Application of Psycholinguistic Features to Authorship Profiling for First Language, Gender and Age Group

ROSEMARY A. TORNEY

This thesis is submitted in total fulfilment
of the requirements for the degree of
Doctor of Philosophy

School of Engineering and Information Technology

Federation University
PO Box 663
University Drive, Mount Helen
Ballarat, Victoria 3353
Australia

Submitted in August 2014

Principal supervisor: Associate Professor Peter Vamplew
Associate supervisor: Professor John Yearwood

Abstract

Much of the fraud committed in cyberspace involves the misrepresentation of the demographic data of the perpetrator via the medium of seemly anonymous text messages. One way to address this issue is to apply techniques from the field of authorship characterisation or profiling which is the analysis of text to determine the demographic profile of the author. Most of the previous research into authorship characterisation has used counts and ratios of lexicographically based features that include words, parts of words and Parts Of Speech (POS) contained within the text. This study examines the effectiveness of classifying the first language, gender and age group of an author using a set of features developed in the psycholinguistic field (the Linguistic Inquiry and Word Count - LIWC), both as a single type feature set and in combination with the lexicographically based features used in previous studies (function words, character bigrams and POS unigrams and bigrams). This study also searched for the smallest, most effective subset of each feature set that was practical, by ranking the features using three feature selection algorithms and systematically reducing the number used. In addition, the study explored the effective lower word limit for accurate classification by reducing the text size by regular increments. LIWC was found to be more effective than a similar number of any of the lexicographic feature types, and to add insight rather than noise when combined with these feature types. This held to be true for both the full and reduced text sizes for all three demographic classes examined. In addition it was found that the size of feature sets could be greatly reduced while still maintaining effective levels of classification accuracy.

Statement of authorship

Except where explicit reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or part from a thesis by which I have qualified for or been awarded another degree or diploma. No other person's work has been relied upon or used without due acknowledgement in the main text and bibliography of this thesis

Signed: ______________________________

Dated: ______________________________

Rosemary A. Torney
Candidate

Signed: ______________________________

Dated: ______________________________

Dr Peter Vamplew
Principal Supervisor

Acknowledgments

I would like to express my sincere thanks to my supervisor. Dr Peter Vamplew, School of Engineering and Information Technology, for the enormous amount of time and support he has given me throughout this project. I could not have completed this task without his support and encouragement. I would also like to thank the Internet Commerce Security Laboratory and Federation University for the use of facilities and support during my candidature.

There have been a large number of people who have lent me both practical and moral support during my candidature who I would also like to acknowledge. Firstly, I would like to thank my original associate supervisor, Dr Liping Ma who began this journey with me. Her practical knowledge and experience, as well as her no-nonsense encouragement, were invaluable in staring my path in the right direction. I would also like to thank Professor John Yearwood who took over as my associate supervisor when Liping was unable to continue. He has also provided assistance and advice that has been instrumental in the completion of this work. I have also been a constant distraction to the statistics department and I would like to especially thank Dr Christ Turville, Dr Savin Chand and Mr Peter Martin for their patience in answering my innumerable questions.

While the practical support provided by the people mentioned above has been vital to the completion of my thesis, my journey to PhD could not have been completed without the pastoral care support from my husband, Mr Barry Torney, and my sister, Dr Theresa Hay. I also received much appreciated support and encouragement from my colleagues Ms Kylie Turville and Ms Sally Firmin and from the members of my meditation group, Ms Kath McHenry, Ms Carmel Callas and Ms Joanne Rossi, all who have helped me keep the little sanity that remains to me.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

During the Cold War, the ARPAnet (Lunceford, 2009) was created to enable military communication in the event of a nuclear strike by distributing processing power and information stores across a number of geographically separated locations. The idea being that if one site was disabled, the others could take up the load and remain viable. This was the forerunner of the Internet and the avenues of information gathering and sharing that have been opened up by the Internet and its associated technologies, as it has developed from these humble beginnings, have been said to have brought about social changes akin to the discovery of fire (Berson, 2003).

1.1. Background

However, alongside the many benefits, the introduction of the Internet has led to cybercrime, criminal activity conducted using technology and the internet (Christensson, 2006), which has become one of the scourges of modern day living. As technology removes barriers to communication of information and knowledge, it also removes barriers to exploitation, confidence tricksters and illegal materials. The cost of cybercrime to the community has grown from little more than an inconvenience in the mid-1990s to an estimated US$ 110 billion in 2013 (Hyman, 2013). However, this is only the cost of the theft of money and goods, it does not include hidden costs. These hidden costs range from economic costs including the impact to legitimate e-businesses due to public confidence in the medium being undermined (Lagazio, Sherif, & Cushman, 2014; Shull, 2014), to social costs, for example the emotional damage to victims of romance scams and the destruction of the innocence of children and the victimisation of minors by internet predators (Edwards, 2012).

Criminal activity occurs when there is a convergence of three elements: a motivated offender, a suitable victim and the absence of a capable guardian (Farrell, Phillips, & Pease, 1995). The element of motivation is common to both real world crime and cybercrime. The difference in the offences is in the second and third elements: the identification of a suitable victim and the recognition of the opportunity that is the absence of a suitable guardian.

A suitable victim is one where there is perceived reward for the criminal activity and little or no threat of defence or retaliation (Clarke, 1994). In cyberspace anyone can be considered to be a “suitable victim”. Internet scammers send out thousands of email baits, attempting to trick the unsuspecting user. It was estimated that in 2009 alone, over 5% of the British public had fallen for one type of scam or another, costing the economy over three billion British pounds (Lea, Fischer, & Evans, 2009). The initial contact can have the appearance of authority and create a sense of urgency. Phishing emails claim to be from a trusted source and state that accounts will be terminated if action is not taken immediately, or make an emotional appeal to the users' sense of charity, business kudos or loneliness as in the Nigerian 9-1-4 and romance scams. People fall for these claims, not because they are stupid or greedy, but because their normal vigilance is negated, or because of a flaw in their evaluation strategies and processes for economic decision making (Cialdini & Goldstein, 2004; Lea et al., 2009).

A capable guardian is one that influences the balance of the perceived rewards of the activity compared to the probable repercussions. It takes only a few successful incidents to show that a particular area, time, or genre of crime is profitable and relatively free of consequences for the occurrences of that type of crime to increase very rapidly (Clarke, 1994). These suitable guardians are missing from cyberspace. There are no law enforcement authorities cruising the virtual streets, there are no alert bystanders to report suspicious activities, and there is no physical crime scene for forensic evidence left by the perpetrators to be collected and analysed.

Much of cybercrime, and especially the confidence scams, is perpetrated via fraud. In most cases the perpetrator is geographically removed from the victim; they could be in a different state, a different country, time zone or even a different hemisphere. Fraud is defined as "deliberate deception, trickery, or cheating intended to gain an advantage" (Collins English Dictionary - Complete and Unabridged, 2003). In the real world, fraud is usually perpetrated by the misrepresentation of goods or services. This relies on social engineering and the technique and bluff of the perpetrator to convince the victim the perpetrator is honest and has the victim's best interests at heart. While goods and services can also be misrepresented in cybercrime, it is much more common for the identity of the perpetrator to be misrepresented to gain the trust and confidence of the victim. In the real world, a person can physically avoid dangerous or high crime areas, and will usually know the identity or at least the demographic characteristics of any persons with whom they interact. Demographic information includes age group, gender, nationality, etc. The misrepresentation of age, gender and even nationality, is particularly insidious when used to lower potential victim's suspicions in preparation for criminal activities (Bogdanova, Rosso, & Solorio, 2013). A perpetrator of one gender can represent themselves as the other gender to gain trust and information, or adults can represent themselves as children or adolescents to insinuate themselves into an online peer group to lure victims into trusting complacency, a process referred to as grooming (Berson, 2003; Gardner, 2005). Most of the time this deception is discovered only after the crime has been committed and the damage has been done. However, if a method of detecting this misrepresentation could be found, the deception could be identified before the crime is committed, thus preventing untold costs in emotional, physical and financial damage.

The internet also removes the influence of appearance, body language and tone from the repertoire of the perpetrators as the entire process has to be done, in most cases, via text. Perpetrators rely on the general anonymity of the internet, and specifically that of text to hide their identity from both their victims and law enforcement agencies. However, text may not be as effective at disguising an author identity as it appears on the surface. While it is widely believed that without physical clues, text is anonymous and untraceable, studies have indicated that this is not the case, and that certain parts of speech are largely out of control of the speaker/writer (Newman, Pennebaker, Berry, & Richards, 2003; Newman, Groom, Handelman, & Pennebaker, 2008; Peersman, Daelemans, & Van Vaerenberg, 2011; Pennebaker & Stone, 2003). Authorship analysis is the study of documents and text to ascertain information regarding the characteristics or identity of the author. The current studies in authorship analysis, or computational linguistics, for the most part, rely on lexicographical features including counts of phonemes, words or parts of speech. However, the field of psycholinguistics puts more emphasis on features which have greater discriminatory power, including the categories of words chosen.

1.2. Research Aims and Objectives

To assist in the apprehension of persons who misrepresent demographic data to perpetrate a fraud we need to be able to identify them. This thesis explores the application of psycholinguistic features to the field of computational linguistics and authorship characterisation to further the research into this area. Its aims are to answer the following questions:

1. Are psycholinguistically based features more effective than lexicographical features for authorship characterisation?
2. Is the theoretical basis for psycholinguistic features sufficiently different from that of lexicographical features that the combination of psycholinguistic features with lexicographical features will be significantly more effective in authorship characterisation than equal amounts of lexicographical features alone?

3. Do the number and/or type of features used in an authorship characterisation classification model have an effect on the success and accuracy of that model?
4. Is there an effective lower limit to the number of features that can be used for a classification model for authorship characterisation?
5. Is there an effective lower limit to the number of words in a text that can be classified using authorship characterisation techniques?

In answering the above questions, this thesis contributes to the following areas:

1. Computational Linguistics

   The research establishes if feature subsets from linguistic theory can be useful in the computational linguistic classification task of authorship characterisation or profiling.

2. Short Message Text Categorisation

   The classification and identification of authors of short texts is becoming more and more relevant to the area of text categorisation with the increasing use of shorter forms of electronic communication and computer mediated communication (CMCs), for example: SMS, twitter, emails and chat rooms. The results from the study reported in the thesis will aid in classification of short texts.

3. Authorship Attribution

   The discoveries made during the course of this research will be applicable to all streams of authorship analysis, including the identification of the author of a given text (authorship attribution) and the association of texts by a given author with one another, even if the author's identity is unknown (authorship similarity detection) as well as identifying demographic characteristics of an author from textual clues in their writing (authorship characterisation or profiling).

4. Law Enforcement in Cyberspace

   This study will determine if computer mediated communications have individual characteristics in the text of the email that can be used to aid identification of the authors. The developments made during this study are applicable to areas within internet law enforcement including the detection and prevention of predatory behaviour in emails and chat rooms, bullying in cyberspace, and other illegal activities that use the written word as their medium.

The thesis is organised as follows:

1.3. Thesis Structure

The next chapter, Chapter 2 gives a brief overview of the surprisingly long history of authorship analysis and the progress of this field of research to date. It discusses the various methodologies and algorithms that have been applied to the different streams within the discipline of authorship analysis. The chapter also discusses the theoretical framework for differences in language use in English between authors of differing first language groups, the rationale behind differences in male and female language use and the mechanisms by which the language use of an individual develops and changes as they age. It finally discusses the motivation behind the different types of feature used, and the corpora that were examined in this study, and how this study differs from the previous work on which it is based.

Chapter 3 is the Methodology chapter. It contains an in depth discussion of the two corpora being used in this study, the International Corpus of Learner English (ICLE) (Granger, 2001) and the Blog Authorship Corpus (Schler, Koppel, Argamon, & Pennebaker, 2006). The details of the pre-processing and preliminary feature reduction methods to facilitate the study are discussed in this chapter. It also gives an overview of the methods used in the study for reduction of feature set and document text sizes.

The following three chapters detail the results obtained for the three demographic classes being studied: first language, gender and age group.

Chapter 4 details the results for the classification on the first language of an author writing in English. It presents the results of the first language classification using a multiclass classifier over the sixteen first language groups present in the corpus, using five feature sets in incrementally large features set sizes and various combinations of feature types. Chapter 4 also presents the effects on the accuracy of first language classification when the numbers of features used in the classifier are reduced, when the text size is reduced and when both feature set size and text size are reduced simultaneously.

The results for the first part of the first language classification, using the full sized feature sets and full sized documents have been published in (Torney, Vamplew, & Yearwood, 2012)

Chapter 5 presents the results for classification of text on the gender of the author. The first sections describe the results of increasing the number of features present in a feature set of a single feature type and the effect on the accuracy of the classifier when different types of features are combined in various proportions. In the following sections the feature set sizes are reduced to ascertain the effect on the accuracy of the classifier with surprising results. In the final sections of Chapter 5 the effect on accuracy of reducing text size is examined, followed by the effect of reducing both feature set size and text size concurrently.

Chapter 6 presents the results of a similar series of experiments for the classification of the age group of an author. The corpus was originally divided into three age groups, but research indicated that two age groups, teens and adults would give a more accurate result. As for the gender classification experiments, the five features were tested individually and then in various combinations to examine the impact on the accuracy of the classifier. The accuracy of reduced sized feature sets were then examined, followed by the impact of reducing text sizes, and finally the impact of reducing both feature set size and text size together.

The final chapter in this thesis is Chapter 7, the conclusion. This chapter will reiterate the discoveries and contributions of the study as they relate to the research questions in Chapter 1. It will also discuss any limitations to the study and directions for future work.

Chapter 2 Literature Review

Much of the fraud perpetrated over the Internet is achieved via text especially the type of fraud that involves misrepresentation of demographic data such as age, gender or nationality (Bogdanova et al., 2013). Communication that involves only text is missing the normal clues that help the receiver build an identity of the sender. For this reason, text may be considered anonymous. However there is a field of study that seeks to tease out clues as to the characteristics or identity of an author of a given text from features of that text. This field of study is authorship analysis.

2.1. History of Authorship Analysis

The field of authorship analysis is based on stylometry which is defined as a “linguistic discipline that applies statistical analysis to literary style” (Abbasi & Chen, 2005). Although usually applied to written text, it can also be applied to other areas such as the spoken word, music and fine art paintings (Juola, 2008). One might consider that stylometry or authorship analysis would be a modern discipline. However there is historical evidence that the analysis of language characteristics to profile the author or speaker was practiced in ancient times. One such example is documented in the Bible in the Book of Judges (Judges 12:5). After losing a battle with their neighbouring tribe, the surviving Ephraimite army attempted to take advantage of the post battle confusion and disguise themselves as members of the successful army, the Gileadites, and sneak back across the River Jordan to regroup. An enterprising Gileadite commander exploited a characteristic difference between the languages of the two tribes and requested that all personnel requesting leave to cross the river speak the word “shibboleth” a word that, in ancient Hebrew, refers to the grain bearing part of a head of a cereal crop. The phonemic differences in the two dialects meant that the word spoken by a member of the Ephraimite tribe became “sibboleth”, a distinct and easily identifiable difference to the Gileadite ear (Juola, 2008). While this is an application of stylometry to the spoken rather than written word, a preference for phonemes that are present in a first language can affect an author’s preference for words and terms in a second or subsequent language (Wong & Dras, 2009). As an example of the changing nature of living languages in general and English in particular, the word “Shibboleth” in modern English now refers to any distinguishing practice or characteristic that is endemic to a particular region or culture, possibly because of this biblical account. Shibboleth is also the name of a single sign-on technology for online identities.

In more recent times, one of the earliest documented applications of authorship analysis to text was a study of the works of Shakespeare by Mendenhall in 1887 (Juola, 2008). The study was undertaken due to the suspicion in some circles that all of the works attributed to The Bard were not in fact actually penned by Shakespeare himself, but by some of his contemporaries, or even plagiarised from much earlier works. Mendenhall concluded, using frequency distributions of words of various lengths, that the works of Shakespeare and Marlowe, a writer who died shortly before Shakespeare's first publication, were strikingly similar (Malyutov, 2005). There have been several investigations into the veracity of Shakespeare's body of work, and a number of alternative authors have been presented including Sir Francis Bacon (Cockburn, 1998), Christopher Marlowe (Malyutov, 2005), the Earl of Southampton (Shakespeare's patron), Edward de Vere, the seventeenth Earl of Oxford and a host of others including Queen Elizabeth the first (Reed, 1987). Most of these studies have used statistical methods comparing the lexical features of the documents in question. These include Yule's characteristic K, which measures the probability that two nouns chosen at random from a text will be the same, Honere's R measurement which measures the number of unique words in a text, and Simpson's index which measures the probability that any two words, not just nouns as in Yule's measurement, chosen at random from a text will be the same (Juola, 2008). The object of these measures is to find one or more discriminators, that is individual stylistic features, that are largely invariant over different passages and over time (Argamon-Engelson, Koppel, & Avneri, 1998).

