct feature. For each document to be classified, a word vector feature representation is constructed based on the frequencies of the words.

To dampen the effect of widely variant frequencies of words in individual documents and in the entire corpus, the frequencies are often weighted using the Term Frequency Inverse Document Frequency (TFIDF) approach (Joachims, 1997). This word weighting schema says that if a word occurs frequently in a document i.e. it has a high term frequency, that word is an important indexing term. On the other hand if a word occurs in many documents, it will be less important as an indexing term and will have a low inverse document frequency.

A number of classification techniques have been used including decision trees (Apte et al., 1998), Bayesian probabilistic approaches (Yang, 1999) and Support Vector Machines (Joachims, 1998). Joachims (1998) compared SVM with four conventional machine learning approaches: the Naïve Bayes classifier, the Rocchio algorithm, an instance based k-nearest neighbour classifier and the C4.5 decision tree/rule learner. The experiments were conducted on two different data sets. The k-nearest neighbour classifier performed best among the four conventional methods on both data sets. SVM using the polynomial kernel function[16] and the radial basis kernel function outperformed the k-nearest neighbour classifier.

Furthermore, the four conventional methods required the "best" features to be selected from the feature set. SVM performed better than the conventional methods even when all features from the feature set were used, indicating that there were few irrelevant features in the problem domain and/or that SVM was still able to discriminate with the irrelevant features included in the feature vector.

[16]Given two points in an input set, and an embedding Hilbert space, the function that returns the inner product between their images in that embedding space is known as the kernel function(Lanckriet et al., 2002).

For this study, the lesson from recent text classification research is that sophisticated machine learning techniques such as the Support Vector Machine provide the most promising avenue to pursue. How the SVM allows learning with many features is further discussed in Section 2.4. As has been mentioned in Section 2.1.4, previous studies in authorship attribution either chose to use a small number of features, or were constrained to use a small number because the classification techniques could not handle a feature space with high dimensionality.

Both Rudman (1998) and Holmes (1998) suggest that combinations of features successful in previous stylometric research should be used to produce a better method of discrimination of authorship. Using SVM as the classification tool could be the solution to the problem of including many features in authorship attribution. The Support Vector Machine was selected as the machine learning tool used in this project due to its reported ability to generalise with many features.

2.3 Sociolinguistics

Sociolinguistics is the study of language as it is actually used in social contexts. There are differences in the way that different social classes speak and write. Such distinctions may be found between:

- people of different gender (Thomson and Murachver, 2001)
- people of different age groups (Rayson et al., 1997)
- people with different educational backgrounds (Baayen et al., 2000)
- people with different ethnic backgrounds or levels of familiarity with the language under investigation e.g. English as a native language vs. English as a second language

It may be possible to determine various sociolinguistic classes, such as gender or age, for characterising or profiling authors of e-mail messages. The approach used for authorship attribution should be applicable for this problem also, i.e. evaluation of features from the e-mail messages followed by machine learning to produce sociolinguistic models. An e-mail message with unidentified authorship could be analysed using the various sociolinguistic models to reduce the number of suspect authors in the set to be further analysed for authorship identification.

Rayson et al. (1997) studied gender differences, age differences (under and over 35 years of age) and social group differences based on occupationally graded social class categories commonly used in market research. The social groups were aggregated into two larger groups, i.e. upper and lower class groups. The gender differences are discussed further in the next section. Different tendencies in word usage were found for both the age and social groups.

Baayen et al. (2000) conducted an experiment where students of different educational level were asked to write texts of around 1000 words in their native Dutch language. The texts written by the students were in three different genres. No authorial structure was discerned between the texts but a difference was noted between the different genres. Some difference was detected in measures of vocabulary richness between the students at different levels in their education.

2.3.1 Gender Differences

It has been established by Ojemann (1983) that different parts of the brain are activated by men and women for some language tasks. Empirical evidence suggests that men and women converse differently even though they speak the same language. These gender differences exist in written communication, face to face interaction and computer

mediated communication (CMC). Many studies have been undertaken on the issue of gender and language use, and gender preferential studies in CMC have been undertaken recently. Very few studies involving e-mail have been done and no studies attempting to automate the classification of gender in e-mail messages have been found in the literature.

Singh (2001) studied transcripts of spoken communication and developed a number of measures of lexical richness. He compared male and female writings using discriminant analysis. He found from the small sample of subjects who underwent the study $(n = 30)$ that male speech was lexically richer and tended to use longer phrases. Female speech used more verbs, shorter sentence structures, used lexical items more repetitively and used nouns and pronouns interchangeably.

Rayson et al. (1997) analysed a conversational corpus and found that men use more swear words and number words e.g. hundred, three etc., while women used the feminine pronouns *she, her, hers* and first person pronouns *I, me, my, mine* with higher frequencies. In general a higher preference for common nouns was found for males, while females preferred proper nouns, verbs and personal pronouns.

In a study undertaken to predict gender from electronic discourse, Thomson and Murachver (2001) suggested that gender preferential language patterns from spoken communication transferred in part to CMC. The many studies of gender have discovered that in general, women are more likely than men to refer to emotions, use intensive adverbs, make compliments, use personal pronouns, ask questions and use more words associated with politeness. Men are more likely to use terms related to quantity, use insults, make grammatical errors and provide more opinions. They found that not all people of the same gender use the same set of gender preferential features. There does not appear to be a pattern to which all men or women conform.

Sussman and Tyson (2000) studied message postings to Usenet newsgroups that were designated as male, female or gender neutral topics. Men's postings were generally longer than those of women, but women initiated new topics more often.

It is also believed that people modify their communication style depending on the type of communication that is being undertaken so that the style is relevant to the situation. Differences exist in formal/informal communication and in male/female communication. Gender-preferential features have been found to be more common in same gender communication, as opposed to mixed gender communication.

In a study based on CMC from Usenet postings, Savicki et al. (1996) suggested that when a larger proportion of men take part in discussion groups, that men use more statements of fact, speak in action based terms, are more argumentative and use more coarse and abusive language. He also postulated that when there are a larger proportion of women in a discussion group, women would use language that is more self disclosing, apologises more and asks more questions. The study showed that only some of the variables measured were significant in distinguishing gender. For males, these were the use of facts and action words, while for females the variables that were significant were the use of personal pronouns and 'we' pronouns.

Herring (1993) conducted a gender preferential communication study based on CMC. The documents studied were postings to Usenet newsgroups. Like the above studies, the study showed differences in the topics that men and women discussed, as well as in the way they discussed them. The messages studied showed that women were more likely to express doubt, apologise, ask questions and to suggest ideas rather than make assertions. Men's postings were more likely to show self promotion, make insults, use sarcasm and make strong assertions. It was suggested that women use a "rapport style" of communication while men use a "report" style.

A study was undertaken by Hills (2000) to determine if males and females could convey a false gender identity in CMC. She studied various aspects of how the participants language changed when attempting to do this. It was found that when trying to establish a false gender identity, the participants exaggerated what they thought were the gender preferential features at the word and clause level. They did not manipulate many of the features or do so particularly well. In general, the participants manipulated the topic of their writings to try to affect their gender identity.

2.3.2 Differences Between Native and Non-Native Language Writers

The use of the English language has spread throughout the world by two main mechanisms (Bhatt, 2001). The first of these was the transplantation of the language, initially to Wales, Scotland and parts of Ireland and then by movement of English speaking people to North America, Australia and New Zealand. These countries have adopted English as their native language. The second mechanism took English to non-English sociocultural contexts such as South Asia, Africa and Latin America. By this second mechanism, English has come into contact with unrelated languages and varieties of English have formed in countries such as India, Malaysia, Singapore and Nigeria. This has led to different ways of teaching English in such countries. The spread of the English language and its diversification to form "World Englishes" by contact with other cultures should leave measurable features in all forms of communication.

One clue for this project is that using a second language makes it more difficult to translate slang or idiomatic expressions from one language to another. Due to different grammatical rules in different languages, non-native writers often translate phrases or sentences literally. They may also make more and possibly characteristic spelling mistakes and grammatical errors.

Some studies into language differences between speakers of different languages have been limited to distinguishing differences in a localised geographic area. Johnstone (1999) studied the differences between speakers from African American, Hispanic and local backgrounds in Texas, USA, using discourse analysis. Johnstone found that all speakers studied were idiosyncratic and that the class, sex, age and region all impacted on the way that they spoke. An investigation of the use of English, Spanish and Creole in the Caribbean was undertaken by LePage and Tabouret-Keller (1985). They found that when people spoke in a certain situation, it was the result of a choice of identity with one group or another.

These studies would imply that the stylistic effects of language background, as measured by text and structure of e-mail, are unknown and could profitably be investigated.

2.4 Machine Learning Techniques

Stylometry makes measures of the discriminatory features proposed for authorship attribution. This reduces the style of a particular author's profile to a pattern. Machine learning is particularly suited to pattern matching problems and was used as a tool in this research for classification of authorship patterns. Machine learning techniques have the ability to predict a classification for an unseen test point, i.e. to generalise about unseen data. The previously discussed work from the humanities area, has not taken advantage of the improvements in the field of machine learning. As shown above, the machine learning technique favoured in previous work has been the neural network.

According to Witten and Frank (2000), "Things learn when they change their behaviour in a way that makes them perform better in the future." A machine learning algorithm attempts to learn from a set of example data in order to generalise about

unseen data. We can train the algorithm by optimizing the learning process via manipulation of the variables of the algorithm itself and of the problem domain. The algorithm must produce some type of model representing the knowledge it has learned, and we must measure its performance or its ability to classify unknown examples to determine how good the model is.

There are a number of different types of machine learning algorithms and some of these are discussed here briefly.

Rule Based Learners Rule based learners attempt to make rules from the feature values in the training data. For each feature in the data, the algorithm determines the frequency of the feature values or discretised bands of feature values and determines the class of instances to which the most common value belongs. A rule is created for each feature that assigns the class from the feature value or range of values and each rule is then tested using each feature. The rules with the lowest error rates are then chosen to classify unseen data.

Decision Trees Decision trees can be induced based on information gain[17] (Quinlan, 1986) using a top-down approach. At each level of the tree starting at the root node, the feature providing the maximum information gain ratio from its classes is selected. This produces a decision tree with minimum structure. Quinlan produced the C4.5 classifier using this approach and has optimised it to perform discretisation of data when the feature values are purely numeric and to cope with missing feature values.

Instance Based Learners An instance based learning algorithm uses a distance function to determine which member of the training set an unknown test instance is

[17]Information gain is the difference between the information value (also known as entropy) of the data before the decision tree is split and the information value of the data after the split.

closest to. This method is particularly suitable for numeric data as the distance function is easily calculated in these cases. The k-nearest neighbour classifier uses the Euclidean distance function and this method has become widely used for pattern recognition problems.

Neural Networks Neural networks are examples of nonparametric methods, meaning that they can construct a representation of a problem from data where an explicit model of the problem domain is difficult to calculate or is unknown. Values for data features are fed to the input nodes of the neural network and are manipulated by transfer functions at each node. The input data is passed through one or more hidden layers of nodes and finally on to a set of output nodes. The input nodes are fully connected to each node in the hidden layer and the hidden layer is similarly connected to each node in the set of output nodes. The transfer functions may be non-linear in nature. The neural network must be trained by adjusting the weights of the connections between the nodes to minimise the error rate of the output nodes with the training data. Unseen test data can be fed into the trained neural network and the output class will be predicted.

Support Vector Machines Support Vector Machines (SVMs) extend the concept of classification with linear models. The input data vector values which are limited to being from two classes and may be non-linear, are transformed from Euclidean space into a new higher dimensional Hilbert space. A model, the maximum margin hyperplane is then constructed in this new space. The maximum margin hyperplane is the model with the greatest separation between the two classes. It is defined by the data vectors which are closest to the hyperplane. These data vectors are termed “support vectors”. Further discussion of the SVM algorithm is given in Section 2.4.1. Test data vectors are similarly transformed into the

new space and are classified by determining which side of the maximum margin hyperplane they are situated on. The model defined by the maximum margin hyperplane is based upon a kernel function. This can be a linear kernel function, a polynomial kernel function or other non-linear functions such as a radial basis function kernel or a sigmoid kernel. The best kernel function is generally determined experimentally.

Rule based learners and decision trees require that the data values for each feature be limited to a set of values or be discretised into a set of ranges for each numeric feature. The discretisation of the data can be critical to the success of these techniques. Instance based learners such as the k-nearest neighbour technique can be computationally intensive as each test instance must be compared with each training instance to find the best matching output class. For these reasons, these machine learning techniques were not considered for use in this research.

This leaves the Naïve Bayes, neural network and SVM as techniques that may be suitable for use in authorship attribution. It was the aim of this research to use a large number of different features to define the authorship profile or pattern.

Neural networks can, depending on the problem domain, be limited to the use of a small number of features. There are two problems that arise when training a neural network with a large number of features. First, because the gradient descent first algorithm is used, if a large dimensionality is used, training is difficult. Second, it is necessary to limit the number of trainable parameters for fear of overfitting the data by producing a network which is too complex. Overfitting is problematic because it can lead to a loss of generalisation ability of the classifier.

2.4.1 Support Vector Machines

The fundamental concepts of Support Vector Machines were developed by Vapnik (1995). The SVMs' concept is based on the idea of structural risk minimisation (SRM). SRM attempts to minimise the generalisation error, i.e. the true error on unseen examples, which is bounded by the sum of the training set error, i.e. the empirical risk, and a term which depends on the Vapnik-Chervonenkis (VC) dimension of the classifier and on the number of training examples. The VC dimension is a measure of the capacity of the classifier model.

The classifier must try to minimise error on the training set to achieve low empirical risk and it must limit the capacity of the model to avoid overfitting the training data to achieve low structural risk. When both are kept low, the generalisation error is limited. The SVM algorithm performs this balancing act even when there are a large number of features measured in the problem domain, i.e. there is a high dimensionality.

SVMs do not suffer from overfitting to the same extent as neural networks, as only the training data vectors that are needed to maximise the separation of the classes are used to define the decision boundary. These vectors are termed the support vectors. Figure 2-2 shows an example of a linear hyperplane with the decision boundary separating the positive and negative classes with maximum margin. While there are many possible hyperplanes that could separate the vectors in a separable problem, the optimal hyperplane is the one that separates the vectors with maximum margin. If certain data vectors are added or removed from the training data, only those that are support vectors affect the decision boundary of the separating hyperplane.

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.

The outcome of text classification research discussed in Section 2.2.5 has shown that the SVM appears to be a better machine learning classifier for the text classification problem domain than those discussed above, including the Naïve Bayes classifier, the C4.5 classifier, the k-nearest neighbour classifier, and the neural network.

Text classification uses the approach of using word count frequencies, often weighted in some manner, as the feature set. Many of the features that were used in this research were based on frequencies of some type and it is believed that SVM should be a suitable classifier for the attribution of authorship of e-mail documents.

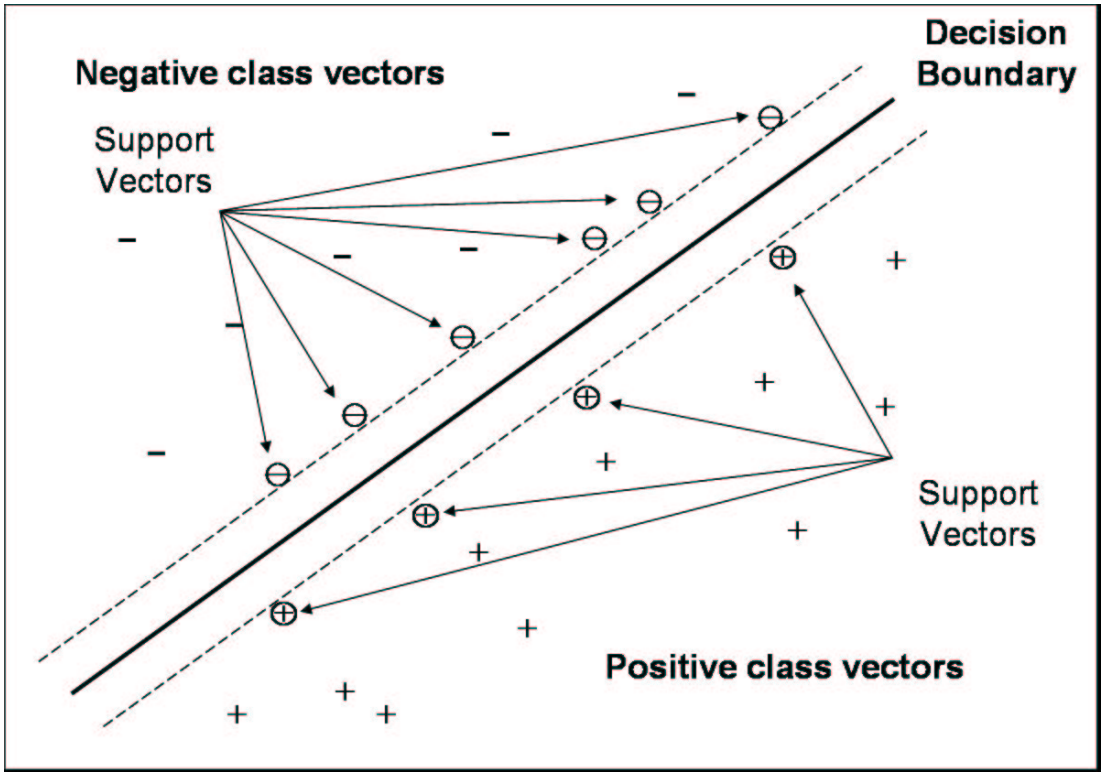


Figure 2-2: An Example of an Optimal Hyperplane for a Linear SVM Classifier

2.5 Chapter Summary

A central claim of stylistics is that there are objectively described patterns which make authorship attribution possible or at least permit the formation of hypotheses that can be pursued by other means (Goutsos, 1995). Authorship attribution has been undertaken on various literary works by finding traits or features that are unique for a particular author. Authorship attribution is a research field of long standing in the humanities but has only recently begun to be applied to CMC.

As this is a study in forensics, the set of possible authors for an e-mail message can not always be immediately confined to a small number of suspected authors. Initially the data set may be quite large. It will typically be necessary to select a smaller number of authors, out of which the best match can be determined. To accomplish this, other linguistic techniques such as authorship characterisation using sociolinguistic filters might have to be applied to limit the suspected author set.

Many authorship attribution studies to date have used only one or a small number of features. All of these features are worth including in this project's initial phase. However, it is expected that more than a few measures will be required to discriminate between authors.

By using as many features as can be identified from a number of areas of linguistics and by using the Support Vector Machine as a classifier for pattern matching, this research should improve the prospects of finding the pattern to distinguish between authors.

The major questions that this research has addressed are as follows:

- Can the work that has been carried out in stylometry research based on literary works be applied to the text contained in e-mail messages?

- What are the best features for authorship attribution for text from e-mail messages?
- How many words are required in e-mail messages to make authorship attribution successful?
- Is there some way of reducing a large list of suspected authors, e.g. 50, to a smaller list of authors, e.g. 5, for selection of one author most likely to have written an e-mail message?

It is necessary to define some level of 'success' for the experiments that will be undertaken. This level should be significantly better than the level of chance. The weighted macro-averaged F_1 value will be used as the ultimate measure for the experiments performed (see Section 3.1.3 for an explanation of F_1). Where possible, the same data set will be used for testing various parameters of the experimental sequence. It will be possible to make comparisons of weighted macro-averaged F_1 values between different experiments to see whether the parameters had a positive or negative influence on the classification of authorship. Ultimately, it is hoped to achieve a weighted macro-averaged F_1 value of 85% or more from the technique.

This research will not attempt to define any new stylometric features. It will instead be limited to finding the best features from those used previously and to use those features with the Support Vector Machine machine learning algorithm to attribute authorship of e-mail messages from the text within those messages.

The SVM has been shown to be a good tool for text classification. Based on this evidence it is thought to be the most promising tool for authorship attribution. Comparisons with other machine learning algorithms will not be undertaken in this thesis.

The language of the data used for any experiments will be limited to English.

Chapter 3 will outline the approach that will be taken to determine whether authorship attribution will be possible with e-mail message text.

Chapter 3

Authorship Analysis and Characterisation

Chapter 2 demonstrated that it is possible to attribute authorship of text using stylometry even though a consensus of opinion on the best measures of an author's style and the best method for discrimination has not been reached. One of the questions that this work must answer is whether or not stylometry is applicable to the attribution of authorship of e-mail messages. Even though the length of e-mail messages is usually less than the suggested minimum word count required for stylometry (1000 words), it is likely that elements of style can be measured in them. With a sophisticated discrimination technique, attribution of authorship of e-mail messages should be able to be achieved.

Much research carried out into non-traditional authorship attribution has come to rely on stylistics and stylometry, applying various numerical techniques such as principal component analysis, and computational techniques such as neural networks. Most of this work has been performed with techniques that can handle only a limited number of features, exposing the problem that one can never be sure whether or not the best features have been selected.

Section 2.4 showed that the Support Vector Machine (SVM) is a good classifier for the analysis of text when attempting to classify its content, because it can be used with many features with less chance of overfitting the data than neural networks.

SVMs have been successfully used in text classification experiments with 10000 or more features (Joachims, 1998), indicating that the technique is applicable to problem domains which use many features. Previous research in authorship attribution has proposed many different types of features to discriminate between authors, but each separate study has relied on only a small number of features or features of one type. This project will use the Support Vector Machine as its classification technique and a combination of many features for discriminating between authorship classes.

E-mail messages contain not only text but also structural formatting features under the author's control. The text based features proposed for this analysis are culled from previous stylometric work. Candidate structural features were sourced from the work done on software forensics, as discussed in Section 2.2.4, where many of the metrics are related to the layout of the source code under investigation. Other possible structural features were derived from the various choices that individuals make when writing a message, e.g. whether or not they include a greeting or format a word or sentence in some particular way.

The issues addressed in this chapter include:

- setting a framework for the conduct of classification experiments using the Support Vector Machine
- choosing candidate stylistic features for solving the problem of classification of authorship of an e-mail message
- determining an experimental sequence for testing whether or not classification of authorship of e-mail messages will be successful

This chapter introduces the experimental work which was undertaken in this research. Section 3.1 discusses how machine learning experiments were conducted and

outlines the relevant performance measures for classification experiments in general. Section 3.2 gives an account of the stylometric features drawn from the literature and how the best features for classification and attribution of authorship of text were determined. Section 3.3 discusses how a baseline of authorship analysis using the SVM was established, and how this transferred to authorship analysis of e-mail messages. The baseline was established using plain text, i.e. not e-mail text, in an attempt to determine the applicability of the techniques being developed to e-mail messages. Section 3.4 discusses how features other than those stylometric ones used in the baseline experiments can be drawn from e-mail messages and how they could impact on the classification efficiency. Features possibly characterising the sociolinguistic classes of an author are discussed in Section 3.5. The data sources and collection methods are discussed in Section 3.6, and Section 3.7 summarises the chapter's contents.