2.2. Methods used in Authorship Analysis

As the field of authorship analysis has matured, and the available computing power has increased, many different methods of analysis have been used. Earlier studies used unitary invariant approaches. A unitary invariant is a single numeric function of a text that is sought to discriminate between authors. Unitary invariant methods did not prove to be consistent across documents and/or genres, and from the early 1960s, multivariate analysis approaches were used (Koppel, Schler, & Argamon, 2009). As the power of computing increased and became more readily available and more cost effective, various approaches have been developed to take advantage of automation. This has led to more sophisticated statistical approaches and machine learning methods that are also achieving respectable results.

Of the statistical methods used in the literature, the most common are Principal Component Analysis (PCA) and Linear Discrimination Analysis (LDA). Stamatatos et al (1999) found that 43% of the differences between authors were in the first two principal components, although the specifics of these components was not specified in their paper. Baayen et al (2002) also found that PCA was very effective. However they found that PCA was more effective when classifying on the educational level of the author rather than the overall style. They found that LDA is more effective for authorship discrimination than PCA, providing that the genre is controlled. Both of these methods were better able to classify on the genre of the text than the authorial style. Naïve Bayes classifiers are a family of simple probabilistic classifiers based on Bayes theorem with strong independence assumptions between the features. That is the features have no relationship with each other and each contribute to the classification independently, however these assumptions are often incorrect, impacting on the accuracy of the classification (Witten & Frank, 2005). Information Gain is a method that determines which feature(s) have the greatest discriminatory power between the classes being classified. It can be used to decide the ordering of the nodes in a decision tree (Witten & Frank, 2005). Argamon et al (2009) had success using a Naive Bayesian classifier and both they and Abbasi and Chen (2008) also successfully applied Information Gain to authorship problems. Van Halteren (1999) created a new machine learning technique called Weighted Probability Distribution Voting (WPDV). During the learning phase this system determines the output class probability distribution for each input feature. During the classification, it takes all the input features and adds the corresponding probability distribution, each multiplied by a weight factor. The primary problem for WPDV is the determination of the correct weights for any specific task. It also creates exceptionally large models with features quickly running into the millions, with memory requirements up to 750Mb. Halteren et al (2005) claimed up to 100% success with the WPDV system for authorship attribution, providing the user accepted “don't know” as a valid result. Their method also attributed a text to a similar group of texts rather than to a specific author.

With the advent of machine learning, many new methods have been applied to the task (Koppel et al., 2009). A number of comparative studies, using a variety of feature sets have compared methods of analysis (Abbasi and Chen, 2005; Koppel et al., 2009; Zheng, Li, Chen, and Huang, 2006). These include both multivariate analysis and machine learning approaches. The machine learning methods used for authorship analysis include neural networks (NNs), Support Vector Machines (SVMs) and decision trees. The SVM divides the data space into classes using hyperplanes in high or infinite dimensions. The best separate is achieved by the hyperplane(s) that have the largest distance from the nearest training data point of any class (Witten & Frank, 2005). In the comparative studies, SVMs have been found to be at least equal to, or more effective than other methods, because they can handle larger, noisier data sets (Abbasi & Chen, 2005; Abbasi & Chen, 2008; de Vel, Anderson, Corney, & Mohay, 2001; Li, Zheng, & Chen, 2006) and others. The particular incarnation of the SVM that has been used in a number of studies is the one supplied by the WEKA tool kit (Witten & Frank, 2005). This specific model of the SVM classifier has been used by (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Argamon, Koppel, Pennebaker, & Schler, 2007; Estival, Gaustad, Pham, Radford, & Hutchinson, 2007; Li et al., 2006) in similar studies to the one detailed in this thesis.

2.3. Streams of Research within Authorship Analysis

The general area of authorship analysis has a number of specialist streams. The two main specialist streams identified by Abbasi and Chen (2005) are those of authorship attribution and authorship characterisation. Other researchers (Baayen et al., 2002; Isard, Brockmann, & Oberlander, 2006; Li et al., 2006) have identified these two and also a third stream, authorship similarity detection. The three streams are described in detail in the following sections.

2.3.1. Authorship Attribution

The field of authorship attribution, as the name suggests, attempts to attribute an anonymous article of text to a known author. The earliest authorship analysis studies were in this area, usually on historical manuscripts, such as the Federalist papers (Mosteller & Wallace, 1963) or literary documents such as the works of Shakespeare (Cockburn, 1998; Malyutov, 2005; Reed, 1987). When undertaking authorship analysis, there is usually a reasonably small pool of potential authors with a number of works known to be created by each author. The stylistic characteristics of each author's writing are analysed, along with the style of the anonymous work and the styles are compared with the closest match being assumed to be that of author of the work (Abbasi & Chen, 2005). More sophisticated approaches may allow the system to report that a document is unlikely to have been authored by any of the candidates. For example, Potha and Stamatatos (2014) concatenated all the sample documents for each given author in the pool of potential authors into one large text. They then created character n-gram profiles of the sample texts, one for each pool author, and compared the character n-gram profile of the anonymous work with the profiles. If a result above a given similarity measure threshold was achieved, the document was considered a match, while below the threshold it was considered not to be a match.

Authorship attribution is based on the theory that every person uses language in a slightly different manner. A given language is learned by copying a proficient speaker or author of that language (Ortega, 2009). An individual’s first language is learned, at least in part, by copying their parents or guardians, and later by copying their peers and conforming to socially accepted norms for their cultural background. Therefore each individual learns a different subset of the language, and becomes familiar and comfortable with the use of the idioms and patterns of that subset. The differences are so subtle that they are not noticed in everyday communications. For example preferences for different synonyms such as the difference between saying a person is a student ‘of’, ‘in’ or ‘at’ an educational institution (van Halteren et al., 2005). Linguists agree that grammar and syntax develop and stabilise at an early age, by eight years at the latest, but that vocabulary continues to develop throughout a lifetime, influenced by occupation, peers, travel, other languages learned, and many other factors (Lippi-Green, 1997). Given that humans are creatures of habit, researchers (de Vel et al., 2001; van Halteren et al., 2005) believe that a pattern of language use that has been effective for an individual in one area would be repeated in others, so that language patterns would cross both genre and media. De Vel et al (2001) have found evidence that syntactic structure is created dynamically and subconsciously during both spoken and written communication and that punctuation is the written equivalent of intonation. Van Halteren et al (2005) explored the possibility of a “human stylome”, an authorial fingerprint, that would consist of stylistic features that could accurately identify the works of a given author. They tested this theory on the Dutch Authorship Benchmark Corpus, a set of essays written by students at the University of Nijmegen. The corpus consists of a series of 72 essays composed by eight Dutch literature students, four in their first year of study, and four in their fourth year. The corpus is controlled for age and native language (all were native Dutch speakers). Van Halteren et al (2005) tested the effectiveness of two different features sets for identifying the authors of the essays: one set of features based entirely on vocabulary and the other based on syntax. Although the vocabulary based feature set was more efficient than the syntactic one, the differences were very slight. This result is in contrast to most theories of first language acquisition that the development of grammar and vocabulary in humans has stabilised by the age of eight. However, Van Halteren et al (2005) concluded that there was indeed an authorial stylome that could be used to match essays to authors, if the author was not trying to copy or mimic another author or style. In their study on the Dutch Authorship Benchmark Corpus, Baayen et al, (2002) discovered that it was possible to identify both authorial style and demographic characteristics of an author. However different methods had more discriminatory power for different characterisations. PCA was more effective for determining the educational level of the authors, while LDA was more effective in identifying authorial style.

The Dutch Authorship Benchmark Corpus has 72 authors. Many authorship attribution studies have considerably fewer authors (Zheng, Li, Chen, & Huang, 2006) however there have been studies that have attempted to use authorship attribution techniques on documents from much larger pools with many thousands of candidate authors (Koppel, Schler, Argamon, & Messeri, 2006; Zukerman & Purcell, 2011). These studies have all found that as the number of authors in the pool of candidate authors increases, the accuracy of the classification methods falls.

The size of the document also impacts on the success of authorship attribution. Earlier work suggested that 1000 words was the minimum size of a document for any reasonable accuracy (Abbasi & Chen, 2008; Chaski, 2001; Corney, 2003) but more recent studies have found that documents of 250 words can also be analysed effectively (RW.ERROR - Unable to find reference:139; Abbasi & Chen, 2008; Layton, Watters, & Dazeley, 2010). Texts as small as twitter feeds and SMS messages have also been used for authorship attribution (Bhargava, Mehndiratta, & Asawa, 2013). Bhargava et al (2013) found that when analysing the authorship of twitter feeds, they could achieve *f*-measures as high as 90%, but the number of authors in the group was quite small (between 15 to 20). When the number of authors was increased to 20 the *F*-measures fell to less than 65%. All the studies mentioned have found that as the size of the documents being examined decreases so does the accuracy rate.

The size of the pool of potential authors also has a profound impact on the accuracy of the exercise. Zheng et al (2006) found a 14% drop in accuracy when number of authors in the pool rose from 5 to 20. Other studies have also found that as the author pool size increases the accuracy of the classification decreases (Luyckx & Daelemans, 2011)

2.3.2. Authorship Similarity Detection

Authorship attribution relies on a limited pool of potential authors with known writing styles with which to compare the writing style of a document of unknown authorship. With the increasing occurrence of online fraud, there is a growing interest in identifying the author of computer mediated communications (CMCs). This form of communication has an effectively unlimited pool of potential authors, including two or more “authors” being the same individual writing under different pseudonyms. There is also rarely, if ever, a definitively “known” document with which to compare anonymous texts (Koppel et al., 2006). In this situation, another stream of authorship analysis, authorship similarity detection is useful. Li et al (2006) define authorship similarity detection as the process whereby anonymous texts are grouped together according to their stylistic patterns. While the identity of the actual author remains unknown with this method, various documents and discussions by the author can be grouped together for further analysis.

In a 1999 study, (Stamatatos et al., 1999) extracted 200 documents, 20 each from ten authors, from a Modern Greek online newspaper to see if they could devise a system that accurately grouped the text by author using only the style of the texts. Although this system was reasonably successful, it was only tested it on articles in the Modern Greek language. Modern Greek is a morphologically rich language, which means that there is little restriction on the order in which the words can be assembled into utterances (Andreou, Karapetsas, & Galantomos, 2008) therefore allowing for more variation in linguistic style between individuals. The technique may not be as efficient on languages with a more limited morphology, such as English, French or Mandarin. Li et al, (2006) found that their research was far more effective grouping English documents into author groups (99% accurate) that those written in Mandarin (93% accurate). Although they did not speculate on the reasons for this, it may be because the acceptable word order in Mandarin is far more rigid than that of English. (Stamatatos et al., 1999) also found that most of the errors in their system were accounted for by documents from one author. They assumed this to be because that author had the smallest amount of text, around 1000 words. The genre of the text could also affect the accuracy of authorship similarity detection. Although (Stamatatos et al., 1999) found that the genre composition of the testing corpus had no effect on the accuracy, (Baayen et al., 2002) found that it had a strong impact. This may again be due to the differences in morphology between the Dutch and Modern Greek languages.

In another authorship similarity study (Abbasi & Chen, 2005) extracted 800 documents from two extremist web sites in the “Dark web” project, 400 each from the White Knights (a branch of the KKK) and the Al-Aqsa Martyrs site. The texts consisted of 20 contributions from each of the 20 most prolific authors on each site. Abbasi and Chen (2005) compared the efficiency of the C4.5 decision tree and an SVM across the two languages (Arabic and English). They found both methods to be effective in both languages, however, while the SVM achieved accuracies of between 88% and 97% for the English data set and between 87% and 94% for the Arabic data set, depending on the number of feature types, the C4.5 tree only achieved an accuracy of 61% to 71% on the Arabic data set. The C4.5 was comparable to the SVM for the English data set with accuracies of between 85% and 90%. The two sites had vastly different styles and contrasting styles. Where the White Knights posts consisted of inflammatory statements and profanity in English, the Al-Aqsa Martyrs posts were based on quotes and supposed evidence to back up their ideals in Arabic. This difference in style may be the reason that analysis on Arabic was less effective than the analysis on English.

Argamon et al (2007) used a corpus that had far greater post authorship editing, but with fewer potential authors, for an authorship similarity detection study. They used the chapters of 20 19th century books from 8 different authors from both USA and UK. Their study found that they could cluster the chapters in correct book and author groups with over 90% accuracy for most of the feature sets they used. They also found that certain features also allowed them to cluster the chapters in to nationality groups with greater than 95% accuracy. In a separate study, Argamon and Levitan (2005) found that it was possible to group the same corpus into nationalities with an 84% accuracy using only function words in the feature set.

Although both Abbasi and Chen (2005) and Argamon et al (2007) experimented with more potential authors than is usual for authorship attribution studies, they still had a reasonably limited author pool. When attempting to identify an author of a document, or cluster documents according to their author in cyberspace, there is potentially a pool of thousands or more. This would add orders of magnitude to the task. Koppel et al (2006) conducted a study using over 18,000 blogs from more than 10,000 authors. The texts were the complete set of postings for each author, for whom the age and gender were self-reported. The blogs were divided into snippets of approximately 500 words. They then attempted to match the snippet to the correct author using 3 different feature sets. While the result was very much greater than chance, (Koppel et al., 2006) agree that their success rate of 20% for matching the correct snippet/author pairs in all three feature sets, and 42% in at least one feature set is not sufficiently accurate for commercial applications. For the second part of their study, they trained an SVM on the pairs (snippet of text and author) that were labelled successful/unsuccessful. The system could have one of three outcomes for a given author/snippet pair: classify it correctly, classify it incorrectly, or state it “didn’t know”. The system attempted to classify the snippet in 31% of cases, and gave the correct answer in 88% of those cases. Koppel et al (2006) conclude that as long as the response “don’t know” is regarded as a valid response, it is possible to achieve reasonably accurate results in authorship analysis. One of the other issues with these results however is the selection of features. The three feature sets used were one consisting of content words, one excluding content words and one based on stylistic features including function words. The difference between function words and content words will be examined in Section 2.5.1. Briefly content words are the nouns and sometimes other parts of speech, that are specific to a topic, such as ‘horse’, ‘cat’ or‘ football’ and ‘gallop’, purr’ or ‘hand-ball’. Function words are the parts of speech that join them all together to make a coherent sentence. Function words include prepositions, articles and pronouns. The second part of the experiment combined the three feature sets. The function words were most successful when used in isolation, but the more successful second part used the combination of features including content words. Using content words can identify authors that are writing about a particular subject. However if the writing is controlled for subject matter and genre, these would not be as successful.

2.3.3. Authorship Characterisation or Profiling

In its classic incarnation, “profiling” is an attempt to extrapolate characteristics of a person from the clues they have left at a crime scene (Alison, Smith, & Morgan, 2003). The first recorded police criminal profile was in 1888 and was given as a description of the criminal that would later come to be known as “Jack the Ripper” (Canter, 2004). Some of the more famous fictional criminal profilers are Sir Arthur Conan Doyle’s Sherlock Holmes and Agatha Christie’s Inspector Hercule Poirot. Both of these characters inspect the scene of a crime and ascertain the characteristics, and eventual identity, of the perpetrator from clues left at the scene. Sherlock Holmes deduces the physical characteristics of the culprit while Inspector Poirot infers their psychological character (Douglas, Ressler, Burgess, & Hartman, 1986; Van Horne, 2013). These two sleuths easily read the evidence and point to the perpetrator without a hitch. However in reality, criminal profiling is not such an exact science. Some criminal psychologists believe that profiling is “as useful to law enforcement as a tea leaf reading” due to either the descriptions being so broad as to include a large percentage of the population or being about nontangible characteristics such as thought process rather than observable behaviours (Alison et al., 2003).

Linguistic profiling is notably more effective because, as well as individuals each having their own unique take on their language (de Vel et al., 2001; van Halteren et al., 2005), the use of language indicates a personal view of the world (Boroditsky, 2001), their obsessions and preoccupations (Mehl & Pennebaker, 2003), demographic characteristics (Gollub et al., 2013) and even mental state or illness (Argamon et al., 2005; Rude, Gortner, & Pennebaker, 2004). The circumstances in which language is used can also affect its use by an individual. Research is revealing that it is possible to discover an insight into an individual’s thoughts, emotions and motives by analysing their language use in speech and writing (Newman et al., 2008). Studies have even shown that people speak differently when attempting to deceive an audience and that these departures from their normal communication style can be observed and measured (Newman et al., 2003). For example in an FBI investigation into a kidnapping case, the supposed “victim” was identified as a party to the crime by the language she used to describe her ordeal. In an interview with investigators, she made statements such as “we went to the forest” and “we stayed in the old cabin” rather than “he took me to the forest” and “I was kept in the old cabin”. The use of the first person plural pronoun indicated that she was complicit in the abduction, whereas the use of singular pronouns would have shown her mental distance from the perpetrator (Adams, 1996).

2.4. Demographic Profiles

Authorship characterisation can indicate the psychological or the demographic profile of an author (Gollub et al., 2013; Koppel, Schler, & Zigdon, 2005). The term demographic is used to describe the quantifiable statistics of any given population. A demographic profile can include things such as education, marital status, income, age or gender and can encompass cultural items such as religion or first language. The demographic information studied in this thesis is age group, gender and first language of authors communicating in English. These three demographic areas have been chosen because they are often the ones misrepresented in online fraud and criminal activity. For example; when an adult presents as a teenager or a male presents as a female to gain the trust of potential fraud victims (Berson, 2003; Bogdanova et al., 2013). Members of organisations that perpetrate romance scams also present as speakers of a language that is not their first language or that of their country of origin to lure potential victims (Rege, 2009; Trevathan & Myers, 2012).

2.4.1. First Language

All languages have their own unique phonology, morphology, syntax and semantics. Studies have shown that the phonology (Tsur & Rappoport, 2007), morphology and syntax (Andreou et al., 2008; Argamon et al., 2009) of first languages impact on expressions and word choices in second and subsequent languages.

Phonology is the study of sounds that words contain. The human vocal apparatus has an incredibly extensive, although finite, set of sounds that it can produce. However each language contains only a subset of these sounds. Most children are born with the ability to produce all the sounds possible for the human vocal apparatus, but eventually lose the ability to produce the ones they do not hear around them, and sometime during adolescence the ability to learn a new set of phonemes, ie, a foreign language as easily as a young child is lost (Ortega, 2009). Children that are exposed to more than one language at birth retain more than one set of sounds or phonemes (Lippi-Green, 1997; Ortega, 2009). This affects the use of a subsequently learned language (Abu-Jbara, Jha, Morley, & Radev, 2013).

Phonemes are most often represented by character bigrams in English text, and an examination of character bigrams used in the written communication of English language students in five non-English speaking countries identified their first language in over 60% percent of cases (Tsur & Rappoport, 2007). This is believed to be the result of the authors first language phonemes influencing their choice of English words, even in written text – an author is more likely to be familiar with, and use, words that they can easily pronounce rather than words that use phonemes not found in their first language (Ortega, 2009). Most linguists consider it "not possible for an adult to substitute his or her phonology (one accent) for another in a consistent and permanent way" (Lippi-Green, 1997)

Morphology relates to how associated words are formed, such as plurals of nouns and conjugations of verbs to match person, tense and number (Finegan, Blair, & Collins, 2000). For example, in English the conjugation of the verb "to be" (an irregular verb) is "I am", "you are", "he is", "they are", while the conjugation of the verb "to play" (a regular verb) is "I play", "you play", ``he plays", "they play". Different languages have different levels of morphology. Modern Greek, for example is considered to be morphologically complex to the extent that is not necessary to have a pronoun subject of a sentence because the conjugation of the verb can give that information (Andreou et al., 2008), whereas in Mandarin, there is no conjugation of the verb at all. Similarly in many languages the plural form of a noun appears differently that the single form of the noun. In English the plural of regular nouns are created by adding an "s" to the end of the word, while some words such as "mouse" or "goose" appear in a completely different form. In contrast to Mandarin where there is again no change to the word to indicate whether it is singular or plural, unless the word is a pronoun. Native English speakers find the complex morphology confusing in Modern Greek because English is not as morphologically complex as Modern Greek (Andreou et al., 2008). French is another language with a high level of morphology - it is possible to spell the conjugations of some verbs more than 16 different ways depending on the tense and person. To continue the example given above the French verb "etre" which means "to be" (also irregular in French) is conjugated "je suis" (I am), "tu es" (you (singular/informal) are), "il/elle est" (he or she is), "nous sommes" (we are), "vous etes" (you (plural/formal) are), "ils/elles sont" (they (male or female) are) for the present tense. The spelling (and pronunciation) is again different for each person for the other tenses. "Etre" is considered an irregular verb, in that each person/tense combination has an almost completely different spelling, however even regular verbs in French have a different suffix depending on the person and tense. In comparison, the most irregular verbs in English, such as "to be" or "to have" have only three different spellings in the present tense. Mandarin, has an even lower morphological complexity than English, in that there are very few plurals and the verbs are not conjugated to match person or tense. When an author from a non-English speaking background writes in English their spelling and grammar can be influenced by their first language (Vajjala & Loo, 2013). Examples include function words, which can be confusing to non-native English users, and letter n-grams that could indicate spelling conventions in first language, where the first language uses the same alphabet as English (Koppel et al., 2005).

Syntax is the order in which words can be validly used to form a sentence. In Modern Greek the word order is very free when compared to other languages such as English, French or Mandarin. Native English speakers do not have difficulty learning the much more syntactically complex Modern Greek, because the syntax of English is an acceptable and legal word order. However Modern Greek speakers find it difficult to express themselves in English because syntactic patterns that are legal in Modern Greek and used for particular stresses and expressions are not legal in English (Andreou et al., 2008). Mandarin has far stricter syntactic rules, in that there is often only one word order that is legal (and understandable to a native speaker). Translating word for word from English to other languages and vice versa, does not always produce a sensible sounding sentence. For example the direct translation for the sentence in Mandarin that means “David and I are going fishing tomorrow” is “I tomorrow and David go fishing”. Author attribution studies using corpora in English have a higher accuracy than the same studies conducted on Mandarin corpora, possibly because the more relaxed syntactic rules of English allow more freedom in communication to create unique expressions (Li et al., 2006; Zheng et al., 2006). (Stamatatos et al., 1999) achieved a very high level of accuracy using a corpus consisting of Modern Greek texts. Modern Greek has very free syntactic rules and almost any word order is acceptable and legal (Andreou et al., 2008) which could result in a very large selection of expressions, allowing authors access to more unique methods to state the same meaning. Syntax can be identified in a limited way by Part-Of-Speech (POS) bigrams. Rare or unusual POS bigrams could indicate an author not familiar with the norms of English syntax. Errors such as incorrect tense or number agreement with verbs and nouns can also be identified by POS bigrams and indicate a non-native speaker communicating in English (Koppel et al., 2005). Other syntactic evidence of a non-native speaker could be over or under use of words that differ from a native English speakers use. For example there are no articles in Mandarin or Russian, so native speakers from these languages tend to omit them when communicating in English. However to indicate a general concept (ie a horse instead of the horse) Russian speakers use the word “some” more often than a native English speaker would.

Semantics is the study of meaning. The word itself is from an Ancient Greek word meaning “significant’. Semantic errors in second language speakers can relate to word that are semantically unacceptable but grammatically correct. For example; using the word “alone” instead of “lonely”, or using the word “uneasy” to mean “difficult” (ie using the prefix negation ‘un’ to indicate not easy) instead of “anxious”. Yang et al (2013) found that this type of semantic error was common for native Mandarin speakers when using English. There are also common sayings or phrases that, when translated are slightly incorrect, such as the French sayings “your beans are cooked” for the English “your goose is cooked” to mean a person is in trouble, or the French “he has holes in his hands” to the English “he has holes in his pockets” to mean an unthrifty person. The more common semantic errors that a non-native speaker makes can include the confusion of opposites (mistaking the meanings of “young” and “old”), gender errors (mistaking the meanings of ‘him/he/his” with “her/she/hers”) and incorrect synonym/context use (mistaking the use of the word “strong” and “harsh”) or other errors in appropriateness or accuracy (King & Dickinson, 2013).

The effect of a prior language’s phonology, morphology, syntax and semantics on a second or later language may clearly indicate a speaker or author is not a native speaker and the original language might be pinpointed using the differences in these four linguistic dimensions between a first language and English where the first language spoken is not English.

2.4.2. Age Groups

Identifying the first language of a non-native speaker is the detection of linguistic characteristics of the speaker or author’s first language as they impact on their production of the second language. However, when looking to identify other demographic characteristics, such as age group, it could be assumed that the cohort will all have similar linguistic backgrounds and there will be no obvious differences in phonemes, morphology, syntax or semantics. Is it possible to discern differences in language use between different age groups? Linguists and psychologists believe it is. Language is a living, changing entity, dealing with concepts strongly related to experience and therefore responding to changes in technology and custom – meaning is being built up constantly by use in different situations (Paradis, 2011). Human beings are creatures of habit, and once a pattern of behaviour or speech is established, unless it becomes manifestly unsuccessful, it will be repeated. Use of language is no exception (de Vel et al., 2001). Therefore, an individual will continue to use the style of language that they grew up with, even though the language style of younger age groups within their culture has developed and changed. Three of the mechanisms by which the semantics of a language can change are metonymization, facetization and zone activation (Geeraertz & Piersman, 2011; Paradis, 2011; Rohrdantz et al., 2011).

Metonymization occurs when one word or phrase can stand in for another . It is used in both spoken and written language to add interest by avoiding the use of the same phrase over and over again, or by shortening a sentence by using one word in place of several ("meal" instead of listing every item consumed, for example). One familiar example of metonymization is the phrase "the pen is mightier than the sword" from the 1839 play Richelieu, by Edward Bulwer-Lytton. It does not mean, of course, that a pen would be preferable to a sword in any combat situation; it means the written word ("the pen") is more persuasive than aggression or military might ("the sword"). Another example could be the statement "The red shirts played well and won the game." No-one would take this to mean that a group of clothing items preformed some activity so convincingly that they were victorious in a competition. The "red shirts" refers to the uniform of the people in the team and they are then known by this distinguishing feature. While the different conceptual meanings of metonymized phrases can stand in for one another in conversation, the different meanings cannot be combined. The sentence "The red shirts won the match and had to be cleaned thoroughly" does not make semantic sense because the two qualia (instances of meaning) create a zeugma or semantic antagonism, when combined (Geeraertz & Piersman, 2011).

Facetization refers to different meanings or senses of the same word. For example the statement "the book is long and boring" refer to two different facets of the word "book". Saying the book is long refers to the physical entity – the amount of text or information it contains, while saying it is boring refers to the content. In the case of facetization, the two qualia do not create a zeugma when used together. Facets are assumed to be aspects of the same concept or sense where metonymizations are separated by boundaries in conceptual space (Paradis, 2004).

Zonal activation is similar to facetization except that instead of a different facet being used to refer to the whole, a different zone of the concept is used, either a part of the concept or a larger concept that encompasses the idea. The examples given by (Geeraertz & Piersman, 2011) include the statement that someone has a cigarette in their mouth, when in fact they only have the filter of the cigarette in their mouth, and statements such as "Washington is insensitive to the people" meaning that the members of the government in Washington are insensitive rather than the city itself.

While (Paradis, 2011) asserts that all changes to language are the result of metonymization, where the replacement word becomes so synonymous with the meaning of the replaced phrase, that the word gains another facet of meaning; However Geeraertz and Piersman (2011) suggest that these three methods all impact on movement in semantic meaning in language. An example of metonymization changing the popular meaning of a word can be seen in the word “gay”. This word actually always had two meanings: “happy and carefree” and “hedonistic and careless”. It is possible that the first meaning was the result of a popular change in the meaning of “carefree” which originally meant almost the same as “careless”. The first meaning was the commonly used one until the second meaning “hedonistic” that had been applied to the homosexual community was adopted by them as their by-word. Now, some social groups are using the word “gay” to mean something that is weak, or wrong, basing this on certain religious view of the homosexual community (Mihalcea & Nastase, 2012). An example of a zonal activation changing the meaning of a word can be seen in the word “bug”. The accepted meaning of the word was a collective term for an insect, however it now also means a fault or problem in a computer. This meaning has come from an incident in the early development of computers when a prototype ceased to function. Later examination found that it had a bug – literally, there was a dead insect blocking a current in the hardware. Zonal activation has meant that the specific problem of a “bug” has become synonymous with any computer problem, either software or hardware (Raymond, 2003). Technology in general has changed the way many words are used and perceived. As recently as the 1980s a “mouse” was a rodent, a “keyboard” was part of a musical instrument and to “surf” was to play on the waves. Each generation speaks about the world with the terms and technologies that were current in their formative years. This leads to a divide between the language use between generations and age groups (Rosenthal & McKeown, 2011). The three change mechanisms mentioned above, as well as changes in social preferences and technical advances combine to give each generation its own unique linguistic style (Kucukyilmaz, Cambazoglu, Aykanat, & Can, 2008; Pennebaker, Mehl, & Niederhoffer, 2003).

As well as language growing and developing over time, the way that individuals use language also changes as they age. A report that analysed the writings of over 3000 participants in 45 separate longitudinal studies (Pennebaker & Stone, 2003) revealed that as individuals aged they used more positive and less negative words and that the older authors also used more present and future tense verbs than past tense. Other trends were a reduction in the use of time related words such as day, time, clock, etc and less reference to self and others within text along with greater cognitive complexity and use of longer words. Even when discussing negative events, older authors tended to negate positive words rather than use negative – ie “I’m not happy” rather than “I’m sad” (Pennebaker & Stone, 2003). A separate study on blog uses also found that older authors use more positive than negative words, and that there was less use of pronouns and greater use of articles (Schler et al., 2006). Schler et al (2006) did in fact use some of the categories of LIWC, however they limited them to nine fairly specific and concrete categories. A number of the categories that were included in LIWC in 2006 appear to no longer be a part of the system such as “sports” and “tv”, although they could conceivably be included in the overall recreation category. Schler et al (2006) also combined these categories with hyperlinks and very specific topic words such as “awesome”, “homework”, “drunk”, “marriage”, “campaign” and “tax. The first three would certainly be more likely to be young people and the second three more likely to be older people. However, if a person wished to present as an older or younger person in the blog or chat, they would be easily able to monitor their communication to avoid using more age specific terms.

The combined effects of changes in semantic meanings over time and individual’s choice of expression over time imply that there is a quantifiable change in both ambient language use and individual’s actual language over time. There have been a number of studies into the effect of age on authorial style. (Koppel, Schler et al. 2006, Schler, Koppel et al. 2006, Argamon, Koppel et al. 2009) have all used the blog corpus compiled by (Koppel, Schler et al. 2006) and tested various feature sets to identify the differences in the writing style of authors of one of three age groups – teens, twenties and thirty plus, finding that there are quantifiable differences in both the tone and the subject of the texts. (Nguyen, Phung et al. 2011) studied 10,000 blogs hand-picked from the Livejournal blogging site. The bloggers were divided into two age groups, 5,000 ‘old’ bloggers consisting of individuals aged between 39 and 60, and 5,000 ‘young’ bloggers, aged between 20 and 22. They found that the two groups displayed significant differences in mood, topic and expression. The particular blogging site Nguyen et al (2011) used allows the blogger to indicate their current mood by attaching one of a 132 predefined mood tags. The tags include both positive (happy, cheerful, grateful etc) and negative (discontent, sad, uncomfortable, etc) mood tags. (Prasath 2010) divided yet another blog corpus into four age groups, teens, twenties, thirties and over forties, and found that the teens age group used more slang and non-dictionary words, and shorter sentences than the other age groups. (Rosenthal, McKeown 2011) compared the writings of individuals that were in college before and after the advent of social media and computer mediated communications (CMCs). They used seventeen different features that fell into three categories: online behaviour (ie number of friends, time of posts, etc), lexical style (including number of emoticons, acronyms, non-dictionary words and standalone punctuation) and lexical content (including collocations in age group, syntactic collocations and POS collocations), and found that they were able to reliably distinguish between the two groups. Many of the studies into the effects of age on language have included content words in their feature sets, using the different topics discussed by different age groups as a method to distinguish their text (Meina et al., 2013; Nguyen, Gravel, Trieschnigg, & Meder, 2013; R. Schwartz, Tsur, Rappoport, & Koppel, 2013), and while this is effective, it will not identify an individual who is attempting to mask their age for fraudulent or criminal purposes (Bogdanova et al., 2013). While Schler et al (2006) did use LIWC in their studies into age and gender, they used only 11 of the 80 features available: Money, Job, Sport, TV, Sleep, Eat, Sex, Family, Friend, positive and negative emotions. In the new version of LIWC used in this work, Sport and TV have been included in the Leisure section, and Sleep has been removed. This work also used the entire number of LIWC features available.

2.4.3. Gender

Do males and females use language differently? The most significant hormonal difference between males and females, testosterone, has been linked to aggression, negative moods, improved special skills, decreased verbal ability, concerns over status and dominance, and more direct thought and action (Pennebaker, Groom, Loew, Dabbs, & Abbasi, 2004). When pairs of males and females performances on a social dilemma type task have been compared, female pairs are seen to develop trust faster, show more concern for social process and to get maximum benefit for all, where males show a certain ruthlessness to win at all costs which hinders the development of trust (Sun, 2008). The study did not allow the pairs to meet face to face, but did allow free chat time away from the task. Female pairs used this time to reinforce their partnership with encouraging, supportive language, while the male pairs used the time to speak of sporting events and actions that could impact on the task. In writing women tend to use more words related to psychological and social processes, while men use more words related to object properties and impersonal topics. Women's language is described as involved and used as a social process, while men's language is categorised as informative and used to convey information (Newman et al., 2008). Female communication also tends to have more softening statements when compared to male communication: for example the difference between a woman asking "would anyone like some food" and a male stating "let's get some food". The meaning is essentially the same, but the tone is different. Female authors also use more uncertainties than male authors, prefixing statement with phrases such as "I wonder if…" and "Does anyone know if…". Male text also uses more articles where female uses more pronouns. Many of the studies that have shown these trends have been conducted on university students, making the age range for the participants reasonably homogenous (Newman et al., 2008).