3.1 Machine Learning and Classification

3.1.1 Classification Tools

The basic tool for classification of authorship selected for this project was the Support Vector Machine (Vapnik, 1995). As discussed in Section 2.4, SVM is recognised as a good tool for text classification and is suitable for classification tasks where there are a small number of data points and a large number of features. The SVM technique is not affected by the sparseness of the feature vectors. The SVM used in this research was SVM^{light} version 3.5, implemented by Thorsten Joachims, which is freely available for research purposes (Joachims, 1999).

The SVM requires a set of training data as the basis for learning. It generates a model, which is used to generalise about or classify other data that is not included in

the training set. The input to train an SVM is a set of training vectors $\{\vec{v}\}$, where $\vec{v} = (\vec{x}, y)$, consisting of a feature vector $\vec{x} \in \Re^N$ and a label for the class number $y \in \{-1, +1\}$ indicating the feature vector's classification as a positive or negative example. An example of some input training data is shown in Figure 3-1.

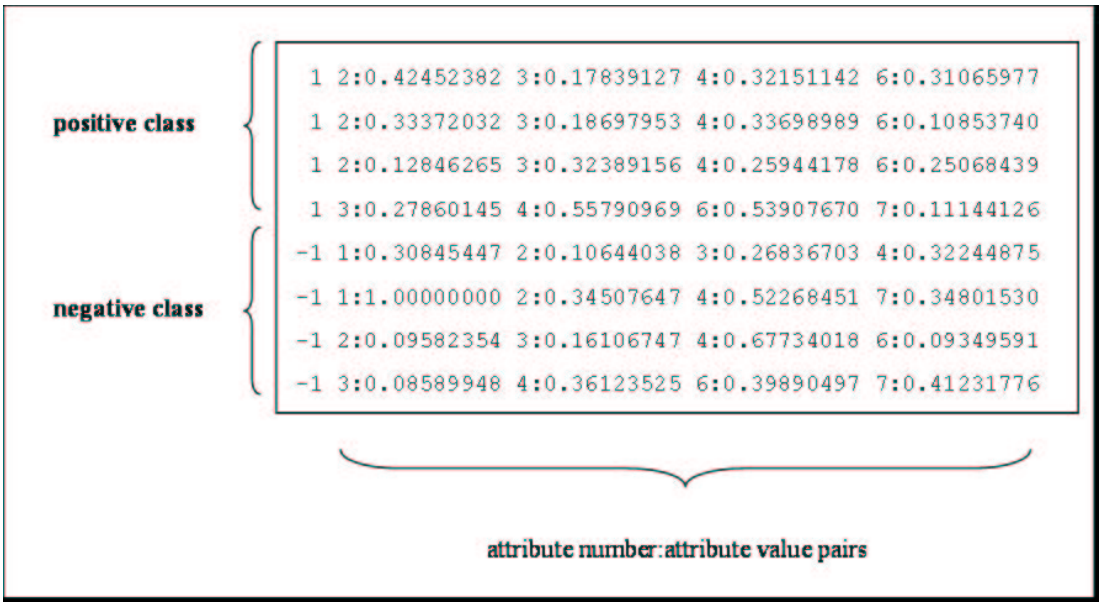


Figure 3-1: Example of Input or Training Data Vectors for SVM^{light}

The input data is fed into the learning module of SVM^{light}, where the decision boundary between the two classes is determined and the support vectors defining the boundary are identified and written to a classification model file.

Unseen test data is analysed for the same features as the training data. Test data vectors are constructed in the same manner as for the input data, except that the class number is not required. Each test data vector is then classified using the learnt model.

The output from the classifier is a single number for each data vector classified. The sign of the output number indicates which class the data vector has been classified into, positive or negative. The magnitude of the output number is an indication of the

confidence of the decision that has been made by the classifier (T. Joachims, personal communication, October 20, 2000). An example of some output data is shown in Figure 3-2.

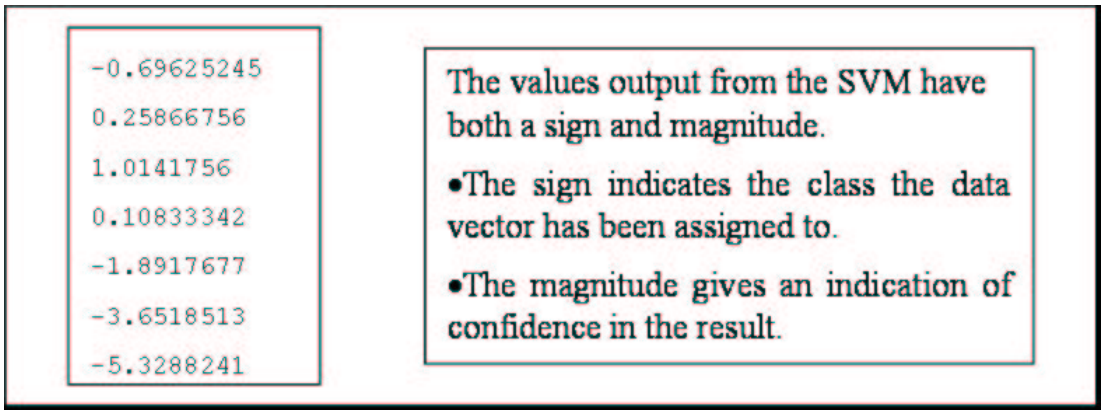


Figure 3-2: Example of Output Data from SVM^{light}

3.1.2 Classification Method

The SVM provides the ability to perform classification of binary class problems only. Forensic attribution of authorship would not, in most cases, be restricted to a selection between only two authors. There are methods, however, for performing multi-class classifications using binary class classifiers. These methods are 'one against all' classification and 'one against one' classification.

In the 'one against all' method, one of the classes is made the positive class and all others are joined to make the negative class. For n classes, n classifier models are learned with each class being made the positive class for its corresponding classifier. Figure 3-3 shows this process when $n = 4$. When this approach is used, the unseen test data will be classified with all n classifier models.

A test data vector, in an ideal classification, will be classified correctly and the output from each classifier will indicate which class it belongs to. If it belongs to one of the authors in the classifier scheme this should be indicated with one positive result and $n - 1$ negative results. If the data vector does not belong to one of the authors in the classifier scheme, this should be indicated with n negative results.

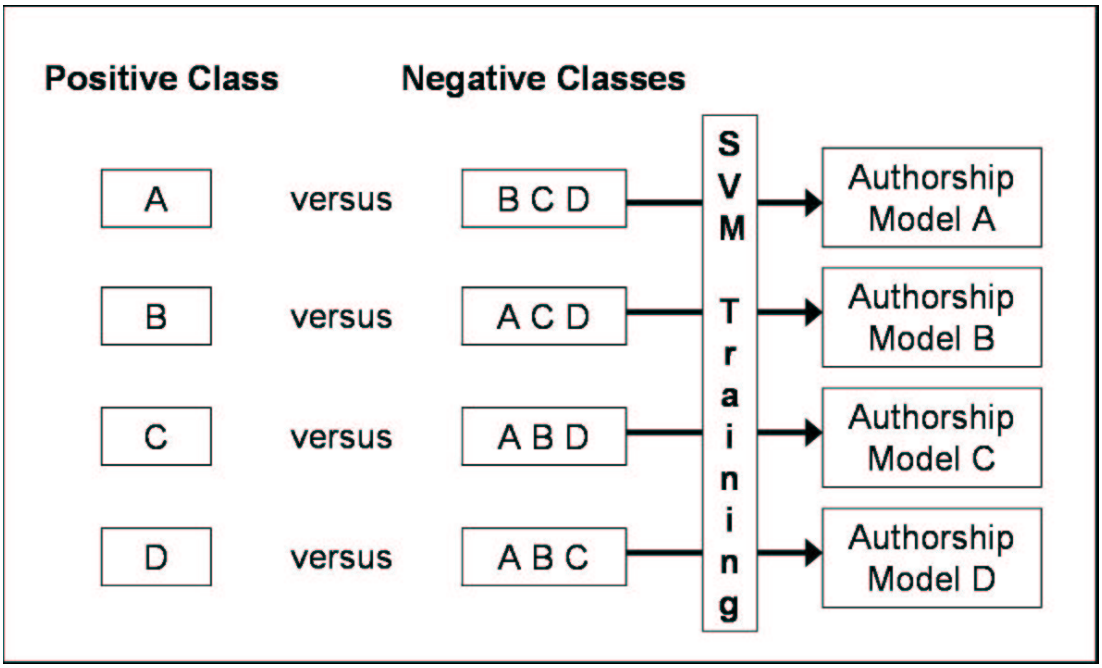


Figure 3-3: 'One Against All' Learning for a 4 Class Problem

In the 'one against one' classification method, only two out of the n classes of data are used to produce a classifier. Each class is paired off with every other class to produce $n(n-1)/2$ classifier models. Figure 3-4 shows this process when $n = 4$. With this classification method, the test example being classified is classified with all classifier models. Each classifier model predicts that it belongs to one of the two classes in that particular classifier model. This prediction is made even if the test example belongs to neither of the classes in the classifier model being used. 'One

against one' classification requires a voting technique to determine which class the test data vector belongs to. The class which receives the most votes is selected as the class to which the test example belongs.

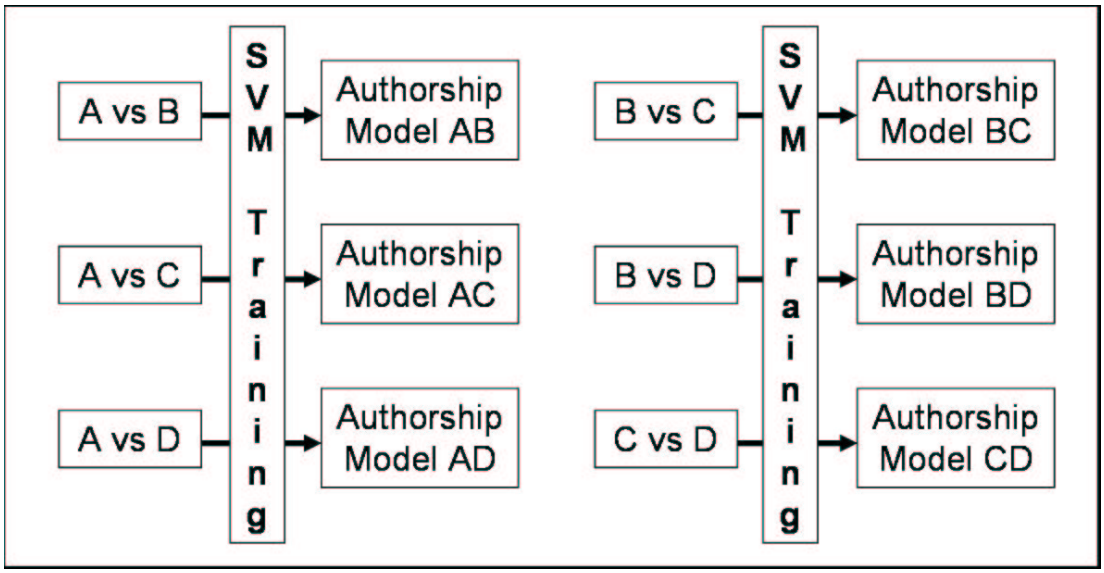


Figure 3-4: 'One Against One' Learning for a 4 Class Problem

The complexity of 'one against all' classification is $O(n)$ with respect to the number of authorship classes being classified for both the learning and classifying phases. In comparison, the 'one against one' classification method is $O(n^2)$. When classifying test data, an authorship class will be assigned regardless of the class of the data. When using either multi-class approach, if a classifier model is learned from authors A and B and the test data does not belong to one of these two classes, a false positive result will be generated no matter which class the classifier selects. When using the 'one against all' approach, there are fewer decisions to be made about which authorship class the test data belongs to, even if it does not belong to any of

the authorship classes in the classifier model scheme. To avoid this computational overhead, the approach used in this project was 'one against all'.

3.1.3 Measures of Classification Performance

There are various measures of classification performance (Witten and Frank, 2000) that can be calculated for each classification experiment. These familiar machine learning measures include:

- Error Rate (E)
- Precision (P)
- Recall (R)
- F_1 combined performance measure

These measures are explained in more detail below. To calculate each of these measures, it is necessary to assign the result of each classification to one of the following four types of result:

- **a true positive result (*TP*)** - the classifier has identified a positive class data point as positive
- **a false negative result (*FN*)** - the classifier has identified a positive class data point as negative
- **a false positive result (*FP*)** - the classifier has identified a negative class data point as positive
- **a true negative result (*TN*)** - the classifier has identified a negative class data point as negative

From the frequencies of these results, a two-way confusion matrix can be constructed as shown in Figure 3-5.

		Predicted Class	
Actual Class		*Yes*	*No*
	Yes	true positive	false negative
	No	false positive	true negative

Figure 3-5: Construction of the Two-Way Confusion Matrix

The error rate is defined as follows:

$$ErrorRate\ (E) = \frac{FP + FN}{TP + FP + FN + TN}$$

Precision and recall measures are taken from the area of information retrieval and are defined as follows:

$$Precision\ (P) = \frac{TP}{TP + FP}$$

$$Recall\ (R) = \frac{TP}{TP + FN}$$

It can be seen from these formulae that, as the number of false positives tends to 0, the precision approaches 1, and as the number of false negatives tends to 0, the recall approaches 1. Obviously, the fewer errors made, either false positive or false negative, the better. The F_1 value combines the precision and recall values into a single value by calculating the geometric mean as follows:

$$F_1 = \frac{2 \times R \times P}{R + P}$$

Each of these measures can be calculated for each authorship class.

To get an indication of the overall success of a classification experiment, the macro averaged error rate and F_1 measure (Yang, 1999) can be calculated across each authorship class using the following formulae:

$$E^{(M)} = \frac{\sum_{i=1}^{n} E_{AC_i}}{n}$$

and

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

where AC_i is the authorship class ($i = 1, 2, \ldots, n$) and n is the number of authorship classes.

To compensate for document frequency, the statistics for each authorship class are inversely weighted by the number of data points in each class (de Vel, 2000). The weighted macro-averaged error rate ($\overline{E}^{(M)}$) and weighted macro-averaged F_1 ($\overline{F}_1^{(M)}$) can then be calculated using the formulae:

$$\overline{E}^{(M)} = \frac{\sum_{i=1}^{n}(1 - w_{AC_i})E_{AC_i}}{n-1}$$

and

$$\overline{F}_1^{(M)} = \frac{\sum_{i=1}^{n}(1 - w_{AC_i})F_{1,AC_i}}{n-1}$$

where w_{AC_i} is the document frequency weight which is calculated using the following formula:

$$w_{AC_i} = \frac{N_{AC_i}}{\sum_{i=1}^{n} N_{AC_i}}$$

and N_{AC_i} is the number of documents in authorship class AC_i, $i = (1, 2 \ldots, n)$ and n is the number of authorship classes.

The weighted macro-averaged error rate and weighted macro-averaged F_1 measure were used to make comparisons between experiments so that increases or decreases in classification efficiency could be measured as the variables in a sequence of experiments are altered.

3.1.4 Measuring Classification Performance with Small Data Sets

When the classification problem has only a small set of data to work with, it can be difficult to provide enough data for disjoint training and testing sets. This is likely to be the case in a forensic study, where there may only be a small amount of data to produce a model of authorship.

In these cases it is possible to use a technique known as k-fold cross validation (Stone, 1974) and (Geisser, 1975) to provide a more meaningful result by using all of the data in the data set as both training and test data. In this technique, the data is split into k folds, which are as equal in size as possible.

A set of classifiers is then learnt from $k - 1$ folds of data, with the remaining fold being used as the test set. This procedure is then repeated so that each fold is held out for testing. The results of the classifications from the k tests are combined to calculate the overall results for the data set. Most commonly, k is set equal to 10.

When the folds are being created, random sampling is performed, so that some classes may be under-represented or over-represented in some folds. The folds can be stratified by sampling from within each class in turn until that class is exhausted. This process produces randomly sampled folds that have a distribution that more precisely mirrors the distribution of classes in the whole data set. This latter process is known as stratified 10-fold cross validation (Kohavi, 1995). Stratification and cross validation can be applied to binary class and multi-class classifications.

Figure 3-6 shows an example of how data could be randomly distributed among k folds with $k = 3$ for the simplicity of the demonstration. In this example Class A has 5 feature vectors, Class B has 6, and Class C has 8. It is not necessary for each class to contain the same number of feature vectors. The feature vectors from Class A are randomly sampled without replacement and distributed to the three folds in turn, i.e. Fold 1, Fold 2, Fold 3, Fold 1, ... When Class A is exhausted, Class B and then Class C are distributed. This process ensures that each fold has a similar representation of feature vectors from the classes involved, and that this distribution is also similar to the data set as a whole.

When k-fold stratification is combined with 'one against all' classification for a multiclass problem, the number of classifier models produced is $k \times n$ where k is the number of folds and n is the number of classes. An illustration of holding out test folds for the above example where $k = 3$ and $n = 3$ is shown in Figure 3-7.

In this example, three training sets are constructed by combining the two folds not used for testing. As each training set contains three classes, three models will be learnt for each training set. The data in the test set, i.e. the fold not involved in training, will be classified using each of the three models from the corresponding training set and

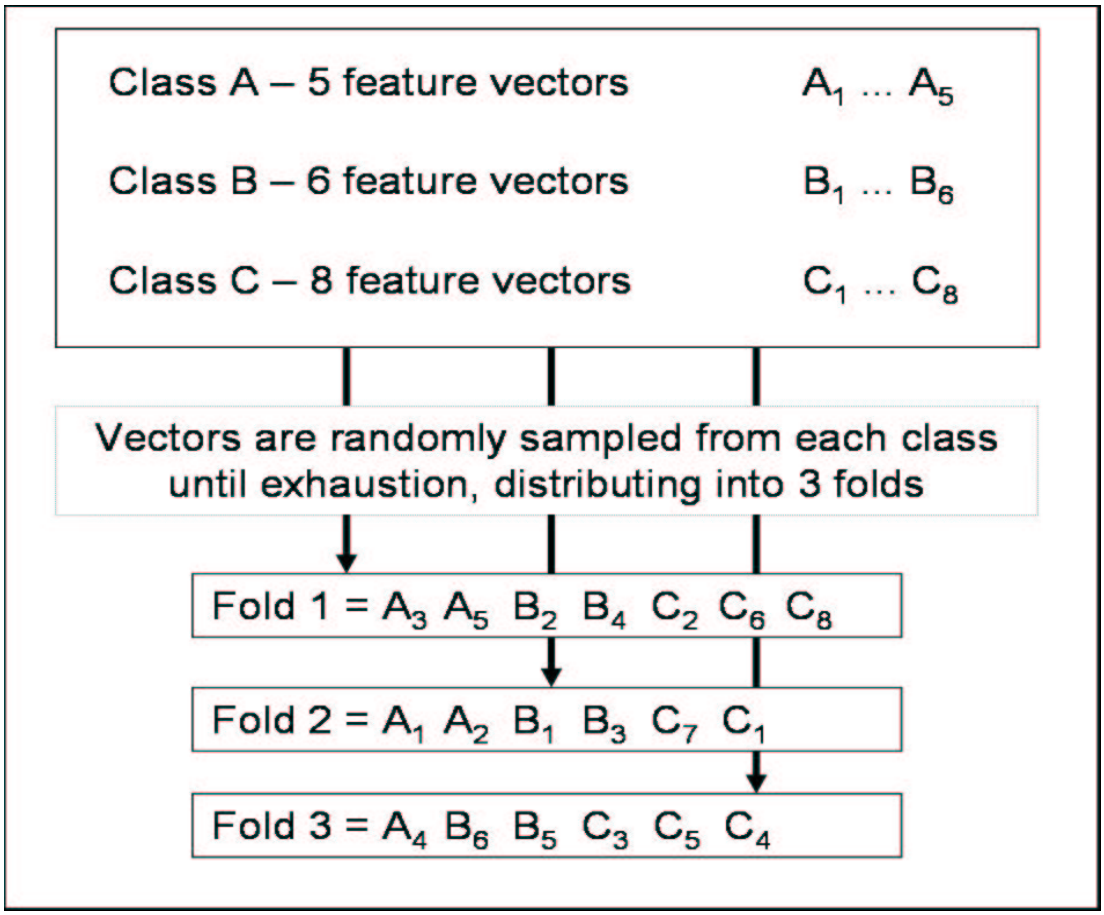


Figure 3-6: An Example of the Random Distribution of Stratified *k*-fold Data

the results of each classification will be collated to produce the combined confusion matrix.

This technique can be used to provide measures of classification efficiency for a set of data points. In a forensic investigation, these measures could be used to tune the models with known data before attempting to attribute unknown test examples.