If male and female authors do use language differently, is it nature, and therefore involuntary, or is it nurture and subject to situational factors? Is the difference due to topic choice, where females using language as a social process and speaking about social interactions would by necessity use more pronouns and less articles than males choosing use language to convey information? Or is the difference due to social conditioning? In many cultures women are encouraged to show emotions such as happiness and tenderness and softer negative emotions such as sadness but it is acceptable for men, but not women to show stronger emotions such as anger. An examination of the journal entries by two patients receiving regular testosterone treatments gave an insight into these questions (Pennebaker et al., 2004). The patients were a biological woman undergoing transgender treatment and a biological male being treated for upper body weakness. Both patients were treated with the testosterone injection every three weeks, and both patients were prolific journal writers. The journals covered a period of two years. In both authors, immediately following the treatment, the writing displayed more "male" characteristics – ie informative, discussing events, with a greater use of articles, but as the three weekly cycle progressed, more "female" characteristics were evident – ie greater use of social process and more pronouns (Pennebaker et al., 2004).

Several studies attempting to identify the gender of authors of short communications such as tweets have been undertaken with varying results. However, many of these studies use the first name or handle of the tweeter as well as content words, identifying the different topic choices of males and females (Bamman, Eisenstein, & Schnoebelen, 2014; Fink, Kopecky, & Morawski, 2012; Ludu, 2014; Ugheoke, 2014). As with the age group characterisation studies, the use of names and topics as features would be effective unless the author were attempting to disguise their gender or the situation was one that was controlled for topic. Schler (2006) did use a limited number of the LIWC features in their work, but some of those features have been combined into more general features in the newer version of LIWC used in this work, and this work included all the features available in the package.

2.5. Feature Selection

Feature selection is an important step that can influence the success of the exercise being undertaken as well as the generalisation of the results. Carefully hand picking a set of features that will discriminate between the classes being examined may give excellent results for that corpus, but may not be applicable or useful in other corpora or situations. Another important consideration is how and why the features are pertinent. Understanding of the principals involved will enable feature sets to be honed to greater efficiency for classification in a broad spectrum of situations, not just the corpus under investigation (Kestemont, 2014).

Each vector in the data set is characterised by its values for a given set of features or attributes. The aim of data mining is to extract information from these features and present it in a coherent and human readable way to facilitate decision making or other activities. Not all the features of a data set are useful for a given exercise, and to include all these features will introduce computational complexity, additional expense in terms of CPU time and memory use and noise which can mask the information being sought. To avoid these negative consequences, the most relevant features are selected for use in the data mining exercise (Witten & Frank, 2005). For the purposes of authorship attribution, the data set is the corpora being examined and the features are various words, parts of speech and other grammatical and linguistic structures. The main types of features used in authorship analysis research to date are lexical, syntactic, structural, content and, idiosyncratic based feature sets (Argamon et al., 2009).

Lexicographical features are the measurements and counts of various constructs within the text. These include records of the length of the words within the text, the number of short or long words used, the number of words ending or beginning with a vowel, or counts of various parts of speech such as verbs, nouns or function words (Juola, 2008; Koppel et al., 2009). Most of the early studies of authorship attribution used lexicographical features. Measures used included an index to measure the probability that two words chosen at random from a text would be the same; comparison of sentence length, and the average number of syllables used. These and other measures were designed to quantify the richness of the vocabulary used in the text and attribute it to authors with similar styles in vocabulary (Juola, 2008).

The word “syntax” is from the Greek word “sу́ntaxis” meaning “to set out together or arrange”. In linguistic terms, the study of syntax is the study of sentence structure as well as the category of a word (ie noun, verb, etc) and how to use it in a sentence (Finegan et al., 2000; Jurafsky & Martin, 2000). It follows that syntactic features would show the part of speech of a word and rewrite rules rather than measurements of how many of a given type of word is present. Rewrite rules (also called phrase structure rules) are a means of breaking sentences and phrases down into their constituent sections. For example S -> NP

VP indicates that a sentence (S) can be broken down into a noun phrase (NP) and a verb phrase (VP). Further rules indicate into what components a NP or VP can be further broken down (Finegan et al., 2000). Van Halteren, Tweedie and Baayen (1996) compared the efficiency of function words and rewrite rules for authorship attribution, however their experiments used only two authors both from the same genre, that of crime fiction. Ten snippets of 2500 words were taken from each text. They compared the efficiency of the 50 most common function words with the 50 most common rewrite rules and found that rewrite rules were more effective.

To identify the syntactic features of a text, the parts of speech used must first be labelled using a Part of Speech (POS) tagger. Many different POS taggers have been used in authorship attribution studies. Van Halteren, Tweedie and Baayen (1996) used TOSCA however many others have also been used in authorship analysis including Arizona, (van Halteren et al., 2005), Brill (Argamon et al., 2007) and QTag (Torney et al., 2012).

Structural features correlate to the setting out and appearance of the text, including the number and size of paragraphs, the space between them and indentation and style of tables or graphs used. Structural features can expand to include fonts and colours of text, hyperlinks and graphics embedded in the text and emoticons in Computer Mediated Communication (CMCs) including emails, chats and blogs (Abbasi & Chen, 2005; Ma, Torney, Watters, & Brown, 2009). While structural features can be informative when the text is in the raw state in which the author wrote it, if the document has been formally produced or edited, any information from structural features could be corrupted by the influence of the typesetting and editing processes.

Content features are typically used by internet search engines to identify relevant documents and pages from the vast selection available in cyberspace (Koppel et al., 2009). They include the names, nouns, and verbs that pertain to a particular subject or topic. While content features are very useful when attempting to identify the subject matter of a text, they are not particularly applicable to authorship analysis. In fact studies have found that features that are effective for content identification are orthogonal to features that are effective for authorship analysis. The more useful a feature is for finding a topic, the less useful it is for identifying author characteristics (Koppel et al., 2006). Some studies have used content features to identify the text of adolescents when compared to adult authors, using the different life events discussed to determine the age group of the author (Argamon et al., 2007). However if the conversation is controlled for topic, with an adolescent and adult conversing in a chat room for example, then content features lose their effectiveness.

The fifth feature type, idiosyncratic, was identified by (Li et al., 2006). Patterns of spelling and grammatical errors could be specific to a particular author and be used to identify their text. The effectiveness of idiosyncratic features such as errors specific to given authors has also been examined in the context of authorship attribution and specifically for security of mobile devices (Saevanee, Clarke, & Furnell, 2011). However unless the text is hand written or has been produced using a fairly primitive text editor or mobile device with the auto-correct disabled, the availability of spelling and grammar checking software embedded in many editing packages would give an author ample opportunity to eliminate these features from their text. On a corpus where errors were not corrected by software, Dahlmeier, Ng and Wu (2013) found that the error rate was less than 4 errors for every 100 word tokens. Given the scarcity of this type of feature, encoding the error types could be prohibitively expensive in terms of time and computing resources.

The main feature sets that have been used in previous studies of authorship analysis are described in detail in the rest of this section.

2.5.1. Function words

Function words are the glue that holds English sentences together. They have little lexical meaning, but instead serve to express grammatical relationships between other words, or specify the mood or attitude of the author. They include articles, pronouns, conjunctions, auxiliary verbs, particles and proto-sentences. Function words are so ubiquitous they are often overlooked by authors and proof readers alike. However these (usually) small and unassuming words provide over half the words that are used in English, even though they only make up 0.04% of the overall vocabulary. Function words are more prevalent in languages that do not use case endings or other forms of inflection to mark grammatical function. For example the phrase “with a sword” in English becomes one word in Latin (“ensi”), a language that makes a great deal of use of case endings and inflections (Kestemont, 2014).

Function words have been used for classification of author style since the modern inception of study of authorship analysis (Juola, 2008). Function words are processed in a different way, and in a different part of the brain, than content words making them largely outside of conscious control of the speaker or author (Newman, Pennebaker, Berry, and Richards, 2003). Newman et al (2003) found that the pattern of use of function words changed in authors or speakers who were trying to deceive. Studies have also found a phenomenon known as language transfer that affects the use of function words in a second language (Tsur and Rappoport, 2007) (Wong and Dras, 2009). Language transfer occurs when grammatical structures or common words from an author or speaker's first language influence their word use in their second or subsequent languages.

In this thesis, the term "function words" has been extended to include many other common, but not topic specific words in English. There are a number of reasons for this. Firstly there is no commonly accepted and freely available and exhaustive list of English function words. Secondly, not all non-topic specific words that are used fit within the strict linguistic definition of "function words". Zheng et al (2006) supplied a list of 150 words that were in their English "function word" feature set, but the list was both not exhaustive and contained high frequency English words that not topic specific but are also not function words. Examples of these words include: "following", "anybody", "anything", "somebody", "something", etc. Therefore the decision was made to follow this lead and use the high frequency, non-topic specific words present in the test sets and use the overarching label "function words" in a similar manner to (Zheng et al., 2006)

2.5.2. Character Bigrams

Character unigrams and bigrams have been successfully used to identify the native language of an author in previous studies (Tsur, Rappoport 2007, Wong, Dras 2009). Character bigrams in particular represent phonemes, which can indicate language transfer from a previously learned language (Tsur, Rappoport 2007). Every language contains a subset of the total set of phonemes that the human vocal apparatus is capable of articulating. Babies are born with the ability to make all these sounds, but the ability to produce the ones not found in the language(s) that they are exposed to is lost, usually before adolescence (Lippi-Green, 1997). Character trigrams and 4-grams have also been used in authorship analysis, however, character n-grams with n > 3 tend to indicate words more than phonemes (Torney, Vamplew et al. 2012) and therefore can indicate topic rather than author information. Tsur and Rappoport (2007) found that by using character bigrams, they could identify the native language of an author 66% of the time over the two sets of five first language groups that they studied. Tsur and Rappoport (2007) also found that counts of character unigrams (ie single characters) can also distinguish between first languages in non-native English authors at a rate considerably higher than chance.

Character bigrams can also identify first and last characters of words used and capture punctuation use (Baayen, van Halteren, Neijt, and Tweedie, 2002). Punctuation tokens are the written equivalent of intonation and as such are useful in capturing the style of a text (Baayen, van Halteren et al. 2002). They have proved to be useful in both authorship attribution and authorship characterization or profiling (Abbasi, Chen 2008, Baayen, van Halteren et al. 2002, Torney, Vamplew et al. 2012, Tsur, Rappoport 2007).

Although character bigrams have proved to be useful in both authorship attribution (Abbasi and Chen, 2008; Baayen et al., 2002; Luyckx & Daelemans, 2011) and in profiling the first language of authors (Tsur, Rappoport 2007), there has been mixed results in their use for profiling other demographic characteristics with (R. Schwartz et al., 2013) finding that character bigrams were not as effective as other features when used in single type feature sets. However Potha and Stamatatos (2014) found that character n-gram profiling was useful in author attribution studies on a Modern Greek corpus.

2.5.3. Part of Speech (POS) N-Grams

POS n-grams utilise the syntactic structure of the text. The activity of POS tagging is the process of going through a document and assigning each word to its POS category. While this activity could, and before computing power and software progressed sufficiently, was undertaken manually, there are now a large number of POS tagging software tools that automatically categorises the words in a text. Each POS tagger has its own strengths and weaknesses. Various taggers have been used including TOSCA (van Halteren, Tweedie, and Baayen, 1996), Brill (Argamon-Engelson et al., 1998; Koppel et al., 2005), and Amazon (van Halteren, 2004) to name a few. (Johnson, Malhotra, and Vamplew, 2006) tested several different taggers and concluded that Qtag (Mason, 2006) was faster and more robust than many others. Being robust, able to deal with misspelled words, non-English words and other informal terms, is particularly important for this study, as the corpora contain a number of words that fit in to these categories Several essays contain words that are not English, but from the authors first language, used to express a concept that is then described in English. POS unigrams (Tsur and Rappoport, 2007), bigrams (Abbasi and Chen, 2008; Koppel et al., 2005; Tsur and Rappoport, 2007), trigrams (Argamon et al., 2009; Koppel et al., 2010), have been shown to be effective in authorship analysis studies. (Koppel et al., 2005; Koppel et al., 2010; Wong and Dras, 2009) both used POS n-grams that were rare when compared to the Brown Corpus to identify unusual word orders and usage in text. While Koppel et al (2005) found that they were useful, they noted that genre could have an impact on what was considered ‘rare’. Wong and Dras (2009) found that rare POS bigrams were not as effective in identifying native language as more common ones.

2.5.4. Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al (2007) developed the Linguistic Inquiry and Word Count (LIWC) program as part of a psycholinguistic study of language and disclosure to aid in the study of various emotional, cognitive and structural components of both written and verbal speech samples. Studies have shown that an individual's mental and physical health impacts on their choice of words (Cohn, Mehl, and Pennebaker, 2004). LIWC relies on an internal dictionary of approximately 4500 words and word stems with each identified in one or more categories. Files are analysed by comparing them, word by word, to the internal dictionary. When a word is found the categories that include that word are incremented. The program has 80 categories: 4 general descriptor categories (word count, sentence length, % of words found in dictionary, and long words – words greater than 6 characters), 22 standard linguistic dimensions (counts of pronouns, articles auxiliary verbs etc), 32 word categories using psychological constructs (affect, cognition, etc) 7 personal concern categories (home, world, leisure etc), 3 paralinguistic dimensions (assents, fillers and nonfluencies) and 12 punctuation categories. A word can be part of more than one category. For example, cried is part of 5 categories: sadness, negative emotion, overall affect, verb and past tense. The output of the program is a vector with the name of the file analysed as the first element, and numeric values for each of the other 80 categories. A list of the LIWC categories and example words is given in Appendix A.

The categories for each word in the LIWC dictionary were assigned using a four step process that was iterated over several years (LIWC manual). The words were initially collected using thesauri, Standard English dictionaries and brain storming sessions among the 3-6 judges to apportion the words to initial categories. The collection of words was then rated by three independent judges. These judges voted on each word to either leave it in the category, remove it from the category or included in another category. They also added words that they believed should be in each category. This phase was repeated a number of times. Finally the original LIWC program was revised and categories that had little or no use were deleted and several new categories were included. The removal and addition of categories were undertaken using a similar process as the original ranking of the words, ie a series of passes where category of each word was assessed by three independent judges (LIWC manual). The in depth assessment of the categories and words gives each category in LIWC greater breadth than those of the other four feature types. LIWC differs most profoundly from the common words and character bigrams feature types in that each feature is an aggregate of different words rather than counts of a single word. While the POS n-grams also contain aggregate counts, LIWC has more informed aggregates than simple parts of speech choices. LIWC mitigates effect of and content words it includes because the words are part of a class of content words not the words themselves. Therefore the topics are included in general rather than specific terms. For example: if a text was speaking about domestic violence, LIWC may credit the family category, the negative emotion category and perhaps the death category, but does not identify the topic of family violence specifically. A text discussing the effect of road trauma or work place injuries on family members could well show the same pattern in the vectors.

Newman et al. (2003) noted in their study on the effect of deception on language use, that one of the markers to identify deception was a change in the pattern of use of function words and used LIWC to confirm this. LIWC has also been used in a number of studies into the effect of trauma, mental health and attitude on written and transcribed spoken language (Cohn et al., 2004; Mehl and Pennebaker, 2033; Pennebaker, Mehl, and Niederhoffer, 2003), however there do not appear to have been any studies into the effectiveness of its application to classification of texts by an author’s demographic characteristics. As LIWC exploits the emotive and cognitive significance of the use of words, it would seem plausible that it could be applied effectively to the problem of authorship profiling on many different demographic characteristics.

(Bamman et al., 2014) used a feature set based on a dictionary that they developed, to identify emotive words, emoticons and words and parts of speech that indicate assent, hesitation and profanity, among other things and found it to be successful. This feature set was not used in this study because many of the features they used including the emoticons were not present in either corpora chosen, and may not be present in all chats in the real world. Other studies have used LIWC to profile twitter feeds, but only in conjunction with content based features and metadata pertaining to twitter behaviour with mixed results. While most found that LIWC was not as effective as the content and metadata features (Fink et al., 2012; Ludu, 2014; Nguyen et al., 2013), content words could be faked by a person wishing to conceal their gender or age group. This work will explore whether psycholinguistic features (using LIWC as an example) are more efficient at determining the first language, gender and age group of an author without adding topic specific content words that can be easily faked.

2.6. Corpora

There have been nearly as many corpora used for authorship analysis as there have been authorship analysis studies. They range from blogs (Koppel, Schler, and Argamon, 2008), emails (de Vel, Anderson, Corney, and Mohay, 2001) twitter feeds (Bamman et al., 2014; Fink et al., 2012; Ludu, 2014), message boards, forums and chat rooms (Abbasi and Chen, 2005; Estival, Gaustad, Pham, Radford, and Hutchinson, 2007; Li et al., 2006) to formal texts and essays (van Halteren et al., 2005). Every author has a large number of different demographic characteristics. To study a particular demographic characteristic, it is necessary to find a corpus that has that demographic characteristic identified for the author of each text, with a sufficient representation of each value or class of the characteristic. A deciding factor in choosing a corpus is that it must also be the actual writings of the authors and not have been subject to any post transcription editing that could mask the authors original writing style (Baayen et al., 2002). One of the problems associated with authorship analysis is the availability of suitable corpora and while in many cases the corpora use are not ideal, they are the best available (Rudman, 2010). Incorrect corpus selection could give excellent results but results that are only applicable to that corpus and that cannot be extended to the general populace.