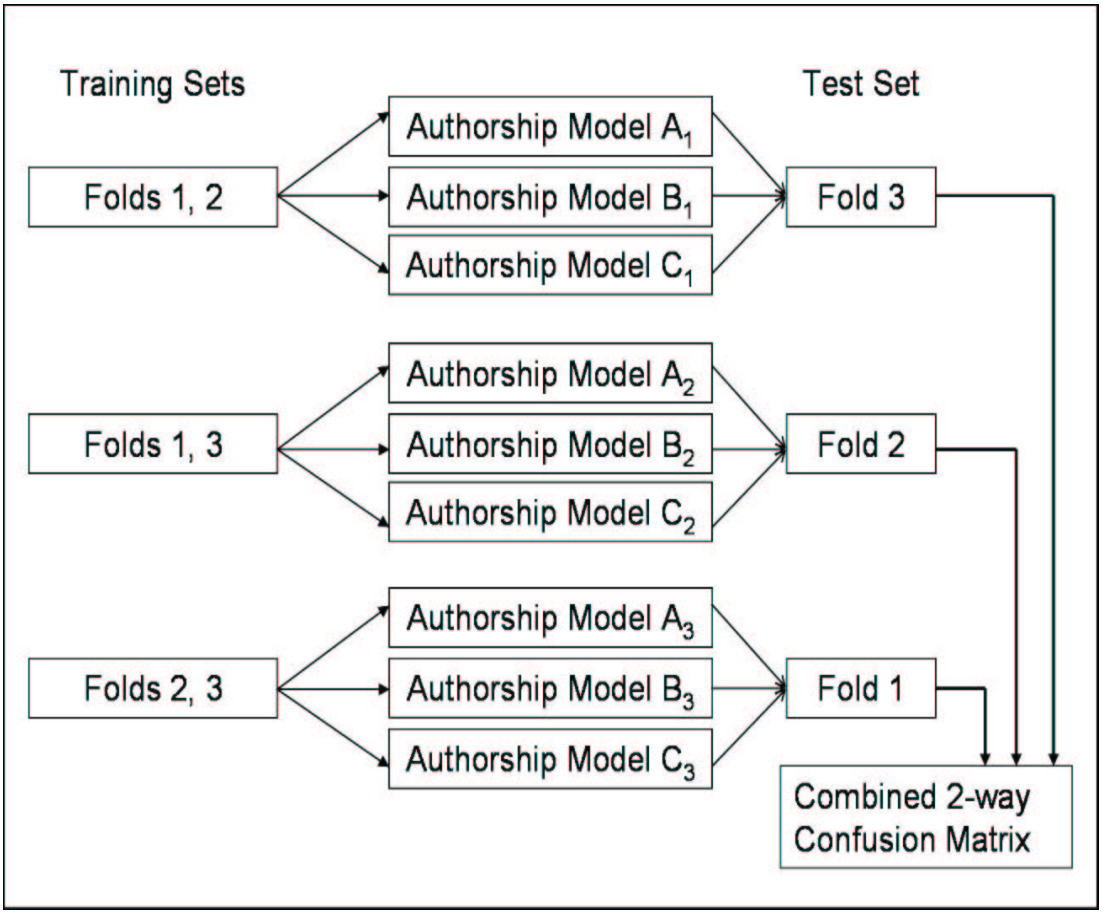


Figure 3-7: Cross Validation with Stratified 3-fold Data

3.2 Feature Selection

The attribution of authorship of an e-mail message would only be possible if a set of discriminatory features could be identified. The features identified from previous authorship attribution and stylometric research as outlined in Chapter 2 have been grouped into sets of features with a similar level of granularity, e.g. document, word and character levels. This was a novel approach to the selection of features for text or authorship classification. This approach fits well with the SVM classifier as the latter is able to handle large numbers of features.

Methods for measuring the effectiveness of feature sets with respect to each other and for determining which features were used for the classification of e-mail are discussed in detail in Section 3.1.3. To determine the best set of features to use, initial tests were based on data that is not e-mail and will investigate the effectiveness of each of the feature sets listed here, individually and in combinations with one another.

The sets of features include but are not limited to:

Document based features The candidate document based features are the average sentence length and the proportion of blank lines.

Word-based features The number of different words (types) and the total number of words (tokens) can be counted to calculate a type:token ratio. The number of words used once (*hapax legomena*) or twice (*hapax dislegomena*) can be counted. A set of metrics (Tweedie and Baayen 1998) based on the values of types and tokens in a document were also used as word-based features. This feature set is displayed in Table 3.1. A full definition of each of the lexical measures is included in Appendix A. In Table 3.1 N is the total number of

words in the document and V is the total number of different word types in the document.

Character-based features These features will include the frequencies or proportions of white space, digits, upper case characters and punctuations. The character based feature set is shown in Table 3.2. In this table, C is the total number of characters in the document.

Function word ratios The frequency of each of a set of function words can be calculated. A list of 120 common function words sourced from Craig (1999) will be used. The list of function words is shown in Appendix A.

Word length frequency distribution The frequency distribution of word lengths across a document can be determined and the relative frequency of each word length can be used as an individual feature. Word lengths between 1 and 30 will be considered.

Collocation Frequencies The frequencies of various collocations of words can be calculated and the ratio of these frequencies to the total number of occurrences of the first or base word in the collocation can be used as feature values.

Letter 2-grams The frequency of letter 2-grams can be calculated and ratioed with the total number of 2-grams in a text. It was noted in Section 2.1.4 that Kjell (1994b) suggested that the 50 or 100 most significant 2-grams should be used for classification. In this project, all possible 676 letter 2-grams were used as features, as the SVM is able to work with a large number of features.

A problem for some of the feature sets is that there is a great variety in the text size or word count between different e-mail messages. This would mean that if simple

frequencies were counted, there would be no way of finding similarities between data points from the same authorship class when one e-mail might have significantly more words in it than another. For these feature sets it is necessary to normalise the chosen features in some manner. Features that rely on counting words should be normalised against the total number of words N in an e-mail message. Character-based features should be normalised against the total number of characters C in the text of the e-mail message. This is also necessary for text chunks sampled from plain text documents even when the number of words in each text chunk is constant.

Feature Number	**Feature Description**
W_1	Average word length
W_2	Vocabulary richness i.e., V/N
W_3	Total number of function words/N
W_4	Total number of short words/N (word length ≤ 3)
W_5	Count of *hapax legomena*/N
W_6	Count of *hapax legomena*/V
W_7	Guirad's R
W_8	Herdan's C
W_9	Herdan's V
W_{10}	Rubet's K
W_{11}	Maas' A
W_{12}	Dugast's U
W_{13}	Luk"janenkov and Neistoj's measure
W_{14}	Brunet's W
W_{15}	Honore's H
W_{16}	Sichel's S
W_{17}	Yule's K
W_{18}	Simpson's D
W_{19}	Entropy measure

Table 3.1: Word Based Feature Set

Feature Number	Feature Description
C_1	Number of characters in words/C
C_2	Number of alphabetic characters/C
C_3	Number of upper-case characters in words/C
C_4	Number of digit characters in words/C
C_5	Number of white-space characters/C
C_6	Number of spaces/C
C_7	Number of spaces/Number white-space chars
C_8	Number of tab spaces/C
C_9	Number of tab spaces/Number white-space chars
C_{10}	Number of punctuation characters/C

Table 3.2: Character Based Feature Set

3.3 Baseline Testing

In order to determine a good set of stylometric features identified by previous workers in stylometrics and authorship analysis and also to evaluate and tune SVM^{light} for optimum performance, it was decided to perform a series of baseline tests on data that was not e-mail. Two sources of data were considered. Project Gutenberg (n.d.) provides plain text versions of books online that are not protected by copyright. A corpus of novels from different authors was collected from Project Gutenberg (n.d.) for use in these initial tests. A corpus of PhD theses on Information Technology topics was also used for these initial tests. These sources were chosen as being representative of a particular genre, novel or thesis. Details of the data are given in Section 3.6.

3.3.1 Feature Selection

For the baseline testing of plain text chunks only the character based, word based, function word, word length frequency distribution and 2-gram feature sets, outlined in Section 3.2 and detailed in Appendix A, were used. This was done to limit the

amount of testing performed in the initial stages of the experimental work. These purely stylistic feature sets were tested individually and then in combinations of two and more sets to determine the best combination of sets of features for attribution of authorship on the baseline data. If the addition of a set of features to the mix caused a reduction in classification efficiency, it was considered that this feature set was not adding to the overall classification of authorship.

When the best combination of feature sets was found, this was used as a starting point for experimentation with e-mail data and the addition of e-mail specific features.

The literature on authorship classification as outlined in Section 2.1 suggests that a minimum of 1000 words should be used for discrimination of authorship so that the effect of the feature sets on classification reliability can be critically evaluated. The text data to be analysed in these experiments, therefore, was split into chunks of 1000 words for the evaluation of the best features.

3.3.2 Effect of Number of Data Points and Size of Text on Classification

With machine learning classifiers it is not possible to create a model with only one or a few data points, in this case, samples of text. Since e-mail data might be sparse, it was necessary, therefore, to determine the minimum number of data points in the form of feature value vectors. As noted in the previous section, the literature suggests that a minimum of 1000 words for each chunk of text be used for classification. E-mail messages rarely contain this number of words. It was necessary, therefore, to study the authorship classification efficiency of smaller chunks of text in comparison to the baseline level of 1000 words, to find how small a chunk can be used and still obtain a good overall result.

Using the outcomes from the experiments in Section 3.3.1 , the most reliable set of features were used. Text chunk sizes of 500, 200, 100 and 50 words were compared to the baseline level of 1000 words.

The number of document chunks was also a variable in this experiment and the number of chunks or data points was varied between 10 and 100. Chunking was limited by the size of the documents being used, as the text chunks had to be sampled independently. We expected to find that as the text chunk size was reduced and the number of text chunks lowered, the level of correctness of classification would also be reduced.

The results of the experiments carried out in this baseline evaluation are reported in Chapter 4.

3.4 Application to E-mail Messages

Following the baseline tests on text, the project moved on to e-mail data. It was necessary to determine if SVM classification of authorship could be successful when applied to e-mail messages, given that they contain a smaller amount of text and that each message contains a variable number of words.

It was possible to apply the feature evaluation and machine learning classification techniques in two ways to e-mail data.

Most intuitively, each e-mail message could be considered to be a data point in its own right in any classification experiment. Values for the selected set of features could be calculated by analysing each e-mail message separately.

Alternatively, the text from a collection of an author's e-mail messages could be concatenated and then sliced into chunks of text with an equal number of words. This approach limits the analysis of authorship to only the text within the messages.

Features that relate to the e-mail message's structure could not be included in such analyses. Both approaches were trialled in the experiments on e-mail messages.

The ultimate goal was to determine if the authorship of an anonymous e-mail message can be attributed to an author for which sufficient existing data, i.e. some minimum number of e-mail messages which have a sufficient number of words in them, has been collected for purposes of building a model of that author's overall authorship style.

3.4.1 E-mail Structural Features

E-mail is produced as a result of a user's interaction with some program which creates an e-mail message. These programs or User Agents (UA) produce e-mail messages in a format that must adhere to the *de facto* standard as outlined in RFC822 (Crocker, 1982). While the format is not consistent between different UAs, there are certain pieces of information which are included in all e-mail messages. These include a set of header attributes and header values. The headers are separated from the text of the e-mail message by a single blank line and can be considered to be metadata. The metadata will not be considered in this study as it is mostly generated by the UA and not by the author of the e-mail message.

An e-mail message may be an original message, a reply message or a forwarded message. This can be evaluated simply and used as a message type feature by giving a discrete feature value of 0, 1 or 2 for the three possible options listed. These possibilities are captured in feature E_1 in the following.

An example of an e-mail message is shown in Figure 3-8 which shows that an e-mail message may include greeting or salutation text, farewell text, a signature and one

```
Return-Path: <b.hodgkiss@qut.edu.au>
Received: from tu01m3.qut.edu.au (mail.qut.edu.au [131.181.127.183])
        by pigeon.qut.edu.au (8.9.3/8.9.3) with ESMTP id PAA22973
        for <corneym@pigeon.qut.edu.au>; Fri, 28 Jun 2002 15:53:46 +1000 (EST)
Received: from DIRECTORY-QUT by mail.qut.edu.au (PMDF V5.2-33 #46659)
 id <0GYE12Y01J1LV7@mail.qut.edu.au> for corneym@pigeon.qut.edu.au
 (ORCPT rfc822;m.corney@qut.edu.au); Fri, 28 Jun 2002 15:53:45 +1000 (EST)
Received: from hr-ss-bh.qut.edu.au
 (s153h32.hr.adms.qut.edu.au [131.181.153.32]) by mail.qut.edu.au
 (PMDF V5.2-33 #46659) with ESMTPA id <0GYE12VFPJ1LVP@mail.qut.edu.au>; Fri,
 28 Jun 2002 15:53:45 +1000 (EST)
Date: Fri, 28 Jun 2002 15:52:50 +1000
From: Brian Hodgkiss <b.hodgkiss@qut.edu.au>
Subject: Re: League scrums
In-reply-to: <5.1.0.14.2.20020628144532.02f6a698@pop.qut.edu.au>
X-Sender: hodgkiss@pop.qut.edu.au
To: Malcolm Corney <m.corney@qut.edu.au>, "qut.forum."@qut.edu.au
Message-id: <5.1.0.14.2.20020628154619.020a9048@pop.qut.edu.au>
MIME-version: 1.0
X-Mailer: QUALCOMM Windows Eudora Version 5.1
Content-type: multipart/alternative;
        boundary="==========26282765==_.ALT"
References: <4.3.2.7.2.20020628142834.00a9c860@pop.qut.edu.au>
X-UIDL: bff1b730a1c6024d103cc1384998a450

Malcolm :

From your explanation it is fairly easy to divine the purpose of "feed" as in "loose head and
feed" however what is meant by a "loose head", surely this is not merely a reference to Gorden
Tallis' oratory prowess or placid temperament? Could you enlighten us on the (dare I say it)
state of origin of this bizarre terminology please?

Cheers

Brian

At 02:53 PM 28/06/2002 +1000, you wrote:

> Hi Eugene,
>
> This may or may not be a good idea depending on the use of the scrum.  Prior to the rule
> changes where loose head and feed was given to the team feeding the scrum and when there was
> a requirement for the ball to be fed in the middle of the front row, there was a skill
> required by the hooker in winning the ball.
>
> Since this time however, there is still a good reason (IMHO) for having the scrum.  It
> provides the attacking team, in most cases the team feeding the scrum, with the opportunity
> to use their back line and have only 7 defenders to beat instead of all 13.  So the team who
> committed the error (knock on) or who wins the feed due to a well placed 40-20 kick gets the
> opportunity to go on the attack.
>
> I will never forget the scrum move where Steve Renouf was passed the ball after a scrum win
> to Brisbane and ran 80 odd meters in the Grand Final of 92 or 93 when Brisbane played St
> George.  Pure joy.  If all 13 players are able to defend after they have made a mistake,
> there is less emphasis on ball security.
>
> I say bring back the scrum for 6th tackle mistakes instead of the hand over!
>
> Mal Corney
>
>> At 02:28 PM 28/06/2002 +1000, Eugene Ross wrote:
>>
>> I know very little about rugby league but do know that hookers in a scrum are supposed to
>> 'hook' for the ball when it is thrown in the middle of the tunnel. Why bother with a scrum
>> when the ball is no longer being thrown in the middle? Why does the League persist with
>> this idiotic nonsense? I say abolish the scrum.
>> ER
>>
---
Outgoing mail is certified Virus Free.
Checked by AVG anti-virus system (http://www.grisoft.com).
Version: 6.0.338 / Virus Database: 189 - Release Date: 14/03/2002

--------------------------------------------------------------------------------
Bbbbb Hhhhhhhh
SUPERANNUATION - Human Resources Dept.   PHONE  : +61 7 5555 9999
Queensland University of Technology        FAX  : +61 7 5555 8888
Gardens Point Campus
E-MAIL     : B.Hhhhhhhh@qut.edu.au

SNAIL      : 2 George Street,
             Brisbane
             Australia  4059
--------------------------------------------------------------------------------
```

Figure 3-8: Example of an E-mail Message

or more attachments. The presence or absence of the text items above and the count of the number of attachments can be used as e-mail structural features ($E_2 \ldots E_5$).

If the e-mail message is a reply message it may contain portions of text that are requoted from the original or reply message being replied to. There are different ways that an author may choose to include this requoted text. Some authors place all of the requoted text at the beginning or the end of the message, whilst others intersperse their reply among the requoted text. The position of requoted text was identified as a feature of interest by de Vel (2000). Here, this concept has been extended, as seven possible combinations of original text (O) and requoted text (R) have been identified. A discrete value has been assigned for each possible combination and this value will be used in the feature vector as feature E_6, on the assumption that an author may have a preferred method of replying to an e-mail message. Table 3.3 outlines the different ways that an e-mail message can be composed from original and requoted text. Forwarded messages generally contain no original text from the author and cannot be analysed further.

Combination	**Assigned Value**	**Explanation**
no text	0	no text
O	1	original text only
R	2	requoted text only
OR	3	original text followed by requoted text
RO	4	requoted text followed by original text
ORO ...	5	original and requoted text interspersed - original first
ROR ...	6	requoted and original text interspersed - requoted first

Table 3.3: Possible Combinations of Original and Requoted Text in E-mail Messages

Different UAs use different methods to indicate that portions of an e-mail message are requoted. Requoted text in a plain text e-mail message is usually marked with a

Feature Number	Feature Description
E_1	Reply status
E_2	Has a greeting acknowledgement
E_3	Uses a farewell acknowledgement
E_4	Contains signature text
E_5	Number of attachments
E_6	Position of re-quoted text within e-mail body

Table 3.4: List of E-mail Structural Features

'$>$' symbol at the start of each line of the requoted portion of text. Some UAs use a different character, such as a ':' or '$|$' to indicate requoted text.

The latest generation of UAs now employ HTML formatting elements to format e-mail message text and the metadata indicates whether or not the e-mail message contains HTML formatting. These UAs use a HTML tag pair:

```
<BLOCKQUOTE> ... </BLOCKQUOTE>
```

to indicate requoted text. When parsing an e-mail message for analysis of the value of its features, it will be simple enough to ensure that only original text is analysed, and also to determine which combination of original and requoted text is being used.

A set of e-mail structural features was compiled and is shown in Table 3.4.

To aid in the parsing of various structural elements from each e-mail message an e-mail grammar was proposed. This grammar is shown in Figure 3-9.

3.4.2 HTML Based Features

As mentioned above, many e-mail user agents compose e-mail messages as HTML documents and embed HTML tags for text formatting. The HTML tags had to be removed from the text before the text could be stylistically analysed. If authors have

```
Email = Header [Body].
Header = HeaderLine {HeaderLine}.
HeaderLine = FieldName ":" FieldContent.
Body = Content {Attachment}.
Content = NewLine [Greeting] Message [Farewell] [Signature].
Message = [OriginalText] {RequoteText OriginalText} [RequoteText].
RequoteText = RequoteCharacter string | RequoteOpeningTag string RequoteClosingTag.
RequoteCharacter = ">" | "|" | ":".
RequoteOpeningTag = "<BLOCKQUOTE>".
RequoteClosingTag = "< /BLOCKQUOTE>".
NewLine = cr | cr lf | lf.
FieldName = string.
FieldContent = string.
OriginalText = string.
Attachment = string.
Greeting = string.
Farewell = string.
Signature = string.
```

Figure 3-9: E-mail Grammar

a hard wired style, it is plausible that text formatting and layout is also a part of this style. Some authors may use different formatting elements preferentially over others and this could be used in conjunction with other text based features to discriminate between the authorship of e-mail messages. HTML tag pairs related to formatting the text and under the control of the author e.g. bold, italics, colour, font, font size etc., can be counted, normalised and used as a feature set.

The list of HTML tag features that have been proposed for use in attribution of e-mail message authorship is shown in Table 3.5. In this table, H is the total number of HTML tags in the document.

3.4.3 Document Based Features

Two document based features were added to form the document based feature set. The definition of these features is shown in Table 3.6. These features were the average

Feature Number	Feature Description
H_1	Frequency of <BIGGER> / H
H_2	Frequency of <BOLD> or <B> / H
H_3	Frequency of <CENTER> / H
H_4	Frequency of <COLOR> / H
H_5	Frequency of <FONT> / H
H_6	Frequency of <ITALIC> or <I> / H
H_7	Frequency of <UNDERLINE> or <U> / H

Table 3.5: List of HTML Tag Features

Feature Number	Feature Description
D_1	Average sentence length (number of words)
D_2	Number of blank lines / total number of lines

Table 3.6: Document Based Feature Set

sentence length and the ratio of blank lines to total number of lines in the body of a message.

3.4.4 Effect of Topic

If the classification of authorship of e-mail messages is to be successful, it will be necessary to investigate whether this classification is affected by the topic of the message. It would be of little practical forensic use to limit training data to messages on the same topic. It is possible that the topic of the messages could affect some features, as topic words will belong to the author's regular vocabulary, or could be badge words.

To show that classification is not affected by topic, it was necessary to obtain a corpus of e-mail messages from a small group of authors writing e-mail messages on a limited set of disparate topics. This corpus had to contain some minimum number

of e-mail messages from each author, and each e-mail message had to contain a sufficient number of words, as determined by the baseline experiments discussed in Section 3.3.2, to make the classification meaningful.

To measure the independent effect of topic, classifier models were learnt for a limited group of authors, using only the messages from one of the topics. Messages from the other topics were then used as the test data for the learnt classifier models from the original topic. Generalisation performance of these classifiers was measured using the same series of performance indicators as before, except that there was no need to perform k-fold stratification on the data set, as the test messages were not used for training.

3.5 Profiling the Author - Reducing the List of Suspects

In a forensic investigation of authorship, the possible list of suspects may be quite numerous. The experiments suggested in Section 3.4 are aimed at selecting one author from a small group of authors. The size of the small group could range from two to possibly as many as ten authors.

There is a clear forensic need to have some way of reducing a large list of suspect authors to a manageable number. Experiments are proposed here to investigate the possibility of filtering or profiling the suspect list to produce a smaller number of authors, from which an anonymous e-mail or e-mails can be classified. This falls under the heading of authorship characterisation as discussed in Section 2.1

An alternative approach is to perform authorship analysis as suggested in Section 3.4 on arbitrarily selected small groups, narrowing down the number of possible suspect authors by keeping each author identified as a possible positive match, whether

it is a true positive or a false positive identification, in the suspect list. With this approach there would need to be multiple iterations through the suspect list until a small enough group of suspects is formed to be able to identify the best single suspect. This approach is shown diagrammatically in Figure 3-10.

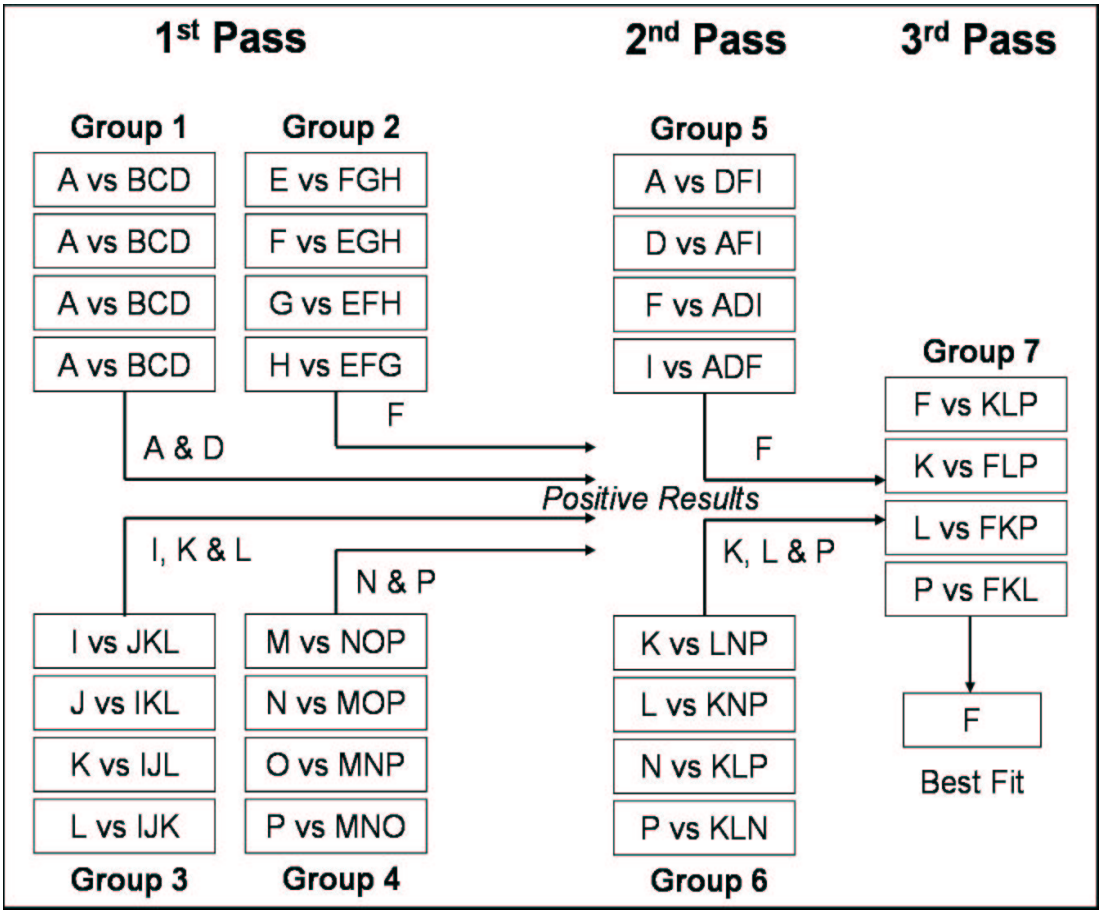


Figure 3-10: Reducing a Large Group of Suspects to a Small Group Iteratively

3.5.1 Identifying Cohorts

As discussed in Section 2.3 some common sociolinguistic groups can be identified as follows:

- Gender
- Age group
- Education level
- Language background

In most cases these characteristics can be classified into a small number of cohorts e.g. male and female for the gender characteristic and English as native or second language for the language background characteristic. These simple groupings lend themselves to machine learning with SVM, which performs binary classifications.

This raises the prospect of later experiments, if this simple approach is successful, of breaking up authors who have English as a second language into cohorts that have a similar language background. For example European languages are quite different from Asian languages and it is to be expected that authors with a European heritage will write English differently to those with an Asian heritage.

In this body of research, we examined gender classification first, as there was some indication in the literature (Thomson and Murachver, 2001) of gender specific features as discussed in Section 2.3.1.

As the available e-mail data provided clues to the language background of the authors, it was decided to also trial the technique on the language background characteristic as a second sociolinguistic classifier.

3.5.2 Cohort Preparation

A corpus of e-mail messages was prepared by collecting personal e-mail over an eighteen month period. This gave the data set referred to here as *inbox*. Individual messages in the data set were then designated as suitable or unsuitable for use

by manually checking each e-mail message. Messages designated as unsuitable included messages that were requoted jokes or stories that masqueraded as original text messages i.e. they did not contain requoted text. Similarly, where messages were notifications of seminars, abstracts were included in these messages that would not have been written by the author of the message. Cleaning the *inbox* data set was a lengthy but worthwhile endeavour as it added confidence to the results to be generated from the experiments performed.

The gender and language background of over 800 authors contributing to the *inbox* data set was investigated using public domain knowledge. Assigning the correct gender to the authors was relatively simple. Common sense also prevailed, with many authors' first name being obviously male or female. Where any doubt remained as to the gender of the author, this author's e-mail messages were removed from the data set.

Language background was assigned in many cases on personal acquaintance with the author. In other cases, the author's e-mail address, showing e.g. a European country's domain and name, was used as an indicator of whether English was a first or second language for the author. There were many cases where, even though the author had a typically non-English name, he/she was a resident of Australia and an assignation of language background simply could not be made. Where doubt existed, the language background of the author was not assigned and this author's e-mails were not used for language background tests.

The classification of these sociolinguistic features resulted in the production of a data set referred to in the following as *cleaned inbox*. Details of the number of authors and the profiles of the e-mail messages are given in Section 3.6.

3.5.3 Cohort Testing - Gender

As outlined in Section 2.3, Thomson and Murachver (2001) suggest that there are various gender-indicating features which can be measured in electronic discourse. For women, these include the use of adverbs and adjectives, the use of the personal pronoun and a tendency to apologise. For men, gender indicating features were references to quantity and grammatical errors. The proposed set of gender specific features for cohort testing is shown in Table 3.7.

Feature Number	Feature Description
G_1	Number of words ending with *able* /N
G_2	Number of words ending with *al* /N
G_3	Number of words ending with *ful* /N
G_4	Number of words ending with *ible* /N
G_5	Number of words ending with *ic* /N
G_6	Number of words ending with *ive* /N
G_7	Number of words ending with *less* /N
G_8	Number of words ending with *ly* /N
G_9	Number of words ending with *ous* /N
G_{10}	Number of *sorry* words /N
G_{11}	Number of words starting with *apolog* /N

Table 3.7: Gender Specific Features

The experiments on gender cohorts addressed the following questions:

- Is there a minimum number of words that an e-mail message must contain so that gender can be classified?
- How many e-mail messages are required per cohort to classify gender?
- Which stylistic and e-mail structural features contribute to classification of gender?

- Do the gender specific features indicated in Table 3.7 improve classification performance?

3.5.3.1 Effect of Number of Words per E-mail Message

A number of e-mail messages in the corpus have very few words and there are some with no words at all. These messages will have an impact on the values of the measured features. Due to the low number of words in some of the e-mail messages, many features will have zero values or, at best, meaningless values that are not indicative of an author or a cohort. A lower limit of 50 words per message was applied to these experiments.

The aim of this experiment was to extract cohorts that have a set minimum number of words to determine the effect on training and ultimately on generalisation performance. Gender cohorts were produced with minimum word count limits of 50, 100, 150 and 200. The cohorts produced were randomly sampled to produce gender cohorts containing 250 and 500 e-mail messages. Stratified 10-fold cross validation experiments were conducted on these cohorts to determine the effect of minimum word count.

3.5.3.2 The Effect of Number of Messages per Gender Cohort

It is unlikely that a small number of e-mail messages or a small number of authors could define male and female gender characteristics. It was necessary to determine how many e-mail messages were required in each gender cohort to produce gender models.

Gender cohorts with minimum word counts of 50, 100, 150 and 200 words were produced. The number of e-mail messages in each cohort is shown in Section 3.6.

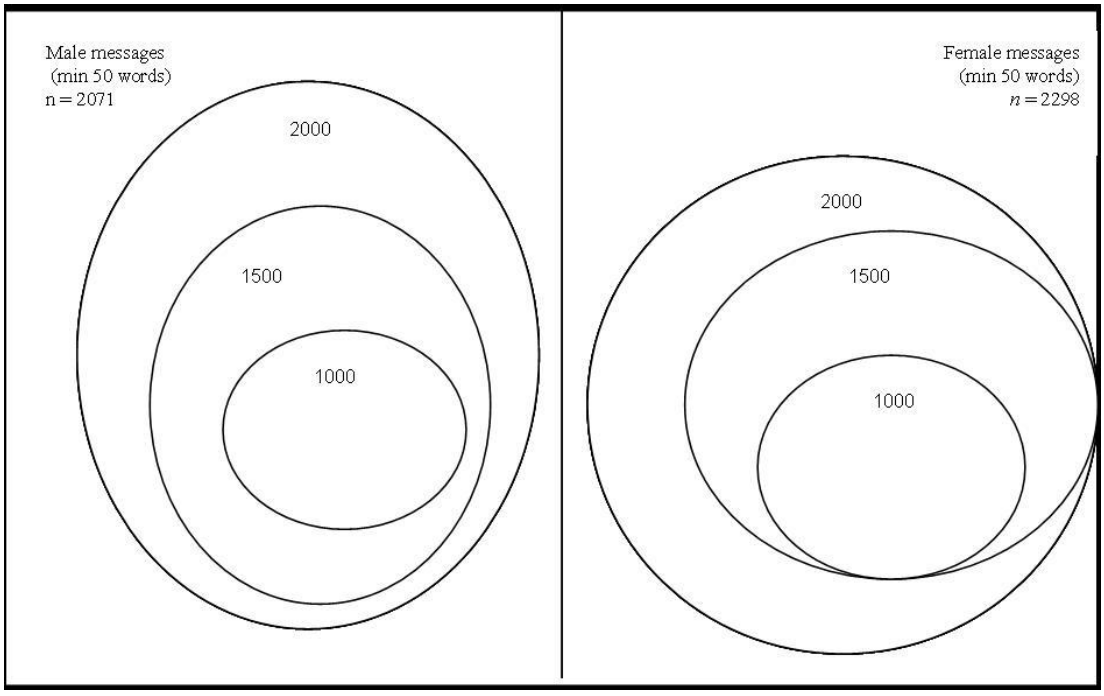


Figure 3-11: Production of Successively Smaller Cohorts by Sub-sampling

As can be seen, there are fewer e-mail messages in each cohort as the minimum word count increases.

To ensure that the e-mail messages used in the experiment were not biased by any particular author, the cohorts were randomly sampled from the cohort of all messages which had the specified minimum number of words. Each successively smaller cohort in these experiments was subsampled from the larger parent sample. The male and female data sets were produced with an equal number of e-mail messages, down to a minimum of 50 messages per data set. Figure 3-11 displays the sampling process diagrammatically.

3.5.3.3 Effect of Feature Sets on Gender Classification

It would be interesting to determine the subset of the full set of features that produce the best discrimination between the gender cohorts. This would require an exhaustive feature set sub-selection experiment. The approach was rejected because of time constraints. Instead, as a broad-brush approach to feature set sub-selection, experiments were performed where each current feature set was removed, one at a time, from the full set of features to determine the effect (positive or negative) on classification results.

3.5.4 Cohort Testing - Experience with the English Language

A similar set of experiments as outlined in Section 3.5.4 for determining gender was run to determine if language background could be discriminated. It was necessary to add more authors to the *cleaned inbox* data set to produce a sufficient quantity of ESL authored e-mail messages. This produced the *language* data set, details of which are included in Section 3.6.

The experiments undertaken were based on determining the number of data points and minimum number of words required to develop a characterisation model of language background.

3.6 Data Sources

The data sets used for the research reported in this thesis are outlined below.

Data Set Name: *book*

A small selection of books were obtained from Project Gutenberg (n.d.). The books and authors are shown in Table 3.8. The text from the books was randomly sampled

from each source document in such a way that chunks of non-overlapping text were prepared which had a constant number of words. The number of chunks of text and the number of words per chunk were variables used in the sampling of text from the source documents.

Author	Title	Year	Word Count
Jane Austen	Pride and Prejudice	1813	123,249
Jane Austen	Sense and Sensibility	1811	120,743
Sir Arthur Conan Doyle	The Memoirs of Sherlock Holmes	1894	88,882
Joseph Conrad	Nostromo	1917	170,726
Charles Dickens	Great Expectations	1861	188,844
Robert Louis Stevenson	Treasure Island	1883	70,161

Table 3.8: Details of the Books Used in the *book* Data Set

Data Set Name: *thesis*

Three PhD theses were obtained for use as a different source of plain text documents. All tables, figures, source code sections and any other parts of the documents that were not plain text were removed from the documents before they were used. As for the *book* data set, the text from the theses was randomly sampled from each source document. The total number of words in each thesis is shown in Table 3.9.

Authorship Class	Word Count
A	29,391
B	24,880
C	33,984

Table 3.9: Details of the PhD Theses Used in the *thesis* Data Set

Data Set Name: *inbox*

This data set consists of the author's e-mail messages collected over a period of 15 months. The messages are multi-topic with no minimum word count restrictions. There are over 10,800 messages from over 800 authors in this data set. The authors come from a range of backgrounds: academics, students, general staff and personal acquaintances. To qualify for this data set, each author had to have 10 or more messages to contribute. Many messages were unsuitable for use in tests as they contained non-original text that was not marked as requoted in the message. These e-mail messages were removed as discussed in Section 3.5.4 to produce the separate *cleaned inbox* data set (see below).

Data Set Name: *email4*

The *email4* data set was a subset of the *inbox* data set. This data set was created from four authors who had a significant number of e-mail messages in the *inbox* data set. Details of the number of e-mail messages in each authorship class are shown in Table 3.10.

Authorship Class	Number of E-mail Messages
A	36
B	65
C	86
D	66

Table 3.10: Details of the *email4* Data Set

Discussion Topic	Authorship Class			Topic Total
	A	B	C	
movies	15	21	23	59
food	12	21	25	48
travel	3	20	15	38
Author Total	30	62	63	155

Table 3.11: Distribution of E-mail Messages for Each Author and Discussion Topic

Data Set Name: *discussion*

It was necessary to obtain data for experiments relating to the effect of topic on authorship attribution, which contained e-mail or newsgroup messages from a small number of authors who discussed a series of disparate topics. A possible source of such messages is the Usenet newsgroups. However, it was not possible to find a set of authors who frequently contribute to the same set of newsgroups. It became necessary to solicit e-mail from a group of authors on three separate topics to each of which the authors felt they could contribute messages.

The e-mail was produced in three topics of discussion by three authors. The topics were *food*, *movies* and *travel*. E-mails had a minimum word count of approximately 100 words. This data set contained 155 e-mail messages. Details of the number of messages contributed by each author in each topic are shown in Table 3.11. The assumption was that each topic would contain a common vocabulary, e.g. restaurant, cook, recipe etc. for the *food* topic.

Data set name: *cleaned inbox*

This data set was created specifically for authorship characterisation from the *inbox* data set that was cleaned by removing any unsuitable e-mail messages as discussed in

Minimum Number of Words	Male Cohort	Female Cohort
0	3479	4514
50	2071	2298
100	1257	1072
150	842	585
200	564	384

Table 3.12: Number of E-mail Messages in each Gender Cohort with the Specified Minimum Number of Words

Section 3.5.4. Each message in this data set contained original text from the author of the message. This data set contains 8820 messages from 342 authors.

Data Set Name: *gender*

This data set consists of e-mail messages selected from the *cleaned inbox* data set. To qualify for this data set, the gender of the message had to be determined from public domain or personal knowledge as discussed in Section 3.5.4. Some details of the number of e-mail messages containing a minimum number of words are shown in Table 3.12. It should be noted, and is further discussed in Section 5.4.1 that gender data present problems with the message size.

Data Set Name: *language*

This data set was created from the *cleaned inbox* data set which had extra ESL authored e-mail messages added to it. The language background of each author in the data set was determined as discussed in Section 3.5.4. This data set contains over 9700 messages from 827 authors. There is no minimum limit on the number of e-mail messages an author has to have written to be included in this data set. Some details of the number of e-mail messages containing a certain minimum number of words

Minimum Number of Words	ENL Cohort	ESL Cohort
0	7158	1743
50	3926	706
100	2128	357
150	1311	231
200	878	161

Table 3.13: Number of E-mail Messages in each Language Cohort with the Specified Minimum Number of Words

are shown in Table 3.13. It should be noted as for the *gender* data set that minimum message size presents some problems.

3.7 Chapter Summary

This chapter detailed how the analysis of e-mail message authorship would be carried out. It described the necessity for specific sequences of experiments establishing baseline parameter values, and also described the preparation of suitable data sources.

The experimental plan for testing aspects of anonymous e-mail authorship began with baseline testing with data that was not e-mail. Experiments designed to test the effect of the features discussed in Section 3.2 on classification performance were examined. The effect of the number of words per message and the number of messages on classification performance was also included in the baseline experimental plan.

The results of these experiments and a discussion of their impact are described in Chapter 4.

Once the baseline parameters were established, e-mail message data was tested and the question of whether or not the structural features available in e-mail messages could assist in classification of authorship was addressed. It was also planned to test the effect of topic on classification using solicited e-mail message data.

The determination of e-mail author gender and an author's language background may be useful in a forensic investigation for characterising an author's sociolinguistic background. Section 3.5 discussed how author characterisation would be undertaken. Various factors such as the number of e-mail messages in a cohort and the minimum number of words a message requires to show cohort discrimination would be investigated to determine the effectiveness of author characterisation.

The results of the experiments undertaken using e-mail messages as the data source and a discussion of these results form the basis of Chapter 5.

Chapter 4

Calibration of Experimental Parameters and Baseline Experiments

Chapter 3 discussed the plan and scope of the experimental process that was used to determine the feasibility or otherwise of analysing authorship features and identifying authorship of e-mail messages. The list of the features to be evaluated was outlined and justification was given as to why the Support Vector Machine (SVM) algorithm would be used for this work.

The framework for machine learning experimental conduct was discussed in Section 3.1. It was decided that the 'one against all' approach for multi-class classification experiments would be used, and that the error rate, precision, recall and F_1 values would be calculated for each test to determine the classification performance for a given set of inputs. It was further decided that the weighted macro-averaged error rate and weighted macro-averaged F_1 values would be used to compare tests where a parameter of the test is being studied. Due to the small number of data points in the corpus, it was decided in Section 3.1.4 to use stratified 10-fold cross validation to extend the evaluation of the data set for the experiments.

The previous chapter also discussed how it would be useful to have a tool that could be used for characterising an author's sociolinguistic profile or, assigning various

cohort attributes such as gender, language background, age and education level to an author. It was decided due to the difficulty in obtaining data that cohort profiling would be examined on the gender and language background cohorts only at this stage. If the testing showed some discrimination between these sociolinguistic classes it would be worth attempting analysis of further cohort profiles in the future.

This chapter details the experiments that were undertaken to develop a systematic basis for classifying the authorship of e-mail messages for forensic purposes and reports their results. This was done by first conducting a series of experiments designed to reveal baseline values for successful SVM authorship attribution of plain text chunks (not e-mail), thereby setting the constraints on feature sets, text size and number of messages. These baseline experiments set the framework for the core of this project, the task of identifying useful features contained in e-mail text. Results of these core experiments are discussed in Chapter 5.

A list of the experiments reported in this chapter is shown in Table 4.1.

4.1 Baseline Experiments

A random collection of e-mail messages would not be ideal for evaluating various basic authorship attribution parameters, as the messages would not have a constant length. The usual length of e-mail messages is less than the minimum requirement of 1000 words suggested in stylometry literature (see Section 2.1). It was decided, therefore, to conduct initial authorship attribution experiments on data that was similar to previous stylistics studies viz. free text or novels. By using data that is not e-mail, the effect of variable text size between data points can be eliminated.

The aim of these experiments was to determine whether or not the stylistic features identified in the literature review and outlined in Section 3.2 were suitable for

Experiment Number	Reported in Section	Experimental Details
B_1	4.3.1	Effect of different feature sets, 1000 words per text chunk, *books* data set
B_2	4.3.1	Effect of different feature sets, 1000 vs 100 words per text chunk, *books* data set
B_3	4.3.2	Effect of different feature sets, 1000 words per text chunk, *thesis* data set
B_4	4.3.3	Effect of collocations, 200 words per text chunk, *thesis* data set
B_5	4.4.1	Effect of number of words per text chunk, *thesis* data set
B_6	4.4.2	Effect of number of data points per class, 200 and 500 words per text chunk, *thesis* data set
B_7	4.5.1	Effect of SVM kernel function, 200 words per text chunk, *thesis* data set
B_8	4.5.1	Effect of degree on polynomial kernel function, 200 words per text chunk, *thesis* data set
B_9	4.5.1	Effect of gamma value on radial basis function kernel, 200 words per text chunk, *thesis* data set
B_{10}	4.5.2	Effect of SVM cost parameter, 200 words per text chunk, *thesis* data set

Table 4.1: List of Baseline Experiments

authorship discrimination on plain text documents using an SVM. The *book* data set and the *thesis* data set described in Section 3.6 were used in these tests to identify, independently of the e-mail context:

- the most effective features from those identified for authorship classification
- the minimum size of text that results in good classification
- the minimum number of data points per authorship class required for good classification

As foreshadowed in 2.5, it is the aim of this research to achieve correct classification of e-mail data at a level of approximately 85% using the weighted macro-averaged F_1 measure.

The tests reported below were carried out using the 'one against all' method for multi-class classification and stratified 10-fold cross validation, except where noted. All reported results are the average of ten repeated tests. The weighted macro-averaged error rate ($\overline{E}^{(M)}$) and the weighted macro-averaged F_1 value ($\overline{F}_1^{(M)}$) for each test are reported in each case. For the tests reported in Sections 4.3 and 4.4 the default SVM^{light} parameters for polynomial kernel function with degree two were used in the learning phase. The "LOQO" optimiser was used for maximising the margin.

4.2 Tuning SVM Performance Parameters

Some initial experiments were undertaken to determine how SVM^{light} behaved and to determine the effect of some of the arguments to the program.

4.2.1 Scaling

As discussed in Section 3.1.1 the input to train an SVM is a set of training vectors consisting of a feature vector $\vec{x} \in \Re^N$ and a label $y \in \{-1, +1\}$. An important finding from these initial experiments was that all feature vectors must be scaled across each feature x_i in $\vec{x}$. It is possible to have values for different features in the same vector which differ by 3 to 6 orders of magnitude. If these features are left unscaled, the kernel function used in the SVM has difficulty in converging on a separating hyperplane. Training without scaled data is also costly on a time basis and often results in a large number of training errors.

The data is scaled between a lower bound lb and an upper bound ub. The lower bound is normally set at -1.0 or 0.0 and the upper bound at 1.0. Features with a value of 0.0 are ignored by the SVM^{light} classifier, so the lower bound for all experiments in this research was set at 0.0 to reduce the number of active features per input vector.

Scaling is performed by calculating a scale factor s_i and a threshold t_i for each feature in the current data set where:

$$s_i = \frac{ub - lb}{x_{i,max} - x_{i,min}}$$

and

$$t_i = \frac{lb}{s_i} - x_{i,min}$$

The scaled feature values $x_{i,scaled}$ are calculated from s_i and t_i using the following formula:

$$x_{i,scaled} = (x_i + t_i) \times s_i$$

When a model is learned for a particular classifier, any data that is to be classified with that model must be scaled in the same manner using the same scale factor and threshold for each feature.