For this study, the corpus, or corpora, selected have to be in English and have the author's first language, age group and gender tagged. It must also have sufficient numbers of each of the classes being studied. There has been a number of authorship studies conducted on corpora in languages other than English including the Dutch Authorship Benchmark Corpus (Baayen et al., 2002; van Halteren et al., 1996) and the corpus compiled by (Stamatatos et al., 1999) in Modern Greek. Translation of these corpora would erode the authors' original style and the resultant documents would show more of the translators' language characteristics than that of the original authors. Many other corpora that have been compiled for specific purposes are also not suitable for this research. The corpus compiled for the Dark Web project (Abbasi & Chen, 2005) was designed to identify the similarity in the writing styles of extremist authors and as such, the results are most likely not applicable to the general populace. Other specific purpose corpora include one compiled to examine the impact of the September 11, 2001 terrorist attacks on New Yorkers conversation style (Cohn, Mehl, & Pennebaker, 2004), and a set of hand written essays to assess the effect of lying and deception on writing style (Newman et al., 2003).

None of the above corpora are suitable for this research. The ideal corpus would have sufficient numbers of authors from several different first language backgrounds with a large range in ages, with equal numbers of males and females in each age group. There is no corpus currently in existence that meets all of these criteria. Because this research could also be of interest to law enforcement agencies attempting to identify online predators, a corpus consisting of de-identified chats of predators grooming underage victims would also be ideal. There is no publicly available example of this. The closest would be the Perverted Justice corpus, but it is transcripts of law enforcement personnel posing as children to apprehend paedophiles posing as children to lure their victims (Gupta, Kumaraguru, & Sureka, 2012). there are two that are sufficient for the different parts of the study.

One corpus that has been used in several first language authorship characterisation studies (Argamon et al., 2009; Kerremans et al., 2005; Tsur and Rappoport, 2007; Wong and Dras, 2009) is the International Corpus of Learner English (ICLE) (Granger, 2001). The ICLE has been compiled over 10 years of collaboration with several universities. It is a collection of essays written by students of English as a Foreign Language (EFL) in various universities across Europe, Asia and Africa, with 16 separate language backgrounds. The sole purpose of compiling this corpus has been to further the study of language. It has been compiled under very strict guidelines. There is only one essay per student with the average length ranging between 874 (Dutch) to 502 (Chinese). The age, gender, first language and several other demographics of the authors are identified in the metadata of the corpus. The corpus is not artificially balanced for any variable, with approximately 76% of the corpus by females, and the age being relatively homogenous. Granger (2001) speculates this is because the ‘soft’ sciences attract mainly female students, and most of the students are from the age group associated with university students in general, early to mid 20s. Each of the first language section of the corpora has a different number of essays, ranging from 243 (Czech) to 982 (Chinese). The studies that have used this corpus have balanced the essay numbers so that the same number of essays are used for each first language group studied (Argamon et al., 2009; Koppel et al., 2005).   Although the ICLE corpus has been shown to have a topic bias that could impact on the accuracy of any analysis (Tetreault, Blanchard, & Cahill, 2013), the selection of features for this study should eliminate the effect of this bias.

The second corpus to be used in this study is the Blog Authorship Corpus compiled by (Schler et al., 2006). This corpus is a collection of the blogs of over 19,000 authors gathered from blogger.com in August 2004. There are a total of 681,288 posts consisting of more than 140 million words, averaging out to 35 posts and 7,250 words per blogger. The blogs were pre-identified with the age group, gender, occupation and astrological sign for each blogger. While the gender and age group is identified for every blogger, if the occupation was not available, it was marked as ‘unknown’. Although the demographic data for the authors is self-reported, other studies that have used this corpus have accepted the validity of age and gender information (Koppel, Schler et al. 2009, Schler, Koppel et al. 2006, Argamon, Koppel et al. 2009). If any incorrect data is included, it is assumed that it will be treated as noise by the classification process. Before making the corpus publicly available, (Koppel, Schler et al. 2006) divided the corpus into three age groups; 13 to 17 years, 23 to 27 years and 33 to 47 years. The actual age value is not available for each blog, only the age group to which the author belongs. The groupings were to give distinct adolescent, twenties and thirties age groups and eliminate any ambiguity because of mature or immature individuals close to the age boundaries. Each age group is balanced for gender, but the age groups themselves are heavily skewed towards adolescents and twenties, with more than three times as many of each of these two groups than the thirties.

2.7. Summary and Implications for Study

There has been a long and varied history of authorship characterisation, dating from before the advent of computing. Many of the earlier methods relied on painstakingly marking up texts and manually collecting and analysing statistical information collected on lexicographical features. Authorship analysis techniques have improved in both speed and accuracy as computing power has increased and the cost of memory space has fallen. A number of techniques have been examined for effectiveness including Principal Component Analysis, Neural Networks, Naive Bayes methods decision trees and Support Vector Machines. Studies have shown that Support Vector Machines are more effective because they can handle noisy data more effectively than other methods (Li et al., 2006; Zheng et al., 2006). There are three main streams in authorship analysis, however the one pertinent to this study is that of authorship characterisation or authorship profiling (Li et al., 2006; Mala & Geetha, 2007).

Authorship profiling endeavours to identify demographic data about the author of a text by idiosyncrasies, patterns or other clues within the text itself. The demographic features targeted by this thesis are first language, age group and gender. These have been chosen because they are of specific interest to law enforcement agencies(Gupta et al., 2012; Whittle, Hamilton-Giachritsis, Beech, & Collings, 2013). An individual's first language can affect their grammar structures and word choices in second and subsequent languages learned. The unique phonology, morphology, syntax and semantics of the first language can colour the use of the second language (Lippi-Green, 1997; Ortega, 2009). Even if the grammatical structures are technically correct within the second language, it sounds awkward to a native speaker and clearly indicates a non-native language background.

Human beings are creatures of habit, and tend to repeat patterns of behaviour that have been successful in the past (de Vel et al., 2001). The use of language is no exception. An individual will continue to use the language patterns and idioms that they learned as a child as they age, even though the youth of subsequent generations develop new language styles and idioms. Studies have also found that as individuals age their language tone tends to change from a more subjective to a more objective view. These studies have also found that younger people are obsessed with time and tend to use more negative emotive terms, whereas older people use more positive and fewer time specific terms (Pennebaker et al., 2003). The combination of changes in tone while other speech patterns remain the same could easily identify the age group of the author from stylistic pointers in their writing.

Anecdotally, men and women communicate using different linguistic tones and styles. Studies have found that these differences do exist and can be measured. Females tend to use language as a social bonding tool, while males use it as a vehicle to exchange information (Mehl & Pennebaker, 2003). A study also indicted that these differences are the result of naturally occurring hormonal differences rather than social expectations (Pennebaker et al., 2004). The unique linguistic patterns in male and female language use could be used to identify the gender of the author of a piece of text.

To identify the differences in language use between any of the demographic groups, the correct features must be identified and labelled. Feature selection is an important step that can heavily influence the success or failure of the classification of the text. Many different types of features have been used in the past with varying degrees of success. These types include lexicographical feature, syntactic features, structural features and features that measure the number and type of errors in a text. The most common way to identify these feature types are with function or common words and character bigrams (lexicographical) and POS *n*-grams (shallow syntactic). The other feature types require different types of manual mark-up methods. However, there is another feature type that, to date, has not been used in the authorship profiling arena, that of LIWC. This is a feature set that has been collated by a number of psychologists to indicate the deeper psychological implications of word choice and linguistic style. This thesis explores the accuracy of this feature set for authorship profiling, both when compared to and when combined with the other, more commonly used feature types.

The final piece of the puzzle for authorship profiling is the corpora used for classification. The corpora used must have sufficient, balanced and identified examples of all possible values for the demographic class being studied. The text must also be in as close to raw form as possible with no editorial amendments or corrections applied post transcription. The corpus must also be free of other influences that may identify the demographic classes by a mechanism other than language style – for example topic. There was no one corpus that was suitable for all three demographic classes being studied. Two corpora were chosen, the ICLE for the first language study and the Authorship Blog Corpus for the age and gender studies. Both these corpora are freely available and have been collated specifically for the study of language use. They have also both been previously used in similar studies, therefore providing a good bench mark for the effectiveness of the methods used in this study.

The Linguistic Inquiry and Word Count (LIWC) program (Pennebaker, Booth et al. 2007) can be used to analyse the psychological significance of features in the text, which can impact on linguistic choices. It was developed as a means to study a number of psychological characteristics in an individual's language, including emotional, cognitive and structural characteristics, and has been used in psychological research. The published works using LIWC for demographic profiling have used it in conjunction with content and metadata based features for twitter feeds. However there is little or no research using LIWC as a single type feature set or in combination with other features based solely on authorial style features rather than content features. This study will apply the LIWC feature set to the problem of profiling the first language of an author using the ICLE corpus (Granger, 2001) and to the problem of identifying the age group and gender of an author using the Blog Authorship Corpus (Schler et al., 2006)

This study will also examine issues of feature reduction and the impact of document length on the accuracy of profiling.

Chapter 3 Methodology

This chapter will discuss the processes used to facilitate the examination of the research questions. The aim of this research is to isolate a robust set of features that can identify the first language, age or gender of an author from the text that the write. Because many electronic communications are very short texts, the research also explores the effect of reducing both the size of the text and the size of the feature set on the accuracy of the classifier. It is hypothesised that the LIWC feature set will increase the accuracy of the classification in all cases as both a standalone feature set and when combined with other feature types. A series of experiments were designed to identify the difference between the LIWC feature set and similar numbers of the other feature types, and then to measure the effect of adding LIWC to other single type feature sets as well as the effect of adding or removing it from combinations of the other feature types. Experiments were also designed to examine the effect on classification accuracy of reduction of the text size, the feature set size and both the text and feature set sizes simultaneously.

This research project used an empirical methodology. Empirical research derives knowledge from actual experiments rather than theory or belief, from observable and measurable phenomena. The research examined the effectiveness of psycholinguistically derived features when compared to the traditionally used lexicographical features in authorship characterisation classification exercises. The accuracy of the features was measured in both single type and combined type feature sets. The effect of ranking the features using different feature selection algorithms and selecting the top n features was also examined as well as the effect of reducing the number of words in the texts being classified.

This chapter will also discuss the corpora used for the research, the pre-processing necessary to extract the data and the feature types used for the classification exercises. It will also discuss the methods used to initially limit the three larger feature sets to manageable and pertinent feature sets, the ranking algorithms used in the feature reduction experiments and the method used to reduce the text size of the documents in the text reduction exercises.

Lexicographic features, which can extract syntactic, semantic and phoneme information, are obtained by counting the number of times a particular feature appears in a given document, whereas the psycholinguistic features used are based on the evaluation of the use of words and word stems. Whereas any word, word fragment or part of speech would only be included in a single lexicographical feature in a specific feature set, it could be included in several psycholinguistic features. The lexicographic features used for this research were function words, character bigrams, POS unigrams and POS bigrams. The psycholinguistic feature set used was LIWC (Pennebaker, Booth, & Francis, 2007).

The classification exercises were executed on two different corpora, one for the author first language classification exercise and one for the gender and age group and classification exercise. The following section gives detailed descriptions of both corpora used.

3.1. Corpora Used

One of the problems associated with authorship analysis is the availability of corpora, and that while the corpora used are often not ideal, they are the only ones available (Rudman, 2010). The choice of corpus could impact on the outcome of the classification exercise. For example if the subject matter boundaries of a corpus matched the class boundaries, content-based features could be very successful in classifying the corpus, but may only be useful for that particular corpus and not be transferable to a larger, less specific corpus. It is necessary to find a collection of texts that have the classes identified and sufficient numbers of documents in each of the demographic classes being examined. Every author has a large number of different demographic characteristics. To study the features that best identify these characteristics within an authorship analysis system, it is necessary to obtain a corpus that has the characteristics being studied identified for each author, with significant representation of each output class. It is exceedingly difficult to find a corpus that has more than one or two characteristics identified for the authors within these guidelines, and prohibitively time consuming to create one.

After some investigation, two corpora were chosen for this study. The International Corpus of Leaner English (ICLE) (Granger, 2001) for the first language section and The Blog Authorship Corpus (Schler et al., 2006) for the age and gender section.

The ICLE corpus was the largest corpus available at the time this study commenced. It has 16 separate first languages with sufficient examples of each class. This corpus has also been used in previous first language profiling studies and therefore gives a better comparison between the results from this study with the previous ones. However the ICLE corpus does not have a balanced representation of age groups or gender, therefore another corpus had to be sourced for those sections of the experimentation. The Blog Authorship Corpus has been compiled specifically for this type of research. It has balanced numbers of both and gender classes with sufficiently large examples of each. This corpus has also previously been used in age and gender classification experiments and therefore facilitates direct comparison between this and previous studies. The corpora were sourced and converted into appropriate formats for the experiments. Both corpora also needed some pre-processing, which is detailed in the following sections

3.1.1. International Corpus of Learner English (ICLE)

The International Corpus of Learner English (ICLE) has been compiled over more than ten years of collaboration between several universities across Europe, Asia and Africa. It is a collection of essays written by students of English as a Foreign Language (EFL) in various universities in several countries across these continents. The sole purpose of compiling this corpus has been to further the study of language. It has been compiled under very strict guidelines. The essays are between 181 and 3366 words long, with an average of 384 to 874 words per essay within first language group. and there is only one essay per student. Although the age, gender and first language, along with several other demographics of the authors are identified in the metadata of the corpus, it was only used for the language study for the following reasons. The corpus is not artificially balanced for any variable, with approximately 76% of the corpus by females and the age being relatively homogenous. This is probably because the ‘soft’ sciences attract mainly female students, and most of the students are from the age group associated with most university students, early to mid 20s (Granger, 2001).

The ICLE corpus is supplied with a complete software package that allows the user to select essays using various criteria. Using the selection criteria on the ICLE interface, it is possible to select essays on first language, gender and age, as well as several other criteria (time in English speaking country, time studying English, etc). It is not possible to extract one essay at a time. Therefore the essays were extracted in language/gender groups. To make sure that all the essays were extracted, a total of 48 queries were run. Each of the 16 language groups had three queries run on it, one for male, one for female and one for unknown gender. The files were saved under separate names in a single directory. The names consisted of the two letter language code used by the ICLE and a letter indicating the gender: ‘m’ for male, ‘f’ for female and ‘u’ for unknown. The files were all saved as .txt files. This resulted in 41 text files, since seven of the languages did not have any files with unknown gender.

The text files created with this exercise had the essays clearly delineated with each essay being separated from the next by two blank lines and each one having a unique title line: eg "<ICLE-BG-SUN-0001.1>". The first four letters indicate it is an ICLE essay, the second group of two letters are the code for the language group used by the ICLE group (see Table 1), the third group of three letters indicates the university where the student wrote the essay and the digits are the unique code for the student.

Language	Two Letter code	Language	Two Letter code	Language	Two Letter code	Language	Two Letter code
Bulgarian	bg	Finnish	fi	Japanese	jp	Spanish	sp
Chinese	cn	French	fr	Norwegian	no	Swedish	sw
Czech	cz	German	ge	Polish	po	Tswana	ts
Dutch	dn	Italian	it	Russian	ru	Turkish	tr

Table 3-1: List of language codes of first language groups present in ICLE corpus

The essay files were further processed to decompose them into separate files, each containing one essay. To split the essays a text handling program was written in Python. The inputs for the program are the name of the directory that holds the combined essays, and the name of the directory that will hold the individual essays. A search through a random selection of the combined essay files revealed that no line in the body of the text started with the character string "<ICLE" therefore that string was used as the trigger to separate the text. The program uses the essay title to get the information to create the file name for the essay file. An example of an essay file name is "bg_f_27_25.txt". The first group of two letters is the language code from the combined file name, in this case, Bulgarian. The second, single letter, group indicates the gender of the author, again from the combined file name. Of the two numbers, the first is a counter created by the program to keep track of the number of essays produced and to enable a count back if any problems occurred in processing, the second number is the code from the essay title line, based on the unique code given by the ICLE group.

The sixteen language groups also have different numbers of essays in them, ranging from 241 (Czech) to 982 (Chinese). Other studies that have used this corpus have balanced the essay numbers so that the same number of essays are used for each language studied (Argamon et al., 2009; Koppel et al., 2005; Wong & Dras, 2009), and this was the path taken in this study as well. 241 essays were randomly selected from each of the first language groups, creating a balanced corpus consisting of 3,856 files (241 * 16). This meant that while all of the Czech essays were included in the corpus, only approximately one fifth of the Chinese ones were included.

3.1.2. The Blog Authorship Corpus

The Blog Authorship Corpus is a collection of postings from over 19,000 bloggers with approximately 35 posts per blogger with 140 million words in total. The blogs are identified for the age group and gender for each blogger. The demographic data for the authors is self-reported, however other studies that have used this corpus have accepted the validity of age and gender information (Argamon et al., 2009; Koppel et al., 2009; Schler et al., 2006). If any incorrect data is included, it is assumed that it will be treated as noise by the classification process. The corpus is divided into three age groups; 13 to 17 years, 23 to 27 years and 33 to 47 years. The actual ages of the participants are not available. The groupings were to give distinct adolescent, twenties and thirties age groups and eliminate any ambiguity because of mature or immature individuals close to the age boundaries (Schler et al., 2006). Each age group is balanced for gender, but the age groups themselves are heavily skewed towards adolescents and twenties, with more than three times as many of each of these two groups than the thirties.