4.2.2 Kernel Functions

The kernel function selected for the SVM can have a considerable impact on the learning phase of the classification of the data. SVM^{light} provides four standard kernel functions: a linear function, a polynomial function in which the order of the polynomial

can be varied, a radial basis function and a sigmoid $\tanh$ function. Initial investigations into the effect of the default kernel functions on classification performance showed that the second order polynomial kernel function produced the best results. A more thorough investigation of the effect of the kernel function is reported in Section 4.5.

4.3 Feature Selection

A range of experiments was conducted to determine the best set of features to use in initial trials with e-mail data. As Section 3.2 explained, features with a similar level of granularity were grouped together. Each of the feature sets was tested individually to see its relative effectiveness with the same data set. Tests were conducted on the *book* data set and on the *thesis* data set, details of which were outlined in Section 3.6. The results of these tests are discussed separately below.

4.3.1 Experiments with the *book* Data Set

This data set was initially tested using stratified 10-fold cross validation on one book each from five authors to see the effect of various feature sets. The books were sampled to create 50 chunks of text containing 1000 words each.

The books were tested using the character-based features (C), word-based features (W), word length frequency distribution (L), function word (F) and 2-gram feature sets and some combinations of these. The results are shown in Table 4.2. No training errors were encountered in these tests and the proportion of support vectors was approximately 30 to 40%. The results show that the 2-gram feature set is the best individual feature set of those used and that the best results are obtained when all feature sets are added together.

Feature Set	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ (%)
Character Based (C)	7.7	79.6
Word Based (W)	6.2	82.9
Function Word (F)	2.5	93.0
2-grams	1.2	97.1
C+W	3.1	92.2
C+F	1.6	95.9
W+F	1.6	95.4
C+W+F	1.2	96.9
2-grams+C+W+F	0.7	98.3

Data Set: *books*
Feature Set: Various

Table 4.2: Test Results for Various Feature Sets on 1000 Word Text Chunks

To determine how well the learned classifiers could generalise about new test data, all 50 data points for each book were used to create one classifier for each author and the sixth book in the data set, which had the same author as one of the five, was used as a test set. The results of testing for various feature sets are shown in Table 4.3.

It can be seen from these results that Austen's second book is correctly not predicted as any of the other four authors but there are some errors made when Austen's second book is classified using Austen's first book. The error rate is greatest when 2-grams alone were used as the feature set for classification.

This experiment was repeated with 50 chunks of text, each containing 100 words, to see the effect of chunk size on attribution. The results for this test are also shown in Table 4.3. There is a marked decrease in classification performance when the number of words is reduced from 1000 to 100. The results for the 2-gram feature set are much worse when fewer words are used per data point. This is possibly due to the frequencies

of the 2-grams being much more variable over 100 words than over 1000 words. E-mail messages rarely have an average length of 1000 words and this feature set may not be useful in the classification of authorship of e-mail messages.

Chunk Size	Feature Set	Error Rate (%)				
		Dickens	Conan Doyle	Conrad	Austen	Stevenson
1000 Words	2-grams	0.0	0.0	0.0	14.0	0.0
	C+F+W+L	0.0	0.0	0.0	4.0	0.0
	2-grams+C+F+W+L	0.0	0.0	0.0	2.0	0.0
100 Words	2-grams	0.0	0.0	0.0	56.0	0.0
	C+F+W+L	8.0	0.0	6.0	36.0	4.0
	2-grams+C+F+W+L	6.0	0.0	0.0	40.0	0.0

Data Set: *books*
Feature Set: Various

Table 4.3: Error Rates for a Second Book by Austen Tested Against Classifiers Learnt from Five Other Books

One reason for the higher error rates in the classification of Austen versus Austen, especially with 100 word chunks, may be that the two Austen novels have a different group of characters taking part in the story. The frequencies of the 2-grams in the character names will be inflated and there will be a different set of character name 2-grams in each book. This might indicate that 2-grams could be useful features for detecting subject or content but may not be useful for classifying authorship with small chunks of data.

4.3.2 Experiments with the *thesis* Data Set

The text data from the *thesis* data set was split into 1000-word chunks, as for the *book* data set experiments reported above. The number of chunks was not kept constant

between authorship classes. The feature sets were tested individually and in various combinations in a similar fashion to the experiments performed on the *book* data set. The results are reported in Table 4.4.

Feature Set	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ **(%)**
Character Based (C)	8.5	86.7
Word Based (W)	6.9	89.4
Function Words (F)	2.7	95.6
Word Length Frequency Distribution (L)	12.7	81.4
2-grams	0.8	98.8
C+W	3.5	93.9
C+F	0.8	98.8
C+L	5.2	92.1
F+W	0.8	98.8
F+L	0.3	99.6
L+W	4.1	93.8
C+W+F	1.2	98.0
C+W+L	4.1	93.8
C+F+L	0.3	99.6
F+L+W	0.3	99.6
C+F+L+W	0.3	99.6

Data Set: *thesis*
Feature Set: Various

Table 4.4: The Effect of Feature Sets on Authorship Classification

Of the individual feature sets, 2-grams gave the best results with a weighted macro-averaged error rate ($\overline{E}^{(M)}$) of 0.8% and weighted macro-averaged F_1 value ($\overline{F}_1^{(M)}$) of 98.8%. The next best set was function words with $\overline{E}^{(M)}$of 2.7% and $\overline{F}_1^{(M)}$ of 95.6%.

As experienced with the *book* data set, when feature sets were added together the results improved, giving an indication that the more features used, the easier it is to discriminate between authorship classes. In many machine learning classifiers, this

would lead to overfitting of the model to the training data, but as Section 2.4 discussed, this is not the case with Support Vector Machines.

4.3.3 Collocations as Features

The literature indicated that some success in authorship attribution had been achieved when word collocations were used as the features for authorship identification (Smith, 1983). A set of collocations based on function word pairs was compiled and used for a classification experiment. The collocations used in this experiment are listed in Appendix A.

Experiments were undertaken using the *thesis* data set using 200 word chunks of text. The addition of the collocation features to the combination of character based features, word based features, word length frequency distribution and function word feature sets caused a reduction in the classification efficiency for this data. This was most probably due to the low number of words in the text chunks, introducing noisy feature values which lead to difficulties in determining the hyperplane decision surface during the training phase.

Further experimentation with collocations was suspended because of the unpromising results.

4.3.4 Successful Feature Sets

The previous sections have indicated that, in general, using SVM and the stylistic features compiled from the literature has been successful for the attribution of authorship of plain text. When character based features, word based features, word length frequency distribution and function word feature sets were combined, the classification efficiency of the attributions was better than when individual feature sets were used.

The use of 2-grams resulted in effective classification using the *book* data set, but the massive error rate increase as chunk size declined from 1000 words to 100 (Table 4.3) leads to the suspicion that these features might be affected by the topic of the text being analysed. It is also possible that for chunks of text of approximately 100 words, i.e. similar in size to an e-mail message, the 2-gram frequencies may not be consistent enough to be useful features.

The use of collocations as a feature set also resulted in poorer classification when they were added to the stylistic feature sets found successful previously. Collocations as features were not pursued any further.

Of the successful individual feature sets, the function word feature set was the most successful of those found to be unaffected by text content. In all of these experiments, no training errors were recorded, indicating that the features being used were successfully separating the authorship classes. Since function words had been found to be important markers for authorship attribution by Mosteller and Wallace (1964) in their seminal study of the Federalist Papers, this confirms the SVM as a suitable classification method for this problem domain and function word based features as a reliable style marker. These features are independent of the content or topic in a piece of text.

4.4 Calibrating the Experimental Parameters

4.4.1 The Effect of the Number of Words per Text Chunk on Classification

As Section 2.1 showed, the literature on stylistics and authorship attribution suggests that the minimum text size for authorship attribution is 1000 words. E-mail messages typically do not contain so many words. It is necessary then, to try to determine the

minimum number of words that will result in reliable attribution of authorship for e-mail. Using the best sets of features found in Section 4.3, experiments were conducted on the *thesis* data set to determine the effect of the number of words in each text chunk on classification performance. Chunk sizes of 1000, 500, 200 and 100 words were used in these experiments and a number of different feature sets were tested.

Using thesis data chunks Table 4.5 shows the effect of the number of words per chunk on the weighted macro-averaged error rate and F_1 results for the various feature sets listed in Table 4.4. Figure 4-1 shows these results graphically. Only the combinations involving function words were tested as the function words were found to be an important set of features in Section 4.3.

As more features were used, the error rates decreased and the F_1 values increased. This supports the evidence already gathered, that the more features that were used, the better the discrimination between classes became.

Weighted macro-averaged error rates of less than 10% and F_1 values of greater than 90% were achieved with the feature sets used on 200 word segments of data. This was an encouraging result, as many e-mails contain at least as many words as that. As the number of words per text chunk decreased, training errors were still not evident, indicating that the authorship problem is a separable one when these features are used.

It is interesting to note in Table 4.5 that there was little variation in the results of classification with chunk size when 2-grams were used as the feature set. In the *books* experiment for identifying a previously unseen Austen text (Table 4.3) it was clear that 2-grams failed badly as a discriminator. However, within topic it appears very successful. This supports the idea that 2-grams may be good markers for indicating the subject matter of a document. The three thesis documents in the *thesis* data set are about different topics, although all three address information technology topics. Due

Number of Words	Feature Set							
	Character Based (C)		Word Based (W)		Function Words (F)		Word Length Frequency (L)	
	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)
100	14.8	75.6	19.4	69.9	15.8	75.0	-	-
200	12.6	80.2	14.0	78.8	10.2	84.0	22.3	67.4
500	12.2	81.9	11.1	83.5	4.6	93.1	15.8	76.6
1000	8.5	86.7	6.9	89.4	2.7	95.6	12.7	81.4

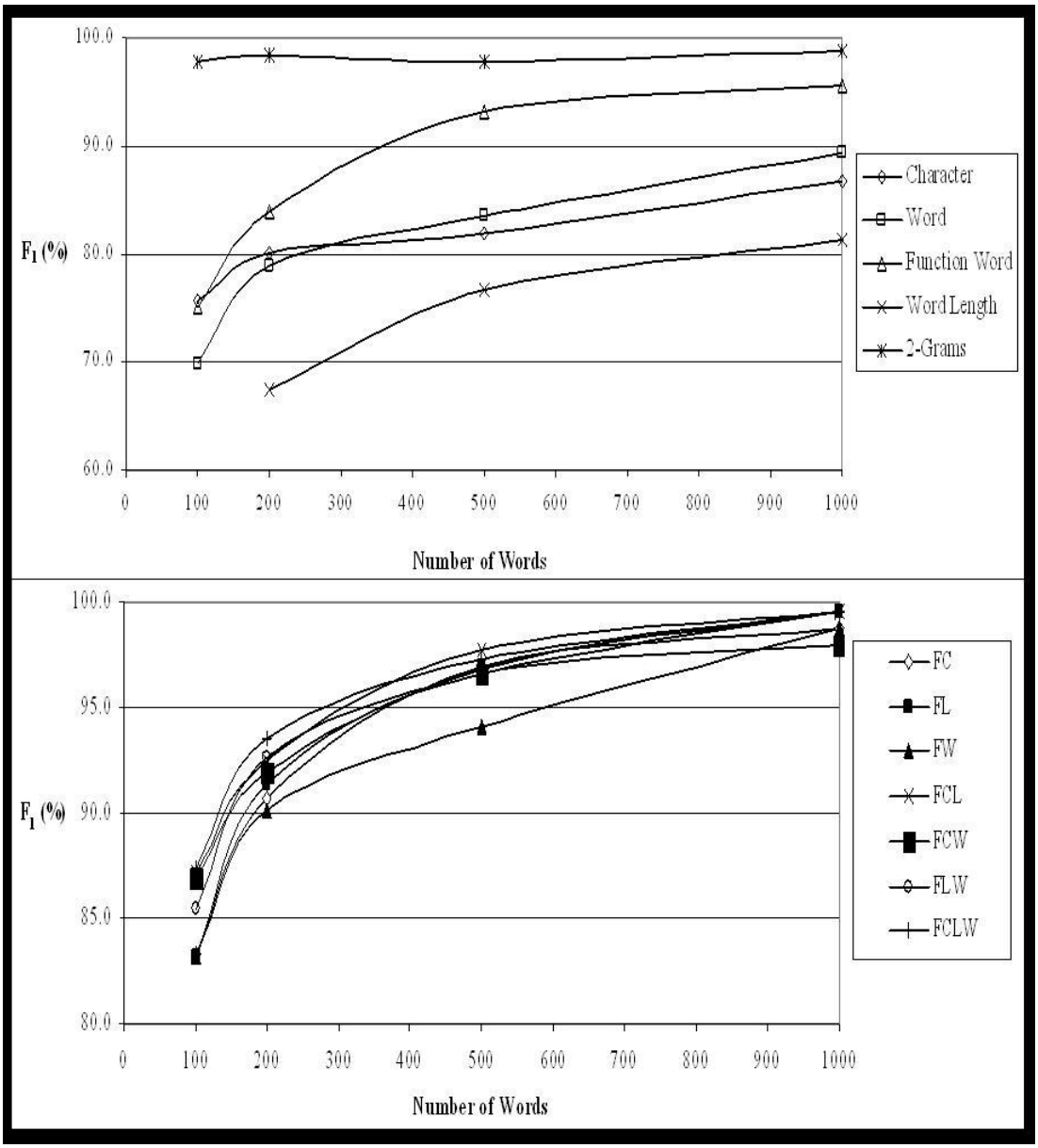
Number of Words	Feature Set							
	2-grams		C+F		F+L		F+W	
	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)
100	1.5	97.7	10.6	83.3	10.9	83.2	10.9	83.2
200	1.1	98.3	5.9	90.7	5.6	91.4	6.5	90.1
500	1.5	97.7	2.0	96.9	2.2	96.8	3.9	94.1
1000	0.8	98.8	0.8	98.8	0.3	99.6	0.8	98.8

Number of Words	Feature Set							
	C+F+L		C+F+W		F+L+W		C+F+L+W	
	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)
100	8.3	87.2	8.5	86.9	9.4	85.5	8.1	87.4
200	4.9	92.5	5.2	91.9	4.9	92.6	4.3	93.5
500	1.6	97.7	2.2	96.6	2.4	96.6	1.8	97.3
1000	0.3	99.6	1.2	98.0	0.3	99.6	0.3	99.6

Data Set: *thesis*
Feature Set: Various

F = Function Words C = Character based W = Word based
L = Word Length Frequency Distribution

Table 4.5: Effect of Chunk Size for Different Feature Sets



Data Set: *thesis*
Feature Set: Various

Figure 4-1: Effect of Chunk Size for Different Feature Sets

to the nature of the results obtained with 2-grams, and the hypothesis that 2-grams are good markers for content but not helpful otherwise, the 2-gram feature set was removed from all further experimentation.

4.4.2 The Effect of the Number of Data Points per Authorship Class on Classification

It may be difficult in a forensic investigation to obtain many, say 100 or more, e-mail messages for building an authorship model. Any successful technique must be reliable with as few e-mail messages as possible. Each e-mail message should also ideally be treated as a separate data point for analysis if possible.

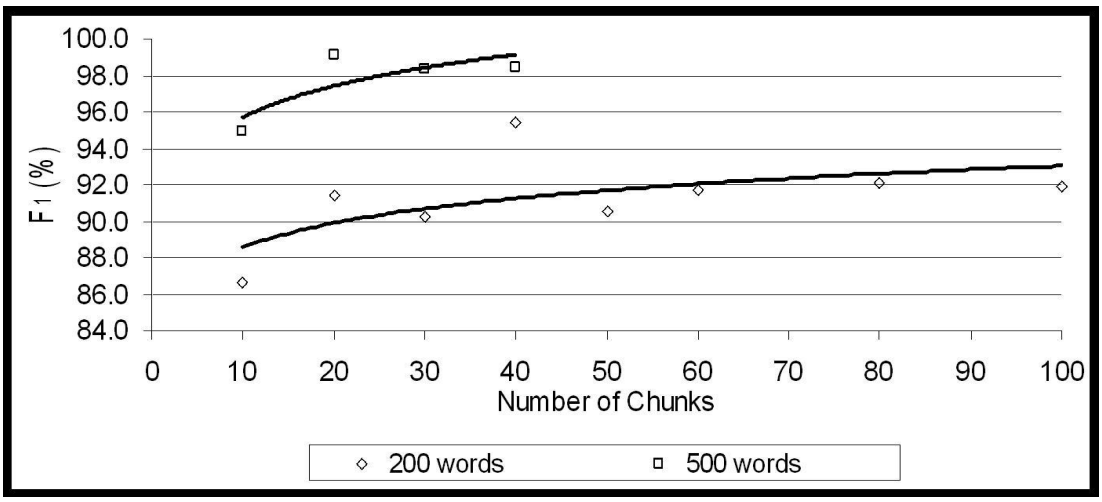
In these experiments, the thesis documents were sampled by splitting them into an equal number of sections or chunks with a constant number of words. Tests were performed with 200 and 500 word chunks. The features used in these tests were a combination of character-based, word-based, function words and the word length frequency distribution feature sets. The results are shown in Table 4.6 and in Figure 4-2.

The results of the experiments show that there is a leveling-off effect after the number of document chunks reaches twenty. This is once again an encouraging result for the analysis of e-mail authorship in a forensic context as it shows that 20 data points may be sufficient for effective classification.

Number of Data Points	200 Word Chunk Size		500 Word Chunk Size	
	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)	$\overline{E}^{(M)}$ (%)	$\overline{F_1}^{(M)}$ (%)
10	8.9	86.6	3.3	94.9
20	5.6	91.4	0.6	99.1
30	6.3	90.2	1.1	98.3
40	3.1	95.4	1.1	98.4
50	6.2	90.5	-	-
60	5.1	91.7	-	-
80	5.1	92.1	-	-
100	5.3	91.9	-	-

Data Set: *thesis*
Feature Set: C+F+L+W

Table 4.6: Effect of Number of Data Points



Data Set: *thesis*
Feature Set: C+F+L+W

Figure 4-2: Effect of Number of Data Points

4.5 SVM^{light} Optimisation

While the stylistic and structural features and other parameters such as number of words per e-mail message and number of e-mail messages define the problem domain, it may be possible to make further improvements in performance if the classification tool is optimised. The SVM^{light} classifier has a number of parameters that can be tuned. The kernel function can be altered, and each kernel function in turn has a number of parameters that may affect the performance of the classifier.

4.5.1 Kernel Function

The SVM^{light} implementation has four standard kernel functions - a linear function, a polynomial function, a radial basis function and a sigmoid $\tanh$ function. Each of these kernel functions was tested with its default parameters to determine the effect on classification efficiency. Table 4.7 shows the results from this experiment where the *thesis* data set was used. The polynomial kernel function with default parameters was found to be the best performed, although the linear kernel function performed nearly as well.

Kernel Function	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathbf{M})}$ **(%)**
Linear	4.4	93.2
Polynomial	3.7	94.4
Radial basis function	18.4	59.2
Sigmoid $\tanh$	33.1	0.5

Data Set: *thesis*
Feature Set: C+F+L+W

Table 4.7: Effect of Kernel Function with Default Parameters

For the polynomial kernel function, the degree of the polynomial and the values for the scalar and the constant can be altered.

A set of tests was run using plain text data where the degree of the polynomial was altered. The results for the *thesis* data set are reported in Table 4.8. The results of a Student's t test, $(n = 10,\ \alpha = 0.05)$ showed that there was no significant statistical difference between 3^{rd}, 4^{th}, 5^{th}, 6^{th} and 7^{th} order polynomial kernel functions. Higher order polynomials had a worse performance and training errors were recorded for these classifiers. The other parameters for the polynomial kernel, the scalars and constants, had little effect on classification performance when tested.

Degree of Polynomial	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ **(%)**
1	4.4	93.2
2	4.0	93.2
3	3.7	94.4
4	3.6	94.5
5	3.3	95.0
6	3.2	95.1
7	3.5	94.7
8	23.2	69.5
9	46.0	22.0
10	46.0	22.2

Data Set: *thesis*
Feature Set: C+F+L+W

Table 4.8: Effect of Degree of Polynomial Kernel Function for the *thesis* Data Set

The radial basis function (RBF) kernel has a default gamma value of 1.0. When this value is reduced by a factor of five, the optimal classification efficiency was found. Results of experiments with the gamma value are shown in Table 4.9. There was,

however, no significant improvement over the 3^{rd} order polynomial kernel function. The RBF kernel was pursued no further in this work.

The tests that were undertaken indicate that the 3^{rd} order polynomial with default scalar and constant parameters should be suitable for authorship classification of text data. While this result cannot be applied universally to all analyses of text authorship, it shows that there was little to be gained in using a more complex kernel function for plain text documents containing 200 words. All further experiments with plain text documents and the initial experiments with e-mail were conducted using the default 3^{rd} order polynomial kernel function. The effect of the degree of the polynomial kernel function on e-mail data was also investigated and is reported in Section 5.2.3.

Gamma Value	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ **(%)**
0.01	4.1	93.7
0.02	4.0	93.9
0.05	3.4	94.8
0.1	3.5	94.7
0.2	3.2	95.1
0.5	5.1	91.6
1.0	18.4	59.2
2.0	33.0	0.4

Data Set: *thesis*
Feature Set: C+F+L+W

Table 4.9: Effect of Gamma on Radial Basis Kernel Function for *thesis* Data

4.5.2 Effect of the Cost Parameter on Classification

Many machine learning classifiers provide a cost parameter, which allows the user to set a threshold up to which the cost of training errors in the learning phase are

acceptable. If the cost parameter is set to a high value, the classifier will attempt to ensure that as few training errors as possible are made. This is done at the expense of training time. Obviously, if training errors are made, it is likely that a higher number of errors in classification will be made when the classifier model is used to test data.

Experiments with the cost parameter, C, were performed with the *thesis* data set using 1000 word chunks of data. The results are shown in Table 4.10. These results showed that a very low value for C was required to have any negative effect on classifier performance. No training errors were made by the SVM learner in any of the experiments until C was reduced to 0.00005. A value for C of 1 was used in all subsequent experiments.

Cost	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ **(%)**
0.00001	33.1	0.0
0.00002	9.6	74.2
0.00005	1.5	97.0
0.0001	0.4	99.3
0.001	0.4	99.3
0.1	0.4	99.3
1	0.4	99.3
10	0.4	99.3
100	0.4	99.3

Data Set: *thesis*
Feature Set: C+F+L+W

Table 4.10: Effect of C Parameter in SVM^{light} on Classification Performance

4.6 Chapter Summary

This chapter has reported the results of experiments undertaken to determine parameters for successful authorship attribution on plain text with a view to optimising SVM performance when e-mail text replaces chunks of plain text. The data used for these experiments was chunks of text from some books and PhD theses. The approach taken was to start with large chunks of text to identify discriminatory features, and to use an SVM for classification. When the most successful set of features was identified, the number of words per text chunk or data point and the number of data points required to attain reliable classification were investigated.

The following findings summarise the results of this chapter:

- Authorship attribution of text documents can be successfully performed using a Support Vector Machine.
- The baseline tests have shown that function words were consistently the best individual feature set independent of topic. The results also showed that better results were obtained by adding extra feature sets to the function word feature set. The 2-gram feature set could identify a previously learned author accurately but could not identify unseen text by the same author with any accuracy. It is hypothesised that this set is biased toward discrimination of content.
- The baseline tests have also shown that it may be possible to attribute text authorship using the chosen feature sets on chunks containing as few as 100 to 200 words, with possibly only 20 data points per authorship class. This is encouraging for tests that will be based on e-mail, as many e-mail messages contain at least 200 words.

- The SVM^{light} tool works quite effectively when the default parameters are used, though slight gains in classification performance can be made if the parameters of the classifier are optimised for the problem domain being used. Experimentation with these parameters found that the 3^{rd} order polynomial kernel function with default scalar and constant values was suitable for the current problem domain.

- Scaling across each feature between 0.0 and 1.0 improves performance. Test data must be scaled with the same threshold and scale factors as those calculated from the training data for each feature.

These results confirmed that the approach used to date could be used as the basis for further research with e-mail data. Chapter 5 discusses the results of the experiments conducted on e-mail message data. Those experiments utilised the findings of the baseline experiments reported in this chapter for setting various parameters.

Chapter 5

Authorship Attribution and Profiling of E-mail Messages

Chapter 4 discussed results for calibration of the parameters for various baseline experiments, where these were conducted on data sources that were not e-mail. This approach was taken so that the chunk size of the text being analysed could be kept constant while other experimental parameters were established.

The selection of the best sets of stylistic features was also undertaken. It was found that of the stylistic feature sets identified and studied, the function word set was the individual feature set that provided the most power for discrimination of authorship. Importantly, though, as suggested by Rudman (1998), when more features were used in combination with one another, the discriminatory potential was increased.

The baseline experiments conducted and discussed in Chapter 4 provided a necessary framework for the conduct of subsequent e-mail authorship attribution experiments. This chapter reports results for the experiments that were undertaken to develop a systematic means for classifying the authorship of e-mail messages for forensic purposes.

Section 5.1 discusses the initial experiments undertaken on e-mail data. E-mail data was used for the experiments discussed in this chapter, and because of this it was

possible to test the impact of e-mail specific features for the first time. Section 5.2 outlines how improvements in the results were obtained. The effect of the topic of discussion in the e-mail messages is established in Section 5.3. In Section 5.4, experimental results for profiling sociolinguistic cohorts are shown. As stated in Chapter 2, it was the aim of this research to achieve correct classification of e-mail data at a level of approximately 85% using the weighted macro-averaged F_1 measure. The results reported in this chapter indicate that this goal was achieved after the addition of the structural features of e-mail messages.

A list of the experiments reported in this chapter is shown in Table 5.1.

5.1 Experiments with E-mail Messages

5.1.1 E-mail Specific Features

Using the knowledge gained from baseline experiments reported in Chapter 4, attention was turned to the classification of e-mail message authorship.

Experiments reported in Section 4.4 established that the combined character based, word based, word length frequency distribution and function word feature sets were found to be the best combination of features to use. Initial experiments in this e-mail phase used this best combination of stylistic text feature sets, the e-mail structural features (E), and the HTML tag feature set defined in Section 3.4.1 (H). In these tests, the e-mail specific feature set and the HTML tag feature set were sequentially added to the stylistic feature sets. The results for these experiments are shown in Table 5.2.

When e-mail data was analysed using the stylistic feature sets only, the $\overline{F}_1^{(M)}$ result was 64.9%. This is lower than the results achieved in the baseline experiments on text chunks containing 200 words. This is likely to be due to the e-mail messages having

Experiment Number	Reported in Section	Experimental Details
E_1	5.1.1	Effect of addition of e-mail specific feature sets, *email4* data set
E_2	5.1.2	Effect of chunking, '200 word e-mail chunks' vs 'individual data points', *email4* data set
E_3	5.2.1	Effect of extra function word features, 100 - 1000 word chunks of text, *thesis* data set
E_4	5.2.1	Effect of extra function word features, *email4* data set
E_5	5.2.2	Effect of part of speech of function words, *thesis* data set
E_6	5.2.2	Effect of part of speech of function words, *email4* data set
E_7	5.2.3	Effect of degree of polynomial kernel function, *email4* data set
E_8	5.3	Effect of topic, *discussion* data set
E_9	5.3	Effect of topic on generalisation capability, classification of *food* and *travel* topic sets using *movies* topic models
E_{10}	5.4.1	Effect of number of data points and minimum number of words on gender cohort classification, *gender* data set
E_{11}	5.4.1	Effect of feature sets on gender cohort classification, *gender* data set
E_{12}	5.4.2	Effect of number of data points and minimum number of words on language background cohort classification, *language* data set

Table 5.1: List of Experiments Conducted Using E-mail Message Data

variable length, leading to more variability in the feature variables from e-mail to e-mail. It was not possible to build a data set containing sufficient e-mail messages from a group of authors where the text length was held constant at $200 \pm 10\%$ words, as there were not enough messages that matched the criteria in the *inbox* data set. Building such

a data set and repeating the baseline work where the number of words per message is held close to constant, c.f. Section 4.4.1, should be considered for future work.

When the e-mail structural feature set and the HTML tag feature set were added separately to the stylistic feature sets, no improvements in classification efficiency were achieved. However, when both were combined with the stylistic features sets there was a marked improvement in classification efficiency. This indicated that neither feature set by itself was sufficient to improve discrimination of e-mail authorship. When acting in concert, there was obviously some interplay between the two sets of features for this group of authors. In all further experiments using e-mail data, both the e-mail structural feature set and the HTML tag feature set were combined with the stylistic feature sets.

As foreshadowed in Section 3.4.3, two document based features (D) were added to the feature sets for tests involving e-mail data. Testing of the document based features was performed using the *email4* data set. Table 5.2 also shows the results of this experiment. The addition of these features was also successful in improving classification efficiency. As the new document based features also improved the classification efficiency, they were included in all further experiments where e-mail data was analysed.

It is not surprising that these document based features add discriminatory power to the separation of authorship classes. The average sentence length was a feature that was one of the first stylometric features used in studies of authorship (Mendenhall, 1887). Even though other studies since then have discounted average sentence length as an authorship discriminator, if this feature is used in conjunction with other features, it should aid classifier performance. The proportion of blank lines is a feature related to a person's sense of formality when using e-mail.

Feature Sets	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ (%)
C+F+L+W	16.5	64.9
C+F+L+W+E	15.5	66.6
C+F+L+W+H	17.0	64.5
C+F+L+W+E+H	8.2	82.4
C+F+L+W+E+H+D	6.9	85.2

Data Set: *email4*
Feature Set: Various

Table 5.2: Classification Results for E-mail Data Using Stylistic and E-mail Specific Features

5.1.2 'Chunking' the E-mail Data

An alternative approach to the analysis of authorship attribution of e-mail messages, especially when the messages contain less than the suggested 200 words (see Section 4.4.1) is to concatenate the text from a series of e-mail messages and then split this into text chunks containing an equal number of words. If this approach is used, it will not be possible to use the e-mail specific features identified in Section 3.4.1, especially those features that are used to indicate the presence or absence of some e-mail structure.

The *email4* data set was stripped of all headers, requoted text, greetings, farewells, signatures and attachments and the e-mail messages for each author were joined together. The combined messages were then split into chunks of 200 words, to produce the data set *chunked email*. The *email4* data set was used as a non-chunked benchmark for comparison of the results and these two data sets were tested in parallel. Function words only and a combination of feature sets, as outlined in Table 5.3, were tested. Training errors were recorded at a rate of approximately 4% during the learning

phase of these experiments. The results show that the weighted macro-averaged F_1 values for both sets of data were similar, indicating that chunking of the data did not contribute any improvement. The weighted macro-averaged error rates, however, were significantly lower.

The F_1 value is, as discussed in Section 3.1.1, the harmonic mean of the precision and recall values. If one of these values for a particular class is much lower than the other, especially if it falls below 50%, the F_1 value falls significantly. When the individual class precision and recall measures for this experiment were inspected, the F_1 results were found to be affected by low recall results, indicating that a large number of false negatives had been assigned during the testing of the data. This may be due to 'unbalancing' the authorship model through highly different numbers of data points in the positive and negative classes in the 'one against all' approach.

When these results are compared with results of the full combination of stylistic and e-mail structural features (Table 5.3) for e-mail messages considered individually, it is clear that even though the chunked e-mail messages have a lower weighted macro-averaged error rate, the results for chunked e-mail messages do not come close to the 85.2% value for $\overline{F}_1^{(M)}$ that was achieved. All further experimentation was carried out using e-mail messages as individual data points.

5.2 In Search of Improved Classification

While the classification of plain text documents in the baseline experiments reported in Section 4.4 met the aim of a weighted macro-averaged F_1 result of 85% or better, the results from the experiments conducted in the previous section using e-mail data only just reached this level. Further investigations were warranted to try and find marginal improvement.

Feature Sets	Separate E-mail Messages		200 Word E-mail Chunks	
	$\overline{\mathbf{E}}^{(\mathrm{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathrm{M})}$ (%)	$\overline{\mathbf{E}}^{(\mathrm{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathrm{M})}$ (%)
F	19.8	57.1	12.5	58.9
C+D+F+L+W	17.3	62.9	10.0	64.4
C+D+F+L+W+E+H	6.9	85.2	-	-

Data Sets: *email4* and *chunked email*
Feature Set: Various

Table 5.3: Comparison of Results for Chunked and Non-chunked E-mail Messages

More and better features could be one approach. Although the function words were consistently the best set of context free features throughout the research, there was no guarantee that the set of function words being used was the optimal set. Further experiments with function words were warranted and are reported below. Another approach could be to tune the parameters of the SVM.

5.2.1 Function Word Experiments

The experimental results have shown that the function word feature set has been the best individual context free feature set for short (approximately 200 word) messages. It was postulated that a larger set of function words may further improve the function word classification performance. A larger set of function words compiled by Higgins (n.d.) was used for testing. The *thesis* data set and the *email4* data set were used to determine the effect of the larger set of function words. Both data sets were used to compare the effect of these extra features on the baseline results and on the e-mail data.

When the tests were run on the *thesis* data set with:

i. function words only; and

ii. function words plus other stylistic features

- as reported in Table 5.4, there is an improvement in classification efficiency for all chunk sizes. The same effect, however, was not seen when the *email4* data set was used. Table 5.5 shows that there was a reduction in classification efficiency when this larger set of function words was used in isolation and no significant improvement when it was combined with all other feature sets.

Feature Sets	**Chunk Size**	**Original Set**		**Large Set**	
		$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ **(%)**
Function Words Only	1000	2.3	0.8	96.4	98.7
	500	4.7	2.5	92.6	95.8
	200	9.6	7.2	84.6	88.5
	100	15.1	11.4	76.1	81.8
Function Words+ C+L+W	1000	0.4	0.4	99.3	99.3
	500	2.3	1.6	96.3	97.4
	200	3.6	4.0	94.5	93.8
	100	7.6	6.4	88.1	90.0

Data Set: *thesis*
Feature Set: Various

Table 5.4: Comparison of Results for Original and Large Function Word Sets for the *thesis* Data Set

5.2.2 Effect of Function Word Part of Speech on Classification

In order to see if some of the function words had more discriminatory potential than others it was decided to split the function words from the new larger set into subsets based on their “part of speech” - pronouns, conjunctions etc. These subsets were tested

Feature Sets	Original Set		Large Set	
	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F_1}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F_1}}^{(\mathbf{M})}$ (%)
F	19.8	57.1	20.2	55.0
C+W+L+F	16.5	64.9	17.3	60.5
F+D+E+H	9.7	79.7	9.6	79.1
C+F+L+W+D+E+H	7.9	82.4	8.2	82.0

Data Set: *email4*
Feature Set: Various

Table 5.5: Comparison of Results for Original and Large Function Word Sets for the *email4* Data Set

as individual feature sets and in combination with other feature sets using the *thesis* data set and the *email4* data set.

The results on plain text data from the *thesis* data set are shown in Table 5.6 and results for the *email4* data set are shown in Table 5.7. The results from both forms of data showed that when function words alone were used as the features, the function word subsets that were best performed were the prepositions and pronouns. When other stylistic and structural features were added to the features used, the auxiliary verb set was one of the best performed feature sets.

No single part of speech subset of function words was as well performed as the original set of function words, indicating that a mixture of parts of speech is required for reliable classification. The original set of function words is a mixture of different parts of speech and this is probably why it has performed so well.

All subsequent testing of e-mail data for this body of research was performed with this original function word set, as there was no compelling evidence shown by these tests to suggest that the extended set of function words would improve classification efficiency.

Function Word Set	Function Words Only		Function Words+ C+L+W	
	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ (%)
Adverbs	32.9	51.0	7.4	88.9
Auxiliary Verbs	34.3	49.9	6.6	90.1
Prepositions	23.7	64.1	6.8	89.9
Pronouns	20.9	68.0	8.5	87.0
Numbers	34.9	27.0	7.7	88.3
Original Set	9.6	84.6	7.2	88.5
Large Set	3.6	94.5	4.0	93.8

Data Set: *thesis*
Feature Set: Function Words

Table 5.6: Comparative Results for Different Function Word Sets for the *thesis* Data Set

5.2.3 Effect of SVM Kernel Function Parameters

The SVM parameters were investigated as part of the baseline experiments conducted on plain text document chunks as reported in Section 4.5. Similar investigations were carried out for e-mail data using the *email4* data set. The baseline experiments showed that the polynomial kernel function with degree equal to three gave satisfactory results. As the polynomial kernel was used for all further experiments with e-mail data, only the parameters of this kernel were investigated. The results of the tests investigating the effect of polynomial degree on classification results are reported in Table 5.8.

These results show that the best classification is achieved when the degree was set to three. This result is again not necessarily universal for all data sets (c.f. Section 4.5). The setting for the degree of the polynomial kernel for all further experiments with e-mail data was three.

Function Word Set	Function Words Only		Function Words+ E-mail Features	
	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ (%)
Adverbs	29.6	31.6	16.4	65.2
Auxiliary Verbs	28.7	41.2	15.4	68.5
Prepositions	27.9	43.9	14.6	69.4
Pronouns	27.9	45.1	15.2	70.2
Numbers	25.4	12.4	14.9	69.4
Original Set	19.8	57.1	9.6	79.7
Large Set	20.2	55.0	9.7	79.1

Function Word Set	Function Words+ E-mail Features+ Stylistic Features	
	$\overline{\mathbf{E}}^{(\mathbf{M})}$ (%)	$\overline{\mathbf{F}}_{\mathbf{1}}^{(\mathbf{M})}$ (%)
Adverbs	10.6	78.6
Auxiliary Verbs	8.8	81.5
Prepositions	10.0	78.4
Pronouns	9.8	79.3
Numbers	11.8	75.8
Original Set	6.9	85.2
Large Set	7.9	82.0

Data Set: *email4*
Feature Set: Function Words

Table 5.7: Comparative Results for Different Function Word Sets for the *email4* Data Set

No training errors were encountered in these tests and as this was similar behaviour to that observed for the baseline tests on plain text data, the effect on classification of the cost parameter, C, was not investigated.

Degree of Polynomial	$\overline{E}^{(M)}$ (%)	$\overline{F}_1^{(M)}$ (%)
1	8.4	82.0
2	7.8	83.2
3	6.9	85.2
4	7.1	85.0
5	7.4	84.5

Data Set: *email4*
Feature Set: C+D+F+L+W+E+H

Table 5.8: Effect of Degree on Polynomial Kernel Function for the *email4* Data Set

5.3 The Effect of Topic on E-mail Authorship Classification

An important aspect of authorship attribution is that the features and the technique being used should be immune to the effect of the topic or subject of the document being classified. A series of experiments was undertaken to ascertain if the topic of text in e-mail messages had an effect on the classification of authorship when being classified by SVM. Details of the *discussion* data set used for topic experiments are contained in Section 3.6.

An initial test was carried out using merged data from all three discussion topics to determine the baseline classification result for the three authors. These tests were performed with all of the features from the stylistic and e-mail structural feature sets. The baseline test was carried out using stratified 10-fold cross validation. The baseline classification results for the three authors are shown in Table 5.9. The largest of the three topic subsets was the *movies* topic and a baseline result was also generated for the three authors on this data set in a similar fashion. The results of this classification

are shown in Table 5.10. The classification results for the whole data set are similar to those of the *movies* topic subset, indicating that topical words do not seem to be affecting classification performance.

Measure	**Authorship Class**		
	1	**2**	**3**
Data Points	30	62	63
Error Rate (%)	7.1	7.7	7.1
Precision (%)	100.0	83.8	93.8
Recall (%)	63.3	98.3	89.6
F_1 (%)	77.6	90.5	91.6
$\overline{E}^{(M)}$(%)			7.3
$\overline{F}_1^{(M)}$ (%)			85.6

Data Set: *discussion*
Feature Set: C+D+F+L+W+E+H

Table 5.9: Classification Results for the *discussion* Data Set

Measure	**Authorship Class**		
	1	**2**	**3**
Data Points	15	21	23
Error Rate (%)	16.9	11.9	6.8
Precision (%)	100.0	79.2	88.0
Recall (%)	33.3	90.5	95.7
F_1 (%)	50.0	84.4	91.7
$\overline{E}^{(M)}$(%)			12.2
$\overline{F}_1^{(M)}$ (%)			73.8

Data Set: *discussion* - *movies* topic
Feature Set: C+D+F+L+W+E+H

Table 5.10: Classification Results for the *movies* Topic from the *discussion* Data Set

In order to further test that topic does not affect the generalisation ability of the models it was decided to use the *movies* topic as the training data and to test these models using the e-mail messages from the other topic sets, *food* and *travel*. If topic does not affect classification performance with the same features as in Tables 5.9 and 5.10, then generalisation performance for each author should be similar to that obtained when all data from each author was used.

A single classification model was then learned for each of the three authors using just one of the topics, *movies*, as the training data set. These models were used to predict authorship classes for the e-mail messages from the other topic subsets, *food* and *travel*. The results of this test are shown in Table 5.11. This inter-topic result shows that the classification of authorship is still approximately 85% successful when e-mail messages from different topics are used. This is another indication that the topic of the e-mail messages does not affect the classification of authorship.

It will be noted from this table that the results are poor for Author 1. The number of e-mail messages from Author 1 in this data set was less than that for the other two authors as shown in Table 3.11. The authorship model for this author learned from the *movies* topic was learned from only 15 e-mail messages, which is less than the recommended minimum of 20 ascertained by experimentation in Section 4.4.2.

5.4 Profiling the Author: Authorship Characterisation

As discussed in Section 3.5, a method to reduce the number of suspects in a forensic investigation involving e-mail authorship attribution is desirable. One approach is to build sociolinguistic profiles for authors. These profiles could then be used to identify the e-mail message author's gender, language background, age group and education level.

Topic	Authorship Class		
	1	2	3
	F_1 (%)	F_1 (%)	F_1 (%)
food	28.6	87.5	88.5
travel	50.0	95.2	100.0

Data Set: *discussion*
Feature Set: C+D+F+L+W+E+H

Table 5.11: Classification Results for the *food* and *travel* Topics from the *discussion* Data Set Using the *movies* Topic Classifier Models

Cohort profiles would have to be learned from a large number of authors to ensure that discriminatory features were cohort based rather than author based.

5.4.1 Gender Experiments

The gender cohort is an ideal starting place for investigations of authorship characterisation. The cohort contains only two classes and some previous research has empirically identified features which distinguish male and female writings.

As in the authorship attribution problem, the number of words in the e-mail messages contributing to the cohort profile will have an effect on the classification results. E-mail messages containing no words will not be able to be discriminated as there is no stylistic evidence contained in them.

Another parameter that will impact on the performance of authorship characterisation will be the number of e-mail messages in the cohort. A small number of messages is unlikely to produce a general model of authorship gender.

Since, as Section 3.6 foreshadowed, very few gender-identifiable messages above 200 words were available, experiments were undertaken where the minimum word

count per e-mail message was varied between 50, 100, 150 and 200, and the number of e-mail messages in the cohorts was varied between 50 and the maximum number of available e-mail messages.

The results of testing these variables are shown in Table 5.12 and are graphically represented in Figure 5-1. Training errors of approximately 5% were recorded during learining. The proportion of support vectors in these tests was between 60% and 80% indicating that these models were more fitted to the training data than those from the authorship experiments on both plain text chunks and e-mail messages.

It can be seen from these results that there is a general trend towards better results when both the numbers of messages per cohort and the minimum word count for each message in the cohorts is increased. This increase in the amount of raw data in the cohorts, as expected, leads to a better model of gender. A cohort size of at least 500 e-mail messages seems to be required for gender classification, although better performance was indicated when a larger cohort of 2000 e-mail messages was used.

A study of the impact of features on gender classification was undertaken by measuring the result for all feature sets and then removing one of the sets at a time to measure the impact. The results of the investigations into the discriminatory power of the various feature sets are shown in Table 5.13. These results show that when e-mail structure features, function words, HTML tags or word length frequency distribution are removed from the full feature set, there is a statistically significant decrease in classification performance. There was no statistically significant gain from the addition of the set of gender based features.

This is a similar result to that from the authorship attribution experiments, where the same combination of feature sets gave the best classification results. There may

be features within the feature sets that do not contribute to the discrimination, but they would have to be determined from an exhaustive feature set sub-selection experiment.

The gender preferential feature set did not improve the classification performance over the gender cohort. These results also showed that the most significant decrease in classification performance came from the removal of the function word features from the feature sets. It may be that some of the function words in the list being used are more powerful discriminators of gender than those identified in the literature to date.