The blog corpus was downloaded from (http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm) in April 2010. There were over 19,000 files downloaded, with each containing the entire set of posts for each blogger. The individual posts were labelled with the date and html tags identifying the start and finish of the post. Because these tags and date stamps could have created noise or bias in the classification the files were pre-processed to remove them and convert the files to a .txt format.

3.2. Pre-Processing of Corpora

Both corpora contained some foreign words and characters. Although it is unlikely that these words would influence the age group or gender classification, and would have no effect on the LIWC, function word or POS based features, they could bias the language classification experiments for the character bigrams, and so needed to be eliminated for the selection of those features. The previous authorship characterisation studies that have used character bigrams have used a combination of letters, numbers and some punctuation (Tsur & Rappoport, 2007; Wong & Dras, 2009). This also influenced the decision to eliminate foreign characters.

3.2.1. ICLE Pre Processing

The ICLE corpus had 3092 character bigrams and 62,913 words. It contains very few non-ASCII letters, other than those used in foreign words, such as the ‘ñ’ character in *señor*. There were a considerable number of instances of this situation, where the author has used a word from their first language. The corpus contained 14,552,538 individual characters, made up of 136 different characters when the corpus was converted to lower case. Given the format of a standard English, QWERTY keyboard, one would expect only 68 different characters and three types of white space in this situation. Manual examination revealed that there were 26 letters (80.12% of the total characters present) 10 numbers (0.11% of the total characters), 6 punctuation characters (1.82% of the characters) and a total of 87 unknown characters or characters from outside the normal letters, numbers and punctuation characters expected. These characters accounted for only 0.39% of the total characters present.

The only feature set that these foreign letters or words would affect was the character bigram feature set. The most effective way to deal with them was to omit the character bigrams that included the foreign character. The rationale behind removing non-English characters (which would undoubtedly be useful in identifying first language) is that in many areas of potential application of this technique such characters may not be available. It was therefore considered undesirable to produce a technique which relied on the presence of these features. The character set used by Tsur and Rappoport (2007) consisted of 26 lower case letters, the digits 0 to 9, and punctuation characters. The exact punctuation symbols used in the previous study were not available so a set of six were included by the pre-processing used in this study; a full stop, a comma, a colon, a semi-colon, a question mark and an exclamation mark. Braces, brackets and other characters were also excluded. This was to aid in comparison to the work by (Tsur & Rappoport, 2007) and because they are not commonly included with punctuation marks. These characters also made up a negligible percentage of the total characters present. White space was also included, but no differentiation was made between tabs, returns and spaces. All other characters that appeared in the corpus were ignored. The documents were converted to lower case letters and then the character bigrams were extracted. If the value was not in the legal character set the character was discarded. The character before the ‘illegal’ character was also then discarded, so the next bigram started after the ‘illegal’ character. So, for example, the word *señor* was broken up into four bigrams: (*space*)*s*, ‘*se*’, ‘*or*’ and *r*(*space*). The bigrams ‘*eñ*’ and ‘*ño*’ were not included because they contain a non-English character ‘*ñ*’. Bigrams that resulted from the omission of any non-English characters were also not included. In the example above, the bigram '*eo*' which would result from the omission of the '*ñ*' character is not considered as a valid bigram in this case since it is not present in the raw text.

3.2.2. Blog Pre Processing

The blogs, being a far less formal medium, had considerably more non-English words and characters. Of the 19,320 blog files, 1,554 included non-English characters and words and many of these had large sections of non-English text. The characters used include non-Latin characters, such as Mandarin characters. There was also the problem that many English words contained non-ASCII characters because the texts were written, in many cases, using a larger character set, so there were a number of different quotation marks, apostrophes and hyphens or dashes. In all, there were 115 non-ASCII characters used, with over two million instances of these characters. Sentences containing examples of each of the non-ASCII characters present were identified and manually examined. If it was a variation on a standard ASCII character, the binary value of the 'illegal' character and the character it was supposed to represent were noted. Part of the pre-processing was to replace the relevant non-ASCII characters with their ASCII equivalent, thus retaining the authors' original intent with quotation marks, apostrophes and hyphens or dashes.
If the non-ASCII character was not accounted for either by a standard ASCII value or the variations character lists, it was treated the same way as the non-ASCII characters in the ICLE corpus, i.e. character bigrams incorporating this non-standard character were omitted. All legal ASCII characters were used in the blog corpus, rather than the restricted set used for the ICLE corpus, because there were a number of braces, emoticons, and other punctuation that were used throughout the blogs that could be an indicator of age and/or gender.

3.3. Classification Methods

This study used the Sequential Minimal Optimisation (Platt, 1998) algorithm of the Support Vector Machine (Fradkin & Muchnik, 2006) implemented by the WEKA suite of machine learning algorithms (Witten & Frank, 2005). While other classification methods are available, (Zheng et al., 2006) compared back propagation neural networks, decision trees and SVM classifiers, and found that SVM are as effective if not more so than the others tested.

While there have been a number of machine learning methods used for authorship analysis, including neural networks (NNs), Support Vector Machines (SVMs) and decision trees, in comparative studies, SVMs have been found to be at least equal to, or more effective than other methods. This is because they can handle larger, noisier data sets (Abbasi & Chen, 2005; Abbasi & Chen, 2008; de Vel et al., 2001; Li et al., 2006) and others. The particular incarnation of the SVM that has been used in a number of studies is the one supplied by the WEKA tool kit (Witten & Frank, 2005). This specific model of the SVM classifier has been used by (Argamon et al., 2005; Argamon et al., 2007; Estival et al., 2007; Li et al., 2006) in similar studies to the one detailed in this thesis. An SVM classifier is inherently a binary classifier, that is they can only compare two classes at a time. To overcome this limitation in multiclass classification problems, the classifier creates several classifiers, one for each pair of classes and combines them into an ensemble (Sazonova & Matwin, 2014). The output from the WEKA SVM shows this clearly. There is a record for each pair of classes, with the final result being the combination of all of the results.

The implementation of the SVM in WEKA comes with several different kernel options, (ie various algorithms for pattern analysis) and the choice of kernel can impact on the results of the classification exercise. Therefore, several preliminary classification exercises were undertaken, on the one third of the data quarantined for the purpose, to test the effectiveness of the various kernels. It was found that the default kernel gave equal or better results than the others and was more computationally efficient. Therefore, the default kernel settings have been retained for this work. All classification experiments used ten-fold cross validation. The classifier evaluation options that were used were the default ones with the addition of the Output Predictions option.   The Output Predictions option gives an output that has the result for each case separately recorded, the actual classification and the predicted classification. Each fold is also separately recorded.

3.4. Feature Selection

There were five different feature types used in this study: LIWC, function words, character bigrams, POS unigrams and POS bigrams. Although content words, structural and error features have been used in other studies (Argamon et al., 2009), content words give the topic more than the author’s style (Koppel et al., 2006), any structural features in the corpora have been removed or changed and error features are too reliant on the particular error checking software, prohibitively expensive in terms of time and computing resources to tag (Tsur & Rappoport, 2007). Of the five feature types used in this study, the POS unigrams and LIWC only have a small number of features, 70 and 80 respectively, but the other three (the function words, character bigrams and POS bigrams) have many thousands, far too many for all available features to be used with the WEKA system. Therefore they were pre-processed to reduce the overall numbers prior to being used for classification. The remaining two feature sets have a limited number of features - the complete LIWC feature set consists of 80 features and the POS unigram has a total maximum feature set size of 70. The feature reduction was achieved by ranking the features within each feature type, and selecting only the top ranked features. The four methods tested in this study are described below. All calculations were done separately for each of the function word, character bigram and POS bigram feature sets, for each of the three classification exercises – first language, age group and gender. To avoid over training and bias, each of the corpora were randomly divided into three sections. One third was used for the feature selection activity while the remaining two thirds were combined and used for the classification exercises.

3.4.1. Corpus Relative Frequency

The corpus relative frequency (*crf*) was developed for this study to calculate the impact of a given term on the corpus as a whole. The frequency for each term was calculated across the corpus as a whole as shown in Equation 1

$$(1) \qquad \mathrm{crf_i} = \frac{\mathrm{cts_i}}{\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{cts_i}}$$

where the corpus relative frequency (*crf*) of term *i* is equal to the corpus term score (*cts*) of term *i* divided by the sum of all the corpus term scores for every term in the corpus (*N* terms). The corpus term score is simply the number of times the term appears in the corpus.

3.4.2. Paired Difference Frequency

Another way of ranking terms is to measure the difference between frequencies of the same term in pairs of classification classes (ie between two different languages or between two different age groups). This method was also developed for this study to find features that had the highest impact in distinguishing between classes with in the corpus. First the class relative frequency (*clrf*) needs to be calculated for each term in each class.

$$(2) \qquad \mathrm{clrf_i} = \frac{\mathrm{clts_i}}{\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{clts_i}}$$

The calculation for the class relative frequency (*clrf*) of term *i* is shown in Equation 2. It is the class term score of term *i* divided by the sum of all the class term scores (*clts*) for every term in the corpus (*N* terms). The class term score is the number of times the term appears in the class. This calculation results in a separate file of terms ranked in frequency order, for each class in the classification exercise (sixteen for the first language classification, three for the age group classification, but only two for the gender classification). The difference between the frequencies of each term in each pair of classes in the classification exercise was then calculated, resulting in the class pair files (Equation 3). The terms were then given a rank number. The program written to rank the terms allowed the user to specify how many terms from the top of the rankings were to be considered. The terms were given a rank number, the highest being the number of terms specified by the user, with each subsequent term being given a rank of one less than the one before. After the program reaches the term with the value 1, the remaining terms are given a value of 0.

(3)
$$\Delta_{\lambda,\mu} = \sum {}^{\mathrm{rank}}{}_{\mathrm{i}}\Delta_{\lambda\mu} \qquad \lambda \neq \mu$$

The pair relative frequency ($\Delta_{\lambda,\mu}$) value given to term *i* is the sum of the ranks in the difference in the frequency of term *i* between class *λ* and class *μ* where *λ ≠ μ*.

3.4.3. Document Relative Frequency

Grieve (2007) found that a very effective word frequency list was obtained by ranking words by the number of documents they appeared in across the corpus, rather than the number of times they appeared within the corpus as a whole (Equation 4). The document relative frequency (*drf*) for term *i* is the count of all the documents *j* that contain term *i*.

(4)
$$\mathrm{drf_i} = \sum_{\mathrm{j=1}}^{\mathrm{N}} \delta_{\mathrm{ij}}$$
$$\delta_{\mathrm{ij}} = \int \begin{matrix} 1 \text{ if term i is in doc j} \\ 0 \text{ if term i is not in doc j} \end{matrix}$$

3.5. Statistical Tests Used

Several statistical tests have been used for the analysis in this study. While a full explanation of each of the tests used is beyond the scope of this document, this section contains a brief discussion of each of the tests used and the reasons they were chosen. For all experiments undertaken in this thesis, the null hypothesis in all cases is that there will be no significant difference between the different feature sets used. The alternate hypothesis is that the use or inclusion of LIWC will improve the accuracy of the classification.

3.5.1. T-Tests

Unless otherwise stated in the discussion, all p values are obtained from an independent-sample one-tailed t-test. This test is used for the comparison of the mean score on a continuous variable of two different groups. It answers the question of whether there is a significant difference between these two means (Pallant, 2011). The samples are considered to be independent because the texts that are classified as correctly for one class variable have no impact on texts that are correctly classified in another; that is that they have no dependence on each other. One-tailed t-tests were used because question being asked is whether LIWC improves the accuracy of the classification, not whether the results will merely be different. One of the assumptions of using a t-test is that the variance within the groups is the same. Although not specifically documented in the thesis, the variance between the groups was tested and found to be suitable for the use of a t-test.

3.5.2. ANOVA

ANOVA (**An**alysis **o**f **Va**riance) is so called because it is used to compare the variance within a group, which would be due to chance, to the variance between that class and another, which would be due to the impact of the independent variable being tested. Where the independent-sample t-test is used for two groups the ANOVA is used to compare three or more groups. A significant F test indicates that the null hypothesis that all the groups are the same, can be rejected. However it does not indicate which of the groups is different. This can be identified by undertaking a post-hoc comparison using the Tukey HSD test (Pallant, 2011). If the interest is in the performance of one of the groups only, pairwise t-tests can be used to further refine the results (Sauro & Lewis, 2012). An ANOVA test was documented in this thesis where there were more than two groups and the results of the test were pertinent to further investigation. As the research questions relate to the impact of LIWC on the classification accuracy, pair wise testing using t-tests were undertaken to indicate the significance of the difference between the group containing the LIWC features and the other group (s) of interest.

3.5.3. Chi squared

The chi-squared test compares the observed frequencies or proportions of cases that occur in each of the categories and tests if it is significantly different from the expected frequencies or proportions (Mann, 2010). Chi-squared tests were only used in this thesis on the confusion matrices for the age group classifications over three classes (Chapter 6). They were used to investigate the poor results shown and to examine if any of the age groups had fewer correctly classified documents than the others and if the difference was significant.

3.6. Feature Types

Four of the five feature types used in this study are lexical in nature. The function words, character bigrams, POS unigrams and POS bigrams are derived from the frequency of given terms within a document. The fifth feature set, the Linguistic Inquiry and Word Count (LIWC), is a psycholinguistically derived feature set that has been used in psychoanalysis, however there is little or no evidence that it has been used in computational linguistics. The first two questions being researched in this study are:

1. Are psycholinguistically based features more effective than lexicographical features for authorship characterisation?
2. Is the theoretical basis for psycholinguistic features sufficiently different from that of lexicographical features that the combination of psycholinguistic features with lexicographical features will be significantly more effective in authorship characterisation than equal amounts of lexicographical features alone?

LIWC was the psycholinguistically derived feature set examined in this study.

3.6.1. Linguistic Inquiry and Word Count (LIWC)

LIWC is a program that has been designed by psycholinguistics to analyse the study of emotional and cognitive language in both written and spoken form (Pennebaker et al., 2007). The program converts text in to vectors of 80 features that consists of 22 standard linguistic features, 32 psychological constructs, seven personal concern categories, three paralinguistic dimensions and twelve punctuation categories. The program itself relies on an internal dictionary of over 4,500 words and word stems. While additional user defined dictionaries can be added to the program, the default dictionary was used for this study. Unlike the other feature types used in this study, a word can be part of more than one category in LIWC. The example quoted in the LIWC documentation is the word “cried”. It can be included in five categories: sadness, negative emotion, overall affect, verb and past tense. Categories also cover many different words. It is anticipated that this flexibility will give the LIWC features more discriminatory power in the language, gender and age group classification exercises.

3.6.2. Function Words

Function words have little semantic meaning in their own right, but instead serve to express grammatical relationships between other words or specify the mood or attitude of the author. Function words are ubiquitous in the English language, so much so that they are often overlooked when proof reading text, but although they make up only 0.04% of the vocabulary of English, they are provide over half the words that are used in communication in English (Kestemont, 2014)

Different studies have used different lists of function words, and different methods of obtaining the lists. Some studies have indicated that they used a list of function words (Koppel et al., 2005) however the lists were not publically available. (Zheng et al., 2006) did include the list of 150 function words that they used, both in English and Chinese. The English listing contained many of the words in the frequency, pair frequency, and document frequency lists produced. In fact the only function words in Zheng et al's list that were not in the top 200 of the lists produced by the methods used in this study were words that did not appear in the corpus at all, such as 'hither' and 'whither'.

Both the frequency and pair frequency methods had a large number of content specific words in the top 300 words for both corpora, which would have had to be manually removed if those lists were used. In the ICLE corpus, these words were closely coupled with a very limited number of topics, and while they could identify the first language groups present, the success was more likely to be due to the essay topics chosen by the language instructors in the universities participating in the ICLE program rather than any idiosyncrasies associated with language transfer from the author's first language. In the blog corpus, there were also a large number of topic specific words, relating to education, employment and hobbies. Again, while these words could very easily distinguish between age groups and genders, the objective is to produce a feature set that identifies the differences in language patterns rather than topic choice and that could be applied to corpora where the topics are homogenous between class groups. As Grieve (2007) suggested, the document frequency method produced a list that had very few content specific words. There were only two in the top 200 in the ICLE corpus, 'country' and 'countries' which can both be associated with a large number of topics, and none in the blog corpus Because the document frequency function word list could be used in its raw form, requiring no pre-processing, this method was chosen to rank the function word features for all classification exercises.

3.6.3. Character bigrams

A character bigram, as the name would suggest, a group of two characters. The character bigram features used in this study represent the phonemes of the words used by the author, and are therefore sets of two characters that appear consecutively within the text. White space has also been included in the character bigrams to indicate the characters use to start and end words. The character bigrams for the first language experiments also include a set of six punctuation characters, to indicate expression. No other characters from the standard ASCII character set were included in the text due to the method of creation and formality of the style. A manual inspection of the ICLE corpus also revealed that the other characters available on a standard QWERTY keyboard represent a negligible percentage of the characters present. The character bigrams for the gender and age group classification experiments include all characters that appear on a standard English keyboard that are also in the standard ASCII character set. The potential character set is expanded for these experiments because the medium for the texts was far less formal than the text used for the first language experiments and these characters could be indicative of language use specific to particular genders or age groups.