Messages per Cohort	**Minimum Word Count**							
	50		**100**		**150**		**200**	
	$\overline{\mathbf{E}}^{(\mathrm{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathrm{M})}$ **(%)**	$\overline{\mathbf{E}}^{(\mathrm{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathrm{M})}$ **(%)**	$\overline{\mathbf{E}}^{(\mathrm{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathrm{M})}$ **(%)**	$\overline{\mathbf{E}}^{(\mathrm{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathrm{M})}$ **(%)**
50	35.5	64.4	37.7	62.2	42.9	57.1	40.2	59.8
100	31.6	68.4	36.0	64.0	43.1	56.8	35.0	65.0
150	33.7	66.3	39.2	60.8	38.9	61.1	36.6	63.3
200	35.2	64.8	38.5	61.5	37.8	62.2	36.1	63.8
250	32.7	67.3	36.8	63.2	34.8	65.2	34.3	65.7
300	33.6	66.4	32.4	67.6	33.4	66.6	32.7	67.3
400	32.5	67.5	31.3	68.7	29.8	70.2	-	-
500	32.7	67.3	29.9	70.1	29.7	70.3	-	-
600	33.3	66.7	29.9	70.1	-	-	-	-
750	30.8	69.2	29.2	70.8	-	-	-	-
1000	30.5	69.4	29.0	71.1	-	-	-	-
1,250	31.0	68.8	-	-	-	-	-	-
1,500	30.2	69.8	-	-	-	-	-	-
2000	27.9	72.1	-	-	-	-	-	-

Data Set: *gender*
Feature Set: C+D+E+F+G+H+L+W

Table 5.12: Effect of Cohort Size on Gender

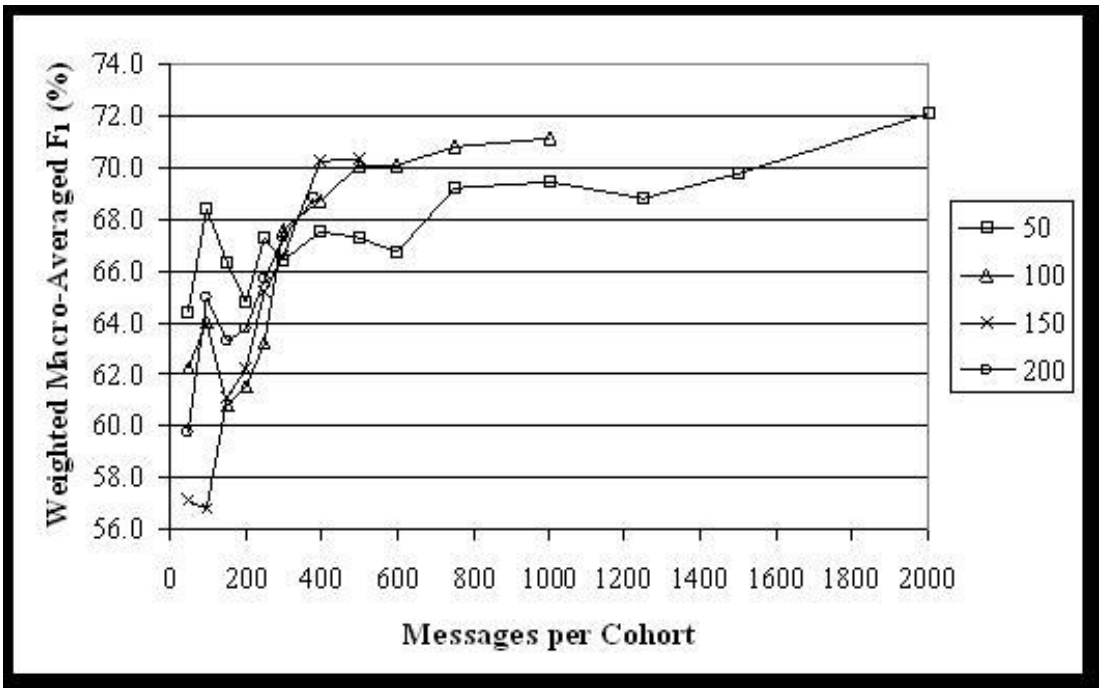


Figure 5-1: Effect of Cohort Size on Gender

Feature Sets	$\overline{E}^{(M)}$ (%)	$\overline{F}_1^{(M)}$ (%)
All	29.9	70.1
Character features removed	30.0	70.0
Document features removed	30.2	69.8
E-mail Structure features removed	31.9	68.1
Function words removed	36.0	64.0
HTML tags removed	30.6	69.4
Word length distribution removed	32.6	67.4
Word based features removed	30.4	69.6
Gender based features added	29.8	70.2

Data Set: *gender*
Feature Set: Various

Table 5.13: Effect of Feature Sets on Classification of Gender

5.4.2 Language Background Experiments

Similar experiments to those conducted for the gender cohorts were conducted for English as a native language (ENL) versus English as a second language (ESL) authored messages. The variables studied were the minimum number of words per message and the number of messages in the cohort.

It was difficult to obtain as many ESL authored e-mail messages as ENL messages. This has restricted the cohort sizes able to be used in the testing. The results of these experiments are shown in Table 5.14 and graphically in Figure 5-2.

Messages per Cohort	**Minimum Word Count**							
	50		**100**		**150**		**200**	
	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{E}}^{(\mathbf{M})}$ **(%)**	$\overline{\mathbf{F_1}}^{(\mathbf{M})}$ **(%)**
50	40.1	59.7	36.4	63.3	26.0	74.0	28.6	71.0
100	34.4	65.6	29.6	70.4	24.0	76.0	28.3	77.3
150	29.7	70.3	30.4	69.5	26.3	73.6	19.7	80.3
200	29.4	70.6	29.7	70.3	25.3	74.7	19.9	80.8
250	30.3	69.7	29.3	70.7	24.7	75.1	-	-
300	27.5	72.5	28.7	71.2	-	-	-	-
400	29.1	70.9	26.9	73.0	-	-	-	-
500	27.5	72.5	-	-	-	-	-	-
600	26.7	73.3	-	-	-	-	-	-
700	25.4	74.6	-	-	-	-	-	-

Data Set: *language*
Feature Set: C+D+E+F+H+L+W

Table 5.14: Effect of Cohort Size on Language

Similar results as for the gender cohort tests were observed. An improvement in classification efficiency is seen as the number of messages in the cohort increases and as the minimum word count per e-mail message increases. The classification

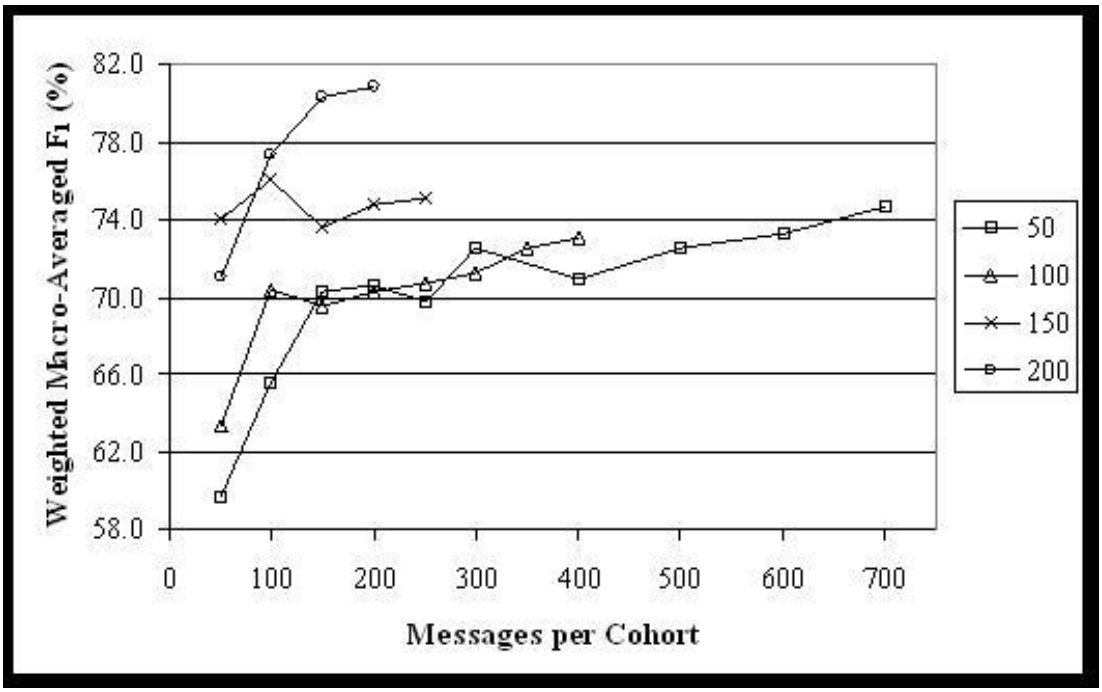


Figure 5-2: Effect of Cohort Size on Language

of language background appears to be more effective than that of gender with the feature sets being used. This is possibly due to the influence of the structure of the native language that an ESL author is familiar with. It may be the case that the native language of the author is more regimented with fewer deviations from formality than English.

5.5 Chapter Summary

This chapter has discussed the experimental sequence used to arrive at the optimal approach to authorship analysis of e-mail messages, and has discussed the possibility of authorship characterisation or cohort profiling for authorship for the gender and language background cohorts.

The major findings from this work include:

- While stylistic features are effective for plain text as shown in Chapter 4, they are not sufficient for e-mail messages. However, the addition of structural features from e-mail messages aids the classification of authorship.

- The effect of topic on authorship attribution was investigated and found to have no effect, indicating that attribution can rely on the style in which authors write and the structure which they add to their messages.

- Authorship characterisation has been attempted and showed promising results for gender and language background cohorts. The classification performance improved as more data points containing higher numbers of words are used. It is interesting to note that the features used for authorship analysis provide the discriminatory power for these cohorts rather than the extra gender specific features identified from the literature.

The final chapter of the thesis provides a summary of the major conclusions and discoveries made by this research and discusses the impact of this work on related fields of study. It also suggests some possible extensions of the work for the future.

Chapter 6

Conclusions and Further Work

This project set out with the objective of showing whether literary stylistics could be applied successfully in a very different domain: identifying the authors of anonymous e-mail. This chapter provides a summary of the major conclusions and discoveries from this research. A discussion of the future implications of this body of research is also given.

6.1 Conclusions

At the end of Chapter 2, the thesis posed these research questions.

- Can the work that has been carried out in stylometry research based on literary works be applied to the text contained in e-mail messages?
- What are the best features for authorship attribution of e-mail messages?
- How many words are required in e-mail messages to make authorship attribution successful?
- Is there some way of reducing a large list of suspected authors, e.g. 50, to a smaller list of authors, e.g. 5, for selection of one author most likely to have written an e-mail message?

We can conclude the following from the results presented.

On the applicability of stylometry research, the project has demonstrated that authorship attribution of text documents can be successfully performed using a Support Vector Machine with a minimum weighted macro-averaged F_1 as large as 85%.

The research has also clarified the best features for authorship attribution of e-mail messages. The baseline tests have shown that function words were consistently the best individual feature set that is independent of topic for the analysis of authorship. The results also showed that better results were obtained by adding extra feature sets to the function word feature set. The 2-gram feature set showed good discrimination but the experimental results indicated that this set is biased toward discrimination of content. The use of word collocations was unsuccessful and it is thought that this is due to the small number of words in a typical e-mail message, leading to noisy feature values. While stylistic features are effective for plain text, they are not sufficient to reach the 85% F_1 target for e-mail messages, but the addition of structural features from e-mail messages aided the classification of authorship.

The project has also determined from the baseline tests that it was possible to attribute the authorship of text with the chosen feature sets on chunks containing as few as 200 to 250 words, with possibly only 20 data points per authorship class. This 200 word limit is feasible for many e-mail messages.

In an attempt to reduce the number of authors to be compared, authorship characterisation by sociolinguistic cohort was trialled for gender and language background cohorts. Although these experiments were hampered by unsuitably small e-mail texts (less than 200 words), enough positive results were gained to make authorship characterisation a promising field for further research. The chosen gender preferential features did not improve classification over the markers already demonstrating success in

individual authorship identification. However, many of the latter clearly have embedded gender bias, most notably among the function words, and need further study.

Much useful experience was gained with the optimisation of SVM^{light}. The SVM^{light} implementation works quite effectively when the default parameters are used, though slight gains in classification performance can be made if the parameters of the classifier are optimised for the problem domain being used. Experimentation with these parameters found that the 3^{rd} order polynomial kernel function with default scalar and constant values was best in the current problem domain.

The effect of topic on authorship attribution was investigated and found to have no effect as formulated in these experiments. Attribution can still successfully be done using markers based on the style in which authors write and the structure which they add to their messages.

6.2 Implications for Further Work

It is clear that the results of this research will be useful only in informal situations requiring authorship identification. Without significant improvements, evidence based on SVM-derived analysis would not reach court room standards. However, the project has laid a firm groundwork for these improvements.

Most research into authorship attribution has used text chunks with at least 1000 words. The work conducted here has significantly lowered this text size. Better classification results can be obtained when more words are used but reliable results have been achieved on plain text documents containing between 200 and 250 words.

The suggestion by Holmes (1998), that the more features used to define an authorship pattern the better, has been validated through use of the Support Vector Machine

learning algorithm. SVMs have not been widely used in authorship attribution studies and could prove to be a better tool for attribution studies than the other machine learning algorithms used to date.

This research has not selected the optimal set of features for building authorship patterns, but a feature set subselection approach as discussed by Witten and Frank (2000) could be used to discover the most significant features for authorship attribution in general or for specific authors.

It is felt that to improve further the results obtained from this research that some extension must be made to the frequencies of character and word based features used here. A possible approach for the future would be to syntactically mark up the text using an automated natural language processing (NLP) tagger. Part of speech tags could be counted and used as features, or a model of an author's preferred grammatical sentence structure could be discovered, e.g. by approximating this structure with Hidden Markov Models.

While the basis of a technique for authorship characterisation has been demonstrated, further improvements must be made if this is to be made more accurate. Further investigations into gender specific or language background specific features are warranted. Grammatical analysis using NLP may be useful for this work as well. While a large number of authors have been used in the generation of cohorts, an even larger number of authors may be needed for the generation of more reliable models. We could expect that a data set of large e-mail messages would give more reliable results.

Further sociolinguistic cohorts could be investigated. Rayson et al. (1997) have discovered differences in the writings of people from different age groups and social class backgrounds. Other cohorts such as education level could also be examined. The investigation of English as a native language versus second language could be extended

to a more refined level based on the country or continent from which the ESL authors come. There is an indication in the literature that there are many "Englishes" spoken around the world (Bhatt, 2001). With sufficient data, models could be generated for these. This opens the exciting prospect that a randomly selected text of more than 100 to 200 words could identify the writer as, e.g., male, at least 40 years old, of an English-speaking background and never educated beyond primary school.

While the work conducted for this body of research has focussed on e-mail messages as the data source, it should be possible to apply the technique to other areas of text analysis. Text plagiarism should benefit from a similar SVM based approach. A second example could be the identification of anonymous writers in Internet Chat Rooms. This is a very different genre of writing from e-mail's, but stylometric authorship patterns would persist here, especially the use of function words.

In 1998 the stylistics researcher David Holmes said that "although as yet, no definitive methodology or technique has emerged, statisticians are coming closer to stylometry's 'holy grail', the fully automated identifier!" (Holmes, 1998). It is hoped that the research work reported in this thesis has brought the stylometric 'holy grail' slightly nearer.

Glossary

Accuracy Accuracy is calculated from the number of correct classifications made by the classifier.

Authorship Analysis Any study that involves studying the authorship of a piece of text. The text may be literature, prose, poetry, communication.

Authorship Attribution Authorship attribution involves the identification of the author of a piece of text. This can be done traditionally using handwriting comparisons or non-traditionally using stylometry.

Authorship Characterisation Authorship characterisation involves determining some sociolinguistic characteristic of an author, such as gender, age, education level etc.

Badge Word A word that is preferred by a particular author relative to other authors.

Collocation A collocation is a combination of two words together or separated by a certain number of words.

Entropy E The entropy of a piece of text is defined as,

$$E = \sum_{i=1}^{N} V_i \Big(- \log_2 \frac{i}{N} \Big) \frac{i}{N}$$

where V_i is the number of types that occur i times in the text and N is the number of tokens.

Error Rate The error rate is the number of misclassifications made on a Test Set. The error rate can be represented as a value between 0.0 and 1.0 or as a percentage.

$\mathbf{F_1}$ Measure The F_1 measure is defined as the break even point between Precision and Recall, when β is set to 1. It is calculated as the harmonic mean of the recall and precision.

$\mathbf{F_\beta}$ Measure The F_β measure is used to calculate a combined measure from the Precision and Recall value. F_β is defined as,

$$F_\beta = \frac{(1+\beta^2)RP}{R+\beta^2 P}$$

where $R =$ recall and $P =$ precision.

Fluke Word A word that is not preferred by a particular author relative to other authors.

Grammatical Accidence The study of changes in the form of words by internal modification for the expression of tense, person, case, number etc.

Hapax Dislegomena *Hapax dislegomena* are words that are used twice in any text.

Hapax Legomena *Hapax legomena* are words that are used once only in any text.

HTML Hyper-Text Markup Language is a language consisting of formatting elements that are used to format World Wide Web pages and e-mail messages.

k-fold Cross Validation k-fold cross validation is used to test the effectiveness of a learnt classifier on a small data set. The data is randomly split across the k folds ensuring that each fold has the same or almost the same number of data points in it. One fold is used as a test set while the remaining $k-1$ folds are used as a training set to learn a classifier. This allows k classifiers to be learnt from the data set. Each fold is used once as a test set. Typically k is set to 10 and 10-fold cross validation is performed.

Machine Learning Machine learning is the field of study that investigates the ability of a learning algorithm to take input and represent it as knowledge, allowing the algorithm to generalise about unseen data.

One Against All A method of producing classifiers from multi-class data sets. For a data set containing n classes, n data sets are produced, with the i^th class $(i = 1 \ldots n)$ being made the positive class and all others part of the negative class. N classifier models are then learnt from the training set. Test data can then be classified by each model to determine which class the test data belongs to.

One Against One A method of producing classifiers from multi-class data sets. For a data set containing n classes, each class in the data set is combined with one other class from the data set in turn to produce $n(n-1)$ training sets to produce classifier models. Test data can then be classified by each learnt classifier model to determine the class to which the test data belongs.

Precision Precision is a measure of classifier performance. It measures the impact of false positive assignments on the classifier's performance.

Recall Recall is a measure of classifier performance. It measures the impact of false negative assignments on the classifier's performance.

Simpson's Index D Simpson's D is defined as,

$$D = \sum_{i=1}^{V} V_i \frac{i}{N} \frac{i-1}{N-1}$$

where V is the number of types, V_i is the number of types that occur i times in the text and N is the number of tokens.

Stratification Stratification can be performed on the data set when performing k-fold Cross Validation. The data from each class in the set are randomly split evenly or as evenly as possible among the k folds. This ensures that each class in the data set is represented in a similar ratio in the Test Set and Training Set.

Stylometry The statistical analysis of literary style. It makes the assumption that there is an unconscious aspect to one's style of writing which cannot be manipulated and possesses features that are quantifiable and may be distinctive.

Support Vector Machine A type of classifier used in machine learning. SVMs were developed byVapnik (1995) to perform a classification of data that is either in a positive class or not in that positive class. From a Training Set, a classifier model can be learnt that can then be used to classify new data as belonging to the positive class or not. SVMs are suited to sparse data sets with many attributes or features. They do not suffer from overtraining. SVMs find the hyperplane which separates the positive and negative training examples with maximum margin. The data points closest to the hyperplane are called support vectors. Originally the kernel function to separate the data was linear but any function can be used

to separate the data and polynomial or radial basis function kernels are typically used. The data is mapped into some higher dimensional space to effect the separation.

Test Set The test set is the set of data that is used to test the Accuracy of classifier that has been learnt on the training set. The test set and the training set must be disjoint.

Training Set The training set is the set of data that a machine learning classifier is learnt on. The training set and the test set must be disjoint.

Type-Token Ratio R As we progress through a text, N increases from 1 to the total number of word tokens in the text i.e N is the number of word tokens. A word token is an instance of a word type. The total number of word types or the vocabulary size is signified with V. The type-token ratio R is defined,

$$R = \frac{V}{N}$$

User Agent An application used to construct, edit, send and receive e-mail messages. Common User Agents include Eudora, Microsoft Outlook, Netscape Messenger and Pine.

Yule's Characteristic K Yule's K is defined as,

$$K = 10^4 \Bigg[-\frac{1}{N} + \sum_{i=1}^{V} V_i \Big(\frac{i}{N}\Big)^2 \Bigg]$$

where V is the number of types, V_i is the number of types that occur i times in the text and N is the number of tokens.

Appendix A

Feature Sets

Definitions

$N =$ total number of *tokens* (i.e., words)

$V =$ total number of *types* (i.e., distinct words)

$C =$ total number of characters

$H =$ total number of HTML tags in the e-mail body

The count of *hapax legomena* is defined as the number of types that occur only once in the text.

Features W_7 to W_{19} are defined in Tweedie and Baayen (1998).