Ranked lists of these bigrams were obtained using the three methods listed in Sections 3.4.1, 3.4.2 and 3.4.3. The top 200 of each listing was tested for each of the three demographic characteristics being examined, on the one third of the relevant corpus that had been quarantined for pre-processing purposes. The most effective list was then selected for the remaining experimentation for the pertinent demographic characteristic.

3.6.4. Part of Speech *N*-Grams

A word is allocated a part of speech (POS) category depending on its syntactic function. POS categories include nouns, verbs, adjectives adverbs, prepositions, etc. Some words can be in more than one POS category depending on their meaning within a sentence. For example the word “can” could either be a noun (meaning “a metal container”) or an auxiliary verb (meaning “to be able”). The position of the word in a sentence will determine to which category that particular instance of the word belongs. The activity of POS tagging is the process of going through a document and assigning each word to its POS category. While this activity could, and before computing power and software progressed sufficiently was undertaken manually, there are now a large number of POS tagging software that automatically categorises the words in a text. Each POS tagger has its own strengths and weaknesses. QTag is a probabilistic tagger, in that it chooses a tag for a word based on the probability of the tag appearing with the surrounding tags within two words, both preceding and after, of the target word. It creates a matrix of tag sequences with the associated frequencies and assigns the most probable tag to the target word (Tufis & Mason, 1998). QTag has been proven to be robust and effective when there are a large number of unknown words in the text (Johnson, Malhotra, & Vamplew, 2006). The two corpora used in this study have the potential to contain a considerable number of unknown words, albeit foreign words, or web slang used by the authors. Therefore it was the one chosen for this research project.

The corpus was tagged using the QTag system, with the “tabular” and “tokenise” settings used. These settings produce a document that has each word/tag pair on a separate line, with one tag per word. There are 77 tags defined by the QTag software, but in both of the corpora used in this study, only 70 were present in the documents they contained. Because this resulted in a feature set of only 70 features, no ordering was necessary for the POS unigram feature set and the entire feature set was used. However, the 70 POS tags present in the corpora produced over with over 3000 POS bigrams in the ICLE corpus and more than 2600 in the Blog Authorship corpus.

The most effective method for ranking the POS bigrams for the first language classification was the pair frequency method discussed in Section 3.4.3. Again the ranking of the features for this section of the study was done using the third of the data quarantined for pre-processing. However, like the character bigrams there was no statistical difference between the methods for the age group and gender classification exercise, so the document frequency ranking was used. This was again due to the ease of production.

3.7. Feature reduction

The third and fourth research questions being examined in this thesis:

3. Do the number and/or type of features used in an authorship characterisation classification model have an effect on the success and accuracy of that model?
4. Is there an effective lower limit to the number of features that can be used for a classification model for authorship characterisation?

Answering these questions requires the feature sets to be systematically reduced in size to judge the impact on the accuracy of the classifier. This reduction is separate from the pruning of features discussed in Section 3.4: Feature Selection. That exercise was a broad based removal of features that were likely to be biased to the particular corpus, or that had little or no productive output due to sparseness or homogeneity of data values. This included content words that would identify topics specific to the contributing schools rather than the first language of the contributing students. The processes in Section 3.4 resulted in a reduced but still large pool of features, ranked by the number of times they appeared in the corpus, the number of documents they appeared in, or the difference in the number of times they appeared in particular classes. While the features chosen by the algorithms in Section 3.4 were useful and gave good results, it is unlikely that all of them were equally useful. In many of the large feature sets, some features could well be removed without an adverse impact on accuracy, but finding the correct subset of features is an important area of study in its own right (Liu, Motoda, Setiono, & Zhao, 2010).

The WEKA data mining package (Witten & Frank, 2005) contains a number of dedicated feature selection algorithms. Several of them were not suitable to the data sets being used in this study. However there were three (information gain, gain ratio and chi squared) that were applicable to the data being examined. Initial testing showed that the information gain and chi squared gave almost exactly the same feature rankings so only the information gain and gain ratio were used from the supplied feature selection algorithms. A third method for ranking the features was also tested: the J48 tree, also supplied by the WEKA package. (Witten & Frank, 2005) recommend the J48 tree as a method for reducing or ranking features, as the features that appear earlier in the tree have more discriminatory power than the features lower in the tree or the features omitted from the tree altogether. Information gain measures the level of impurity or entropy within the values for a given feature. Entropy is the measure of the information content that a feature supplies. Unexpected values, those that appear less often in the values for a feature, have more information than values that appear more frequently. The entropy formula takes the probability of the values as the basis for computing the entropy score (Shannon, 1948). Therefore, the higher the entropy score for a feature, the more information that feature supplies towards the classification of the data. Information gain can indicate how important a given feature within a feature vector will be, that is how useful it will be in discriminating between classes (Witten & Frank, 2005). However, information gain can be biased against features with large numbers of distinct values. The gain ratio algorithm is a modification of the information gain algorithm that reduces this bias. (Witten & Frank, 2005)

In the sections of the experimentation that required the feature sets to be systematically reduced these three algorithms (information gain, gain ratio and the J48 tree) were used to rank the features using the third of the corpora set aside for pre-processing. The resulting feature ranks were used to classify the remaining two thirds of the corpora.

3.8. Text Reduction

The fifth and final question being studied in this thesis is:

5. Is there an effective lower limit to the number of words in a text that can be classified using authorship characterisation techniques?

The trend in communication style is moving away from longer, more formal contact to shorter more frequent and informal contacts (Saevanee et al., 2011). The ultimate short communication form is Twitter, where messages are limited to 140 characters or less – less than 28 words if you accept the traditional ratio for English of five characters per word. There have been a number of studies on Tweets (the message unit for Twitter) but most have either been author attribution rather than author characterisation studies, (for example: (Bhargava et al., 2013; Saevanee et al., 2011; R. Schwartz et al., 2013) and/or have aggregated the Tweets from one author into a larger text for analysis (for example: (Bamman et al., 2014; R. Schwartz et al., 2013). This study will systematically reduce the text of a document to find if there is an effective lower limit after which it is no longer possible to identify the first language, gender or age group of the author at a greater level than that of chance. The level of chance is 6.25% for 16 classes (the first language classification), 33% for three classes (the age group classification) and 50% for two classes (the gender classification).

An author's attention to detail or mood could change as the writing in a text progresses, especially in the two corpora being used in this study. The ICLE corpus (Granger, 2001) is a series of essays written by tertiary English students. While the students may write the essays in their own time, they can also be done under exam conditions. An approaching time limit could lessen the attention to the details of English grammar and syntax. Conversely, the student could 'warm up' during the exam and their grammar and expression proficiency could increase as the exam progresses. There could also be a difference in expression in writing an introduction for an essay compared to the discussion and conclusion of the same essay. The Blog Authorship Corpus (Koppel et al., 2006) texts consist of all the blogs for a given blogger for the collection period. The author's mood or emotion about the topic could be different for different sections of the blog. A method to generate a sample that is not biased to one section of a text is to randomly select sentences from the text (Gamon, 2004). However examination of both corpora showed that there was a very wide spread of words per sentence across the corpora, ranging from sentences of one to three words, up to sentences that covered an entire paragraph or even the whole text, and that the spread was very irregular. The large variation in sentence size would make selecting a uniformly sized reduced corpus extremely difficult. Therefore using randomised sentences from each text to create the reduced text size corpus was not practicable. A method to create uniform chunks of text was required.

Given that one of the features being used in this study is POS bigrams, simply randomly selecting individual words from the texts to make up the required word count would not be appropriate and a random selection of words could also affect the LIWC feature selection. The human memory can hold 7 +/- 2 items, an idiosyncrasy that has been used in previous natural language processing to automatically infer the meaning of words (Watters, 2002). While (Watters, 2002) used chucks of nine words, chunks of eight words were deemed more suitable for this study. Eight words is between the maximum (nine) and average (seven) number of items that can be stored in short term memory and it is an even number so it will be a better fit for the POS bigram feature.

To obtain the reduced word texts, the texts were split up into lists of words, retaining their original order, and any punctuation marks. The words were then grouped into chunks of eight words and stored in a separate list. The number of chunks that would make up the required document size were then randomly selected and combined and saved to a text file that was labelled with the original file name and a number denoting the number of words in the text.

3.9. Summary

This chapter has discussed the methodology which was common throughout all of the research reported in this thesis. The following three chapters detail the experiments conducted on the ICLE corpus for first language and the blog corpus for age group and gender. Each classification required some specialised methodology, and these are detailed in the relevant sections.

Chapter 4 Profiling for First Language of Author

The previous chapter discussed the overall framework for the research including an analysis of the ranking methods for features, the feature types being used, and the corpora being examined.

This chapter looks at the results for the task of the first language classification which was undertaken on a sub section of 3856 essays from the ICLE corpus. The corpus consists of over 6,000 individual essays of tertiary students of English whose first language is one of the sixteen included in the corpus. The full number of essays was not used because each first language group has varying number of essays ranging from the most prolific (from the Chinese first language group with 982) essays to the least (from the Czech first language group with only 241). Table 4.1 gives each first language group, its language code, the number of essays present in each language group and the average number of words, the longest and shortest essay for each first language group.

First Language Group	Language Code	Number of Essays	Average Words per Essay	Longest Essay	Shortest Essay
Bulgarian	bg	300	634	1466	216
Chinese	cn	982	502	918	247
Czech	cz	241	843	1484	408
Dutch	dn	263	874	3366	312
Finnish	fi	261	726	1591	272
French	fr	314	657	2246	278
German	ge	445	541	1478	180
Italian	it	398	571	1088	259
Japanese	jp	366	542	990	400
Norwegian	no	316	640	1536	319
Polish	po	366	640	1098	229
Russian	ru	274	866	3082	184
Spanish	sp	250	775	2801	223
Swedish	sw	471	578	1050	293
Turkish	tr	276	708	1001	500
Tswana	ts	519	394	989	181

Table 4-1: Language code, number of essays and minimum, maximum and average length of text for each first language group (full ICLE corpus)

There was an average of 655 words per essay. The longest essay was a Dutch essay consisting of 3366 words while the shortest essay was a German one consisting of 180 words. The information supplied with the corpus makes no comment on the overall proficiency of any of the first language cohorts or the teaching methods employed at the individual universities involved.

The 241 selected files from the first language groups were randomly divided into three sections. The 80 essays in the first sections from each first language group were combined and used for pre-processing and ranking of features, and the remaining sections of 81 and 80 essays were combined and used for the experimentation. This was done to avoid any selection bias. The function words, character bigrams and POS bigrams from the first third of the corpus were ranked using the methods discussed in Section 3.4.

Previous studies have examined profiling the first language of an author when they are writing in English and their first language is one other than English. The ones that have used the ICLE (Granger, 2001) have only used five of the available sixteen first language classes. Others have used a corpus that was not available at the commencement of this research, the TOFEL 11 corpus (Blanchard, Tetreault, Higgins, Cahill, & Chodorow, 2013). This corpus consists of essays from eleven first language groups, rather than sixteen, but there are seven first language groups common to both corpora. When using over 600 features, a combination of character n-grams, function words rare POS bigrams and errors, Koppel et al (2005) achieved an accuracy of between 50% and 70% when using five first language classes from the ICLE corpus. Wong and Dras (2009), also using 125 essays from five first language groups from the ICLE corpus, (25 essays from each group) and almost 400 features consisting of function words and character n-grtams, achieved 65.14% accuracy in classification. An overall classification accuracy of 31.9% was achieved by Daudaravicius (2013) using all eleven first language groups from the TOFEL 11 corpus using character trigrams as features. Abu-Jbara et al (2013) using the same corpus had a classification accuracy of 43.0% using a combination of features consisting of function words, character n-grams, POS n-grams and errors.

The remainder of this chapter is organised as follows. The next section, Section 4.1, will document the results of the first language classification experiments conducted on the full sized essays in the testing portion of the corpus using both increasing numbers of features and various combinations of the five feature types being tested (LIWC, function words, character bigrams, POS unigrams and POS bigrams). Section 4.2 will present the results from experiments, again using the full sized essays, but using reducing numbers of features for classification, Section 4.3 will give the results when the size of the essays was reduced, Section 4.4 explores the effect of reducing the feature set size and the document size simultaneously and Section 4.5 summarises the results.

4.1. Comparing and Combining Psycholinguistic and Lexicographic Features on Full Sized Essays

4.1.1. Single Feature Types

The aim of this research is to ascertain the efficacy of psycholinguistically based features compared to lexically based features when used for authorship characterisation. To test this, classifiers trained using similar numbers of the five feature sets were compared. Note that only 70 POS unigrams were tested because that is the maximum number available. As can be seen in Figure 4-1, LIWC, the psycholinguistically based feature set, was between 3% (character bigrams, $p = 0.0128$) and 9% (POS bigrams $p < 0.001$) more effective than the lexically based feature sets.

Figure 4-1 gives the average accuracy over all sixteen first language groups for each feature type. There was considerable variation within each feature type across the first language groups. Table 4-2 shows the true positive rate for the individual first language groups. The true positive rate for the top four first language groups for each feature set is bolded and shaded, while the bottom four for each first language group are bolded and cross-hatched.

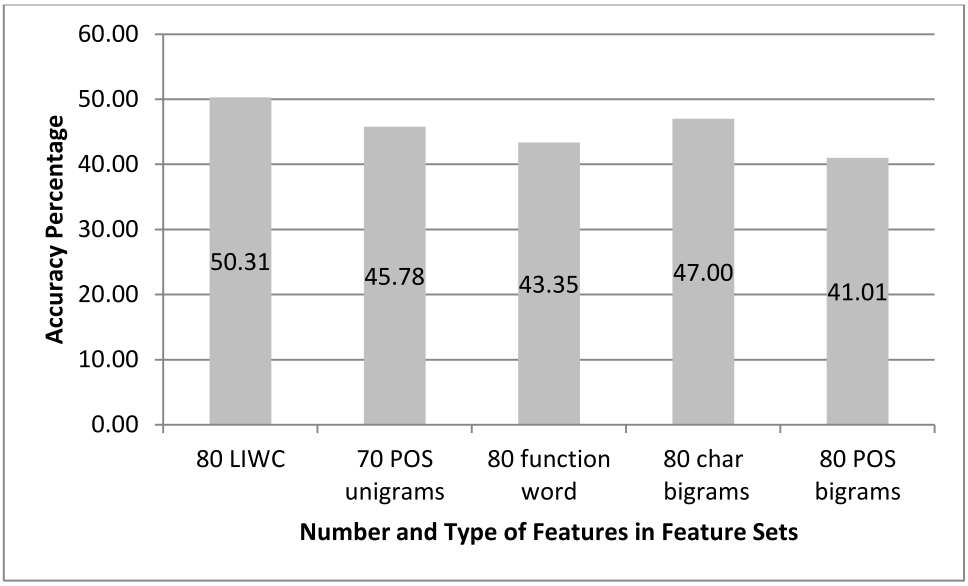

Figure 4-1: Comparison of accuracy for 80 LIWC, 80 function word, 80 character bigram, 80 POS bigram and 70 POS unigram features for first language classification

The first thing to note is that there is a considerable variation between the highest and lowest true positive rate within each feature set. The smallest difference between the best performed language (Chinese with 0.683) and the worst performed language (Swedish with 0.21) is seen in the function words, a difference of 0.466. The largest difference of 0.609 is seen in the POS unigrams, also between Chinese (0.826) and Swedish (0.217). There is a great deal of homogeneity across the five feature types when it comes to the rankings of the first language groups true positive rates. Swedish and Finnish consistently appear in the lower four first language groups, and Dutch and Russian make up the other two lower first language groups for LIWC, POS unigrams and character bigrams. Norwegian is in the lower four true positive rates for function words and POS bigrams, while French is the fourth for function words and German fills that place for the POS bigrams. A similar theme is played out for the top four true positive ratings. Tswana, Chinese and Japanese are in the top four true positive rates for all five feature types, LIWC includes Italian in the top four, POS unigrams and POS bigrams include Turkish, and function words include Polish.

The hypothesis that LIWC would give greater insight into the first language group classification problem because it is based on the psychological basis of word choice rather than ratios of various lexicographical features appears to be supported.

First Language Group	Number and Type of Features in Feature Set				
	80 LIWC	70 POS unigrams	80 function word	80 char bigrams	80 POS bigrams
dn	**0.306**	**0.238**	0.344	**0.256**	0.281
ru	**0.311**	**0.25**	0.373	**0.335**	0.348
fr	0.491	0.447	**0.329**	0.441	0.398
bg	0.569	0.494	0.431	0.531	0.294
cn	**0.783**	**0.826**	**0.683**	**0.758**	**0.752**
ts	**0.794**	**0.663**	**0.65**	**0.788**	**0.669**
no	0.425	0.325	**0.338**	0.406	**0.275**
sw	**0.273**	**0.217**	**0.217**	**0.317**	**0.236**
sp	0.5	0.431	0.388	0.419	0.431
jp	**0.627**	**0.683**	**0.615**	**0.609**	**0.602**
it	**0.646**	0.54	0.391	**0.565**	0.478
tr	0.559	**0.553**	0.491	0.528	**0.584**
cz	0.497	0.484	0.46	0.416	0.354
ge	0.488	0.369	0.425	0.45	**0.256**
fi	**0.28**	**0.317**	**0.298**	**0.23**	**0.211**
po	0.503	0.484	**0.503**	0.472	0.391
Avg.	0.503	0.458	0.433	0.47	0.41

Table 4-2: True positive results for first language classification for 80 LIWC, 80 function word, 80 character bigram, 80 POS bigram and 70 POS unigram features – highest and lowest for each feature set highlighted.