A.1 Document Based Features

Feature Number	**Feature Description**
D_1	Number of blank lines/total number of lines
D_2	Average sentence length (number of words)

A.2 Word Based Features

Feature Number	Feature Description
W_1	Average word length
W_2	Vocabulary richness i.e., V/N
W_3	Total number of function words/N
W_4	Total number of short words/N (word length ≤ 3)
W_5	Count of *hapax legomena*/N
W_6	Count of *hapax legomena*/V
W_7	Guirad's R
W_8	Herdan's C
W_9	Herdan's V
W_{10}	Rubet's K
W_{11}	Maas' A
W_{12}	Dugast's U
W_{13}	Luk"janenkov and Neistoj's measure
W_{14}	Brunet's W
W_{15}	Honore's H
W_{16}	Sichel's S
W_{17}	Yule's K
W_{18}	Simpson's D
W_{19}	Entropy measure

Guirad's R = $\frac{V}{\sqrt{N}}$

Herdan's C = $\frac{\log_{10} V}{\log_{10} N}$

Herdan's V = $\sum_{i=1}^{V} V_i \frac{i^2}{N^2}$

Rubet's K = $\frac{\log_{10} V}{\log_{10}(\log_{10} N)}$

Maas' A = $\sqrt{\frac{\log_{10} N - \log_{10} V}{(\log_{10} N)^2}}$

Dugast's U = $\frac{(\log_{10} N)^2}{\log_{10} N - \log_{10} V}$

Luk"janenkov and Neistoj's measure = $1 - \frac{V^2}{V^2 \times \log_{10} N}$

Brunet's W = $N^{V^{-0.172}}$

Honore's H = $\frac{100 \times \log_{10} N}{1 - \frac{count\ of\ hapax\ legomena}{V}}$

Sichel's S = $\frac{count\ of\ hapax\ dislegomena}{V}$

Yule's K = $10^4 \left[-\frac{1}{N} + \sum_{i=1}^{V} V_i \left(\frac{i}{N}\right)^2 \right]$

Simpson's D = $\sum_{i=1}^{V} V_i \frac{i}{N} \frac{i-1}{N-1}$

Entropy = $\sum_{i=1}^{N} V_i \left(-\log_{10} \frac{i}{N} \right) \frac{i}{N}$

A.3 Character Based Features

Feature Number	Feature Description
C_1	Number of characters in words/C
C_2	Number of alphabetic characters/C
C_3	Number of upper-case characters in words/C
C_4	Number of digit characters in words/C
C_5	Number of white-space characters/C
C_6	Number of spaces/C
C_7	Number of spaces/Number white-space chars
C_8	Number of tab spaces/C
C_9	Number of tab spaces/Number white-space chars
C_{10}	Number of punctuation characters/C

A.4 Function Word Frequency Distribution

Feature Number	Feature Description
$F_1 \ldots F_{122}$	Function word frequency / N

Original Function Word List

This list of function words was sourced from Craig (1999).

a	about	after	all	am	also
an	and	any	are	as	at
be	been	before	best	better	both
but	by	can	cannot	come	comes
could	did	do	does	done	first
for	from	give	go	had	has
have	he	her	here	him	his
how	i	if	in	into	is
it	know	let	like	may	me
men	might	mine	much	must	my
need	no	none	not	nothing	now
of	on	once	one	or	our
out	put	see	shall	she	should
since	so	stay	still	such	take
tell	than	that	the	their	theirs
them	then	there	these	they	this
those	though	time	to	too	up
upon	us	very	was	we	well
were	what	when	which	who	whose
why	will	with	yes	yet	you
your	yours				

Extended Function Word List

This list of function words was sourced from Higgins (n.d.)

Adverbs

again	ago	almost	already	also	always
anywhere	back	else	even	ever	everywhere
far	hence	here	hither	how	however
near	nearby	nearly	never	not	now
nowhere	often	only	quite	rather	sometimes
somewhere	soon	still	then	thence	there
therefore	thither	thus	today	tomorrow	too
underneath	very	when	whence	where	whither
why	yes	yesterday	yet		

Auxiliary Verbs and Contractions

am	are	aren’t	be	been	being
can	can’t	could	couldn’t	did	didn’t
do	does	doesn’t	doing	done	don’t
get	gets	getting	got	had	hadn’t
has	hasn’t	have	haven’t	having	he’d
he’ll	he’s	i’d	i’ll	i’m	is
i’ve	isn’t	it’s	may	might	must
mustn’t	ought	oughtn’t	shall	shan’t	she’d
she’ll	she’s	should	shouldn’t	that’s	they’d
they’ll	they’re	was	wasn’t	we’d	we’ll
were	we’re	weren’t	we’ve	will	won’t
would	wouldn’t	you’d	you’ll	you’re	you’ve

Prepositions and Conjunctions

about	above	after	along	although	among
and	around	as	at	before	below
beneath	beside	between	beyond	but	by
down	during	except	for	from	if
in	into	near	nor	of	off
on	or	out	over	round	since
so	than	that	though	through	till
to	towards	under	unless	until	up
whereas	while	with	within	without	

Determiners and Pronouns

a	all	an	another	any	anybody
anything	both	each	either	enough	every
everybody	everyone	everything	few	fewer	he
her	hers	herself	him	himself	his
i	its	itself	less	many	me
mine	more	most	much	my	myself
neither	no	nobody	none	noone	nothing
other	others	our	ours	ourselves	she
some	somebody	someone	something	such	that
the	their	theirs	them	themselves	these
they	this	those	us	we	what
which	who	whom	whose	you	yours
yourself	yourselves				

Numbers

billion	billionth	eight	eighteen	eighteenth	eighth
eightieth	eighty	eleven	eleventh	fifteen	fifteenth
fifth	fiftieth	fifty	first	five	fortieth
forty	four	fourteen	fourteenth	fourth	hundred
hundredth	last	million	millionth	next	nine
nineteenth	ninetieth	ninety	ninth	once	one
second	seven	seventeen	seventeenth	seventh	seventieth
seventy	six	sixteen	sixteenth	sixth	sixtieth
sixty	ten	tenth	third	thirteen	thirteenth
thirtieth	thirty	thousand	thousandth	three	thrice
twelfth	twelve	twentieth	twenty	twice	two

A.5 Word Length Frequency Distribution

Feature Number	**Feature Description**
$L_1 \ldots L_{30}$	Word length frequency distribution / N

A.6 E-mail Structural Features

Feature Number	**Feature Description**
E_1	Reply status
E_2	Has a greeting acknowledgement
E_3	Uses a farewell acknowledgement
E_4	Contains signature text
E_5	Number of attachments
E_6	Position of re-quoted text within e-mail body

A.7 E-mail Structural Features

Feature Number	Feature Description
H_1	Frequency of <BIGGER> / H
H_2	Frequency of <BOLD> or <B> / H
H_3	Frequency of <CENTER> / H
H_4	Frequency of <COLOR> / H
H_5	Frequency of <FONT> / H
H_6	Frequency of <ITALIC> or <I> / H
H_7	Frequency of <UNDERLINE> or <U> / H

A.8 Gender Specific Features

Feature Number	Feature Description
G_1	Number of words ending with *able* /N
G_2	Number of words ending with *al* /N
G_3	Number of words ending with *ful* /N
G_4	Number of words ending with *ible* /N
G_5	Number of words ending with *ic* /N
G_6	Number of words ending with *ive* /N
G_7	Number of words ending with *less* /N
G_8	Number of words ending with *ly* /N
G_9	Number of words ending with *ous* /N
G_{10}	Number of *sorry* words /N
G_{11}	Number of words starting with *apolog* /N

A.9 Collocation List

and all	and of	and the	and then	all of	are a
are all	are also	are as	are by	are in	are no
are not	are now	are of	are some	are the	are to
as a	as if	as the	as though	as well	at a
at last	at the	be in	be of	be on	be the
be to	can also	can be	can do	can have	can no
can not	can only	can the	could also	could be	could do
could have	could no	could not	could only	could the	did not
did the	did this	did with	do not	do the	do this
do with	for a	for example	for the	get a	get an
get the	had a	had an	had been	had no	had not
had the	had to	has a	has an	has been	has no
has not	has the	has to	have a	have an	have been
have no	have not	have the	have to	in a	in it
in the	in to	is a	is the	it is	may also
may be	may do	may have	may not	may only	might also
might be	might do	might have	might not	might only	must also
must be	must do	must have	must not	must only	of a
of the	on a	on to	on the	shall also	shall be
shall do	shall have	shall no	shall not	shall only	shall the
should also	should be	should do	should have	should no	should not
should only	should the	that is	that it	that the	to a
to be	to go	to the	was a	was an	was as
was in	was not	was of	was on	was the	was to
were a	were an	were as	were in	were not	were of
were on	were the	were to	will also	will be	will do
will have	will no	will not	will only	will the	would also
would be	would do	would have	would no	would not	would only
would the					

Bibliography

I. Androutsopolous, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos. An experimental comparison of Naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In N. J. Belkin, P. Ingwersen, and M.-K. Leong, editors, *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–167, Athens, Greece, 2000a.

I. Androutsopolous, G. Paliouras, V. Karkaletsis, G. Sakkis, C. D. Spyropoulos, and P. Stamatopoulos. Learning to filter spam e-mail: A comparison of Naive Bayesian and a memory-based approach. In H. Zaragoza, P. Gallinari, and M. Rajman, editors, *Workshop on Machine Learning and Textual Information Access at the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 1–13, Lyon, France, 2000b.

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.

R. H. Baayen, F. J. Tweedie, A. Neijt, and L. Krebbers. Back to the cave of shadows: Stylistic fingerprints in authorship attribution. The ALLC/ACH 2000 Conference, University of Glasgow, 2000. WWW Document, URL `http://www2.arts.gla.ac.uk/allcach2k/Programme/session2.html#233`, accessed August 3, 2000.

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*, 11(3):121–131, 1996.

N. Baron. Letters by phone or speech by other means: The linguistics of email. *Language and Communication*, 18:133–170, 1998.

P. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems 9*, pages 134–140, 1997.

D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Physical Review Letters*, 88(4):048702–1–4, 2002a.

D. Benedetto, E. Caglioti, and V. Loreto. On J. Goodman's comment to "Language trees and zipping", 2002b. WWW Document, URL `http://arXiv.org/abs/cond-mat/0203275`, accessed March 12, 2003.

R. M. Bhatt. World Englishes. *Annual Reviews Anthropology*, 30:527–550, 2001.

S. L. Brodsky. *The Expert Expert Witness: More Maxims and Guidelines for Testifying in Court*. American Psychological Association, Washington, DC, 1999.

J. F. Burrows. Computers and the study of literature. In C. Butler, editor, *Computers and Written Text*, Applied Language Studies, pages 167–204. Blackwell, Oxford, 1992.

D. Canter and J. Chester. Investigation into the claim of weighted Cusum in authorship attribution studies. *Forensic Linguistics*, 4(2):252–261, 1997.

CERT. CERTő Advisory CA-2000-04 Love Letter Worm. Carnegie Mellon University, 2000. WWW Document, URL `http://www.cert.org/advisories/CA-2000-04.html`, accessed March 12, 2003.

CERT. CERTő Advisory CA-2001-19 "Code Red" Worm Exploiting Buffer Overflow In IIS Indexing Service DLL, 2001a. WWW Document, URL `http://www.cert.org/advisories/CA-2001-19.html`, accessed March 12, 2003.

CERT. CERTő Advisory CA-2001-22 W32/Sircam Malicious Code, 2001b. WWW Document, URL `http://www.cert.org/advisories/CA-2001-22.html`, accessed March 12, 2003.

CERT. CERTő Advisory CA-2001-26 Nimda Worm, 2001c. WWW Document, URL `http://www.cert.org/advisories/CA-2001-26.html`, accessed March 12, 2003.

C. E. Chaski. Who wrote it?: Steps toward a science of authorship identification. *National Institute of Justice Journal*, 233(September 1997):15–22, 1997.

C. E. Chaski. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1):1–65, 2001.

W. W. Cohen. Learning rules that classify e-mail. In *1996 AAAI Spring Symposium on Machine Learning in Information Access*, 1996.

H. Craig. Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing*, 14(1):103–113, 1999.

C. Crain. The Bard's fingerprints. *Lingua Franca*, 4:29–39, 1998.

E. Crawford, J. Kay, and E. McCreath. Automatic induction of rules for e-mail classification. In *Sixth Australasian Document Computing Symposium*, Coffs Harbour, Australia, 2001.

D. H. Crocker. RFC822 - Standard for the format of ARPA Internet text messages, 1982. WWW Document, URL `http://www.faqs.org/rfcs/rfc822.html`, accessed 26 October, 2000.

P. De-Haan. Analysing for authorship: A guide to the Cusum technique. *Forensic Linguistics*, 5(1):69–76, 1998.

O. de Vel. Mining e-mail authorship. In *Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA, 2000.

H. Drucker, D. Wu, and V. N. Vapnik. Support Vector Machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048–1054, 1999.

R. Efron and B. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

W. E. Y. Elliott and R. J. Valenza. A touchstone for the Bard. *Computers and the Humanities*, 25(4):199–209, 1991a.

W. E. Y. Elliott and R. J. Valenza. Was the Earl of Oxford the true Shakespeare? A computer aided analysis. *Notes and Queries*, 236:501–506, 1991b.

W. E. Y. Elliott and R. J. Valenza. And then there were none: Winnowing the Shakespeare claimants. *Computers and the Humanities*, 30:191–245, 1996.

W. E. Y. Elliott and R. J. Valenza. The Professor doth protest too much, methinks: Problems with the Foster "response". *Computers and the Humanities*, 32(6):425–488, 1998.

W. E. Y. Elliott and R. J. Valenza. So many hardballs, so few over the plate. *Computers and the Humanities*, 36(4):455–460, 2002.

J. M. Farringdon, A. Q. Morton, and M. G. Farringdon. *Analysing for Authorship: A Guide to the Cusum Technique*. University of Wales Press, Cardiff, 1996.

R. Forsyth. Towards a text benchmark suite. In *The Joint International Conference for Computing in the Humanities and the Association for Literary and Linguistic Computing*, Ontario, Canada, 1997.

R. Forsyth. Stylochronometry with substrings, or: A poet young and old. *Literary and Linguistic Computing*, 14(4):467–478, 1999.

D. Foster. A Funeral Elegy: W[illiam] S[hakespeare]'s "best-speaking witnesses.". *Publications of the Modern Language Association of America*, 111(5):1080, 1996a.

D. Foster. Primary culprit: An analysis of a novel of politics - who is anonymous? *New York*, 26 February, 1996 1996b.

D. Foster. Response to Elliott and Valenza, "And Then There Were None". *Computers and the Humanities*, 30:247–25, 1996c.

D. Foster. The Claremont Shakespeare authorship clininc: How severe are the problems? *Computers and the Humanities*, 32(6):491–510, 1999.

D. Foster. *Author Unknown: On the Trail of Anonymous*. Henry Holt and Company, New York, NY, 2000.

J. Gains. Electronic mail - a new style of communication or just a new medium?: An investigation into the text features of e-mail. *English for Specific Purposes*, 18(1): 81–101, 1999.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistcal Association*, 70(350):320–328, 1975.

J. Goodman. Extended comment on language trees and zipping, 2002. WWW Document, URL `http://arXiv.org/abs/cond-mat/0202383`, accessed March 12, 2003.

D. Goutsos. Review article: Forensic stylistics. *Forensic Linguistics*, 2(1):99–113, 1995.

T. Grant and K. Baker. Identifying reliable, valid markers of authorship: A response to Chaski. *Forensic Linguistics*, 8(1):66–79, 2001.

R. A. Hardcastle. Forensic linguistics: An assessment of the Cusum method for the determination of authorship. *Journal of the Forensic Science Society*, 33(2):95–106, 1993.

R. A. Hardcastle. Cusum: A credible method for the determination of authorship? *Science and Justice*, 37(2):129–138, 1997.

S. C. Herring. Gender and democracy in computer-mediated communication. *Electronic Journal of Communication*, 3(2), 1993.

J. Higgins. Function words in English, n.d. WWW Document, URL `http://www.marlodge.supanet.com/museum/funcword.html`, accessed January 15, 2001.

M. Hills. *You Are What You Type: Language and Gender Deception on the Internet*. Bachelor of Arts with Honours Thesis, University of Otago, 2000.

D. I. Holmes. The analysis of literary style: A review. *The Journal of the Royal Statistical Society (Series A)*, 148(4):328–341, 1985.

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

D. I. Holmes and R. Forsyth. The *Federalist* revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–127, 1995.

D. I. Holmes, M. Robertson, and R. Paez. Stephen Crane and the *New-York Tribune*: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3):315–331, 2001.

J. Hoorn, S. Frank, W. Kowalczyk, and F. van der Ham. Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, 14(3):311–338, 1999.

T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *International Conference on Machine Learning*, 1997.

T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. Technical Report LS-8 Report 23, University of Dortmund, 19 April, 1998. WWW Document, URL `http://www.cs.cornell.edu/People/tj/publications/joachims_97b.pdf`, accessed October 15, 2000.

T. Joachims. Making large-scale SVM learning practical. In B. Scholkopf, C. J. C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

A. Johnson. Textual kidnapping - a case of plagiarism among three student texts? *Forensic Linguistics*, 4(2):210–225, 1997.

B. Johnstone. Lingual biography and linguistic variation. *Lannguage Sciences*, 21: 313–321, 1999.

D. V. Khmelev and F. J. Tweedie. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16(4):299–307, 2002.

R. I. Kilgour, A. R. Gray, P. J. Sallis, and S. G. MacDonell. A fuzzy logic approach to computer software source code authorship analysis. In *International Conference on Neural Information Processing and Intelligent Information Systems*, Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems, pages 865–868. Springer-Verlag, Singapore, 1997.

B. Kjell. Authorship attribution of text samples using neural networks and Bayesian classifiers. In *IEEE International Conference on Systems, Man and Cybernetics*, San Antonio, TX, 1994a.

B. Kjell. Authorship determination using letter pair frequencies with neural network classifiers. *Literary and Linguistic Computing*, 9(2):119–124, 1994b.

B. Kjell, W. A. Woods, and O. Frieder. Information retrieval using letter tuples with neural network and nearest neighbor classifiers. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1222–1225, Vancouver, BC, 1995.

J. Klein. *Primary Colors: Anonymous*. Warner Books, 1996.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, 1995.

I. Krsul and E. H. Spafford. Authorship analysis: Identifying the author of a program. *Computers and Security*, 16(3):233–57, 1997.

G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. In *International Conference on Machine Learning*, Sydney, Australia, 2002.

R. B. LePage and A. Tabouret-Keller. *Acts of Identity: Creole-based Approaches to Language and Ethnicity*. Cambridge University Press, Cambridge, 1985.

A. Lohrey. Linguistics and the law. *Polemic*, 2(2):74–76, 1991.

D. Lowe and R. Matthews. Shakespeare vs. Fletcher: A stylometric analysis by Radial Basis Functions. *Computers and the Humanities*, 29:449–461, 1995.

P. Lyman and H. R. Varian. How much information? 2000. WWW Document, URL `http://www.sims.berkeley.edu/how-much-info/`, accessed December 5, 2002.

B. A. Masters. Cracking down on e-mail harassment, 1998. WWW Document, URL `http://www.washingtonpost.com/wp-srv/local/frompost/nov98/email01.htm`, accessed March 12, 2003.

R. Matthews and T. Merriam. Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, 8: 203–209, 1993.

G. R. McMenamin. Style markers in authorship studies. *Forensic Linguistics*, 8(2): 93–97, 2001.

T. C. Mendenhall. The characteristic curves of composition. *Science*, 9:237–249, 1887.

T. Merriam. Marlowe‘s hand in Edward III revisited. *Literary and Linguistic Computing*, 11(1):19–22, 1996.

T. Merriam and R. Matthews. Neural compuation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9:1–6, 1994.

G. M. Mohay, B. Collie, A. Anderson, O. de Vel, and R. McKemmish. *Computer and Intrusion Forensics*. Artech House Incorporated, Norwood, MA, USA, 2003.

F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley Publishing Company, Inc., Reading, MA, 1964.

G. Ojemann. Brain organization for language from the perspective of electrical stimulation mapping. *Behavioral and Brain Sciences*, 6:189–230, 1983.

J. Pitkow, C. Kehoe, K. Morton, L. Zou, W. Read, and J. Rossignac. GVU’s 8th WWW user survey, 1997. WWW Document, URL `http://www.gvu.gatech.edu/user_surveys/survey-1997-10/`, accessed April 14, 2002.

Project Gutenberg, n.d. WWW Document, URL: `http://promo.net/pg/`, accessed September 29, 2000.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

P. Rayson, G. Leech, and M. Hodges. Social differentiation in the use of English vocabulary: Some analysis of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1):133–152, 1997.

J. Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and Humanities*, 31:351–365, 1998.

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *AAAI 1998 Workshop on Learning for Text Categorization*, Madison, Wisconsin, 1998.

P. J. Sallis and D. Kassabova. Computer-mediated communication: Experiments with e-mail readability. *Information Sciences*, 123:45–53, 2000.

A.-J. Sanford, J.-P. Aked, L.-M. Moxey, and J. Mullin. A critical examination of assumptions underlying the Cusum technique of forensic linguistics. *Forensic Linguistics*, pages 151–167, 1994.

V. Savicki, D. Lingenfelter, and M. Kelley. Gender language style and group composition in Internet discussion groups. *Journal of Computer Mediated Communication*, 2(3), 1996.

S. Singh. A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing*, 16(3):251–264, 2001.

J. C. Sipior and B. T. Ward. The ethical and legal quandary of email privacy. *Communications of the ACM*, 38(12):48–54, 1995.

J. A. Smith and C. Kelly. Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, 36:411–430, 2002.

M. W. A. Smith. Recent experience and new developments of methods for the determination of authorship. *ALLC Bulletin*, 11:73–82, 1983.

E. H. Spafford and S. A. Weeber. Software forensics: Can we track code to its authors? *Computers and Security*, 12(6):585–95, 1993.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Ser. B*, 36(2):111–147, 1974.

K. Storey. Forensic text analysis. *Law Institute Journal*, 67(2):1176–1178, 1993.

N. M. Sussman and D. H. Tyson. Sex and power: Gender differences in computer-mediated interactions. *Computers in Human Behaviour*, 16:381–394, 2000.

B. Thisted and R. Efron. Did Shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

R. Thomson and T. Murachver. Predicting gender from electronic discourse. *British Journal of Social Psychology*, 40(2):193–208, 2001.

V. Tirvengadum. Linguistic fingerprints and literary fraud. Computing in the Humanities Working Papers, 1998. WWW Document, URL `http://www.chass.utoronto.ca/epc/chwp/tirven/`, accessed August 14, 2000.

R. N. Totty, R. A. Hardcastle, and J. Pearson. Forensic linguistics: The determination of authorship from habits of style. *Journal of the Forensic Science Society*, 27(1): 13–28, 1987.

Y. Tsuboi. *Authorship Identification for Heterogeneous Documents*. Master's thesis, Nara Institute of Science and Technology, 2002.

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

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

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

S. Waugh, A. Adams, and F. J. Tweedie. Computational stylistics using Artificial Neural Networks. *Literary and Linguistic Computing*, 15(2):187–198, 2000.

I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, San Francisco, California, USA, 2000.

Y. Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retireval*, 1(1):67–88, 1999.

A. Yasinsac and Y. Manzano. Policies to enhance computer and network forensics. In *2001 IEEE Workshop on Information Assurance and Security*, pages 289–295, United States Military Academy, West Point, NY, 2001.

G. U. Yule. On sentence-length as a statistical characteristic of style in prose, with applications to two cases of disputed authorship. *Biometrika*, 30:363–390, 1938.

G. U. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, 1944.

G. K. Zipf. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA, 1932.