When the five separate feature types are compared within each first language group, as shown in Table 4-3, the LIWC feature set gives the highest true positive rate for nine of the first language groups. POS unigrams and function words give the highest for three first language groups each, and character bigrams and POS bigrams give the highest true positive rate in one first language group each. When the lowest true positive rate is compared within first language group, across feature type, LIWC and character bigrams do not give the lowest true positive rate for any language. Function words and POS bigrams give the lowest ranking for seven first language groups each, and POS unigrams gives the lowest for two first language groups uniquely, and shares one (Swedish) with function words. The LIWC feature set gives the highest average true positive rate, the POS bigrams the lowest.

First Language Group	Number and Type of Features in Feature Set				
	80 LIWC	70 POS unigrams	80 function word	80 char bigrams	80 POS bigrams
dn	0.306	**0.238**	**0.344**	0.256	0.281
ru	0.311	**0.25**	**0.373**	0.335	0.348
fr	**0.491**	0.447	**0.329**	0.441	0.398
bg	**0.569**	0.494	0.431	0.531	**0.294**
cn	0.783	**0.826**	**0.683**	0.758	0.752
ts	**0.794**	0.663	**0.65**	0.788	0.669
no	**0.425**	0.325	0.338	0.406	**0.275**
sw	0.273	**0.217**	**0.217**	**0.317**	0.236
sp	**0.5**	0.431	**0.388**	0.419	0.431
jp	0.627	**0.683**	0.615	0.609	**0.602**
it	**0.646**	0.54	**0.391**	0.565	0.478
tr	0.559	0.553	**0.491**	0.528	**0.584**
cz	**0.497**	0.484	0.46	0.416	**0.354**
ge	**0.488**	0.369	0.425	0.45	**0.256**
fi	0.28	**0.317**	0.298	0.23	**0.211**
po	0.503	0.484	**0.503**	0.472	**0.391**
Avg.	**0.503**	0.458	0.433	0.47	**0.41**

Table 4-3: True positive results for first language classification for 80 LIWC, 80 function word, 80 character bigram, 80 POS bigram and 70 POS unigram features – highest and lowest for each first language group highlighted.

The results given by Koppel et al (2005) and Wong and Drass (2009) only gave an overall accuracy, however the results given by Daudaravicius (2013) and Abu-Jbar et al (2013) give detailed results for all first language groups. The most accurately classified first language group by Abu-Jbar et al (2013) was Italian with an *f*-measure of 0.488 with the combined feature set listed above. Italian was also the most accurately classified first language group in the study by Daudaravicius (2013), with an *f*-measure of 0.435. The results when using LIWC on its own for Italian were 0.646, a 20% improvement. The most accurately classified first language group in this study was Chinese, with an *f*- measure of 0.783 for the LIWC feature set. The TOFEL 11 corpus has 100 essays in each first language group, the first language groups in this study consisted of 241 essays. The two studies that used the TOFEL 11 corpus cited above had 52 (Abu-Jbar et al (2013)) and 54 Daudaravicius (2013) essays classified correctly. The Chinse essays had 38 and 32 essays classified correctly respectively. As can be seen from Table 4-4, there were 161 essays of a possible 241 Italian essays and 189 of a possible 241 Chinese essays classified correctly by the classifier using the LIWC feature set.

	dn	ru	fr	bg	cn	ts	no	sw	sp	jp	it	tr	cz	ge	fi	po
dn	86	25	20	16	0	0	13	7	15	3	3	11	9	4	16	13
ru	24	83	12	37	2	0	3	5	7	5	3	8	26	5	4	17
fr	17	9	122	11	0	0	5	8	14	3	7	4	9	4	12	16
bg	7	13	4	144	1	2	7	12	6	5	1	12	9	4	6	8
cn	2	3	1	4	189	6	3	1	2	5	1	8	1	0	3	12
ts	2	0	1	1	8	199	7	3	1	5	1	7	1	1	1	3
no	14	5	9	24	1	2	107	18	5	7	1	4	11	8	13	12
sw	11	6	6	17	1	3	30	78	12	9	6	7	9	19	13	14
sp	13	12	17	23	0	1	7	9	117	2	6	11	8	1	4	10
jp	4	7	1	5	4	5	7	13	0	161	1	14	3	8	1	7
it	8	8	5	9	0	0	2	6	11	3	159	4	2	5	4	15
tr	5	5	5	24	3	1	11	5	5	8	4	139	7	1	5	13
cz	6	25	8	19	0	0	7	10	8	7	2	8	114	4	7	16
ge	6	5	11	9	4	2	5	20	10	2	8	4	7	130	2	16
fi	26	15	8	15	2	2	14	23	8	5	9	12	12	16	60	14
po	9	11	12	19	10	2	2	11	6	1	6	8	11	4	10	119

Table 4-4: Confusion matrix for first language classification for 80 LIWC features

The confusion matrix for the LIWC classification is shown in Table 4-4. The correct classifications are indicated in white text with a black background. The largest error in classifying the target language (the X axis) as another language (the Y axis) are indicated in greyed out cells. So the highest number of Dutch (du) essays incorrectly classified as one other language were the 25 classified as Russian (ru). The biggest error in classification was for Bulgarian (bu) being classified as Russian). As would be expected, the first language groups that are the most accurately classified show the most uneven spread in results, where the poorly classified first language groups have the most even spread. The exception to this was Polish (po) which had 119 essays (49.38%) classified correctly, but the incorrectly classified essays displayed a very even spread between 12 of the other 15 first language groups. Bulgarian (bu) was the most common incorrect classification, while Tswana and Chinese were the least common. This was common across all the confusion matrices for this experiment.

The next section will explore the effect of increasing the number of features per feature set for the function words, character bigrams and POS bigrams.

4.1.2. Increasing Numbers of Features

The 80 feature limit in the previous section was to compare approximately equal numbers of features with the LIWC feature set which has a fixed size of 80 features. Three of the feature sets (the function words, character bigrams and POS bigrams) have many more features available. Many of the previous authorship attribution studies (Argamon-Engelson et al., 1998; Li et al., 2006) have used feature sets containing far more than 80 features each. The same would be expected to apply to authorship characterisation, although this has not been previously tested. To examine the effect of larger feature sets, the numbers of the three larger feature sets were increased to 200 features per set and then in increments of 200 until there were 1000 features in each feature set.

The results are shown in Table 4-4. All three feature types showed a marked increase between 80 features and 200 features, between 8.8% and 9.8%. POS bigrams increased significantly from 400 to 1000 features ($p = 0.0254$) but there was no significant increase between 600 and 1000 features ($p = 0.1002$). Character bigrams showed a significant increase between 400 and 1000 features ($p = 0.0422$) but, again, no significant increase between 600 and 1000 features ($p = 0.4696$). While function words did show a significant increase between 600 and 1000 features ($p < 0.0180$), this could have been due to the increasing number of content words that were appearing in the feature set. The corpus consist of essays written to set topics, and the accuracy could have more to do with identifying the topic the teacher of the particular first language group set rather than any idiosyncrasies associated with language transfer to English from the first language of the students. Some of the more content oriented words appearing in the word list after 600 words include “financial”, “industrial”, “economic”, “unemployment”, “crime”, “punishment”, “jail”, “student”, “school”, “education”, “birth”, “baby”, “child” and “mother”.

Feature Set Type	Number of Features per Feature Set					
	80	200	400	600	800	1000
function words	43.35	52.06	58.05	61.98	65.12	67.19
character bigrams	47.00	55.91	57.94	60.58	60.14	60.72
POS bigrams	41.01	50.82	54.09	54.71	56.81	57.55

Table 4-5: Accuracy percentage of increasing numbers of the same feature type for first language classification

4.1.3. Combining Different Types of Features

Previous studies have shown that a mixture of feature types tend to be more effective in authorship identification than a large number of the same type (Argamon-Engelson et al., 1998; Li et al., 2006). The same would be expected to apply to authorship characterisation, although this has not been previously tested. One of the aims of this research is to discover if a psycholinguistically based feature set is sufficiently different to the lexically based feature sets that combining them will produce a better result than the same number of the lexically based features alone. To examine this, the accuracy of 280 of each of the larger features sets (function words, character bigrams and POS bigrams) was compared with combined features sets of LIWC added to 200 of the base feature sets. As can be seen in Figure 4-4, the inclusion of LIWC increased the accuracy of each of the base feature types by a greater amount than adding 80 more features of the same type. The increases were between 2.8% and 7.5% greater when LIWC was added. All these increases were statistically significant with *p* < 0.0394.

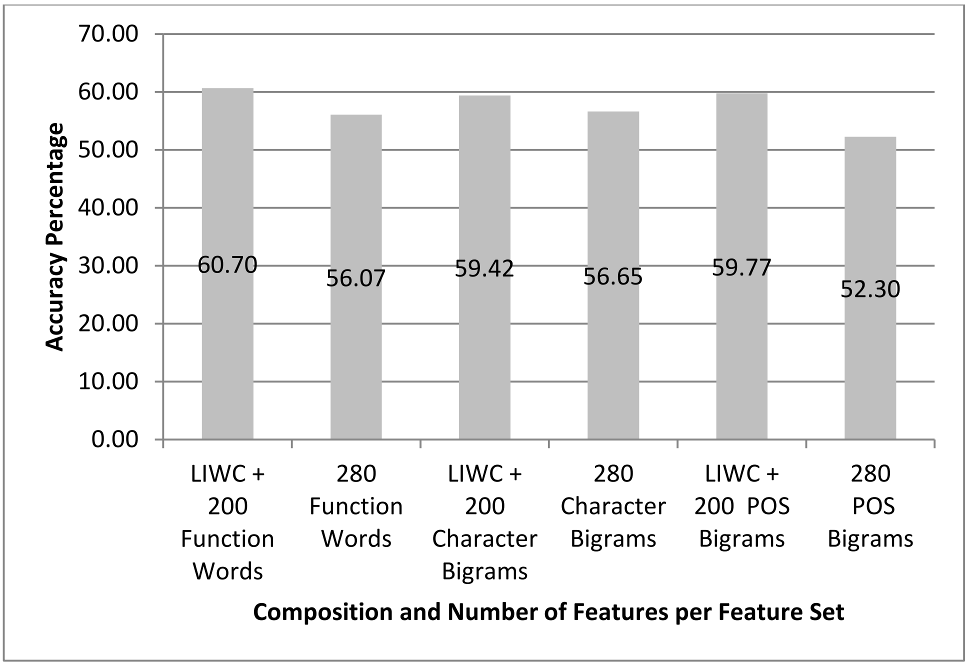

Figure 4-2: Effect on accuracy for the first language classification of adding LIWC to 200 lexicographic features

Table 4-5 compares the true positive rates for the feature sets consisting of 280 lexicographic features with the feature sets consisting of a combination of LIWC and 200 lexicographic features. The highest true positive for each feature set within each first language group is bolded and highlighted, the lowest true positive rate is bolded and crosshatched. Adding the LIWC features to the 200 lexicographic features improved the true positive rate in the majority of cases. The Russian and Japanese first language groups both had a higher true positive rate for the function words only feature set than for the combined LIWC function word feature set. POS bigrams only had a higher true positive rate than the LIWC combined with POS bigrams for the French first language group.

First language group	Number and type of features in feature sets					
	280 LIWC word	280 word	280 LIWC char	280 char	280 LIWC POS	280 POS
dn	**0.488**	0.438	**0.425**	**0.425**	0.469	**0.425**
ru	0.59	**0.602**	0.571	0.447	0.484	**0.466**
fr	**0.634**	**0.559**	0.596	**0.559**	0.565	0.571
bg	0.681	0.575	**0.7**	0.656	0.594	**0.475**
cn	**0.851**	0.845	0.845	0.839	**0.851**	**0.832**
ts	0.85	0.738	0.85	0.806	**0.869**	**0.725**
no	0.556	0.519	**0.581**	0.519	0.569	**0.444**
sw	0.348	0.36	0.348	**0.41**	0.36	**0.248**
sp	0.613	0.6	0.581	**0.55**	**0.638**	0.569
jp	**0.708**	0.745	**0.77**	0.745	0.739	0.714
it	0.64	**0.522**	0.689	0.596	**0.696**	0.609
tr	0.652	**0.621**	0.702	0.696	**0.714**	0.683
cz	**0.609**	0.522	0.565	**0.484**	0.615	0.491
ge	**0.5**	0.469	0.444	0.431	0.475	**0.313**
fi	**0.404**	0.36	**0.298**	0.36	0.354	0.335
po	**0.59**	0.497	0.54	0.54	0.571	**0.466**
Avg.	**0.607**	0.561	0.594	0.567	0.598	**0.523**

Table 4-6: True positive results for first language classification for adding 80 LIWC to 200 lexicographical features – highest and lowest for each first language group highlighted

The highest number of first language groups that had a higher true positive rate in the lexicographic feature set compared to the feature set combined with LIWC was the character bigram feature set where four languages that either had the same or a higher true positive rate for the single type feature set. These were the Dutch and Finnish language groups (the same true positive rate) and the Swedish and Finnish first language groups.

Function words combined with LIWC gave the highest true positive rate across all six different feature sets for seven of the sixteen first language groups. Character bigrams combined with LIWC had the highest true positive rate in three first language groups and five of the first language groups had higher true positive rates in the POS bigrams-LIWC combination feature sets. Nine of the first language groups had the lowest true positive rate in the POS bigram only feature set, four had the lowest in the character bigram only feature set and three first language groups had the lowest in the function word only feature set. There were also two languages that had the lowest true positive rate in feature sets combined with LIWC. The Dutch first language group had equally low true positive rates in character bigrams, character bigrams combined with LIWC and POS bigrams. The Japanese language group had the lowest true positive rate in the feature set that consisted of function words combined with LIWC.

First language group	Number and type of features in feature sets					
	280 LIWC word	280 word	280 LIWC char	280 char	280 LIWC POS	280 POS
dn	**0.488**	**0.438**	**0.425**	**0.425**	**0.469**	**0.425**
ru	0.59	0.602	0.571	0.447	0.484	0.466
fr	0.634	0.559	0.596	0.559	0.565	0.571
bg	**0.681**	0.575	0.7	0.656	0.594	0.475
cn	**0.851**	**0.845**	**0.845**	**0.839**	**0.851**	**0.832**
ts	**0.85**	**0.738**	**0.85**	**0.806**	**0.869**	**0.725**
no	0.556	0.519	0.581	0.519	0.569	0.444
sw	**0.348**	**0.36**	**0.348**	**0.41**	**0.36**	**0.248**
sp	0.613	0.6	0.581	0.55	0.638	0.569
jp	**0.708**	**0.745**	**0.77**	**0.745**	**0.739**	**0.714**
it	0.64	0.522	0.689	0.596	0.696	0.609
tr	0.652	**0.621**	**0.702**	**0.696**	**0.714**	**0.683**
cz	0.609	0.522	0.565	0.484	0.615	0.491
ge	**0.5**	**0.469**	**0.444**	**0.431**	**0.475**	**0.313**
fi	**0.404**	**0.36**	**0.298**	**0.36**	**0.354**	**0.335**
po	0.59	0.497	0.54	0.54	0.571	0.466
Avg.	0.607	0.561	0.594	0.567	0.598	0.523

Table 4-7: True positive results for first language classification for adding 80 LIWC to 200 lexicographical features – highest and lowest for each feature set highlighted

Table 4-6 shows the same data as Table 4-5 but with the highest and lowest true positive rates for the first language groups within each feature set indicated. Again the highest true positive rates are bolded and highlighted, the lowest are bolded and crosshatched. The least accurately classified first language groups are consistent across all feature combinations: Dutch, Swedish, German and Finnish. Three of the most accurately classified languages are also consistent across all feature sets: Chinese, Tswana and Japanese, however, while Bulgarian was in the top four for the combined LIWC-function word feature set, the other five feature combinations had Turkish included in the top four.

When the accuracy figures for the 1000 feature single type feature sets (Table 4-4) are compared with the accuracy figures for the feature sets consisting of LIWC and 200 features from the same feature sets (Table 4-6), the effect of adding LIWC can be seen clearly. The most striking effect is in the case of POS bigrams where the 280 LIWC-POS bigram feature set‘s accuracy of 59.8% is significantly increased over the 1000 POS bigram feature set accuracy of 57.5% (*p* = 0.0481). The 1000 character bigram feature set gives an accuracy that is not significantly different from the 280 LIWC-character bigram feature set (*p* = 0.2149), a feature set that is only 28% of the size. Although the 1000 function word feature set does give significantly higher accuracy than the 280 LIWC-function word feature set, as discussed earlier, the higher function word results could be due to increasing numbers of content words included in the feature set that skew the results.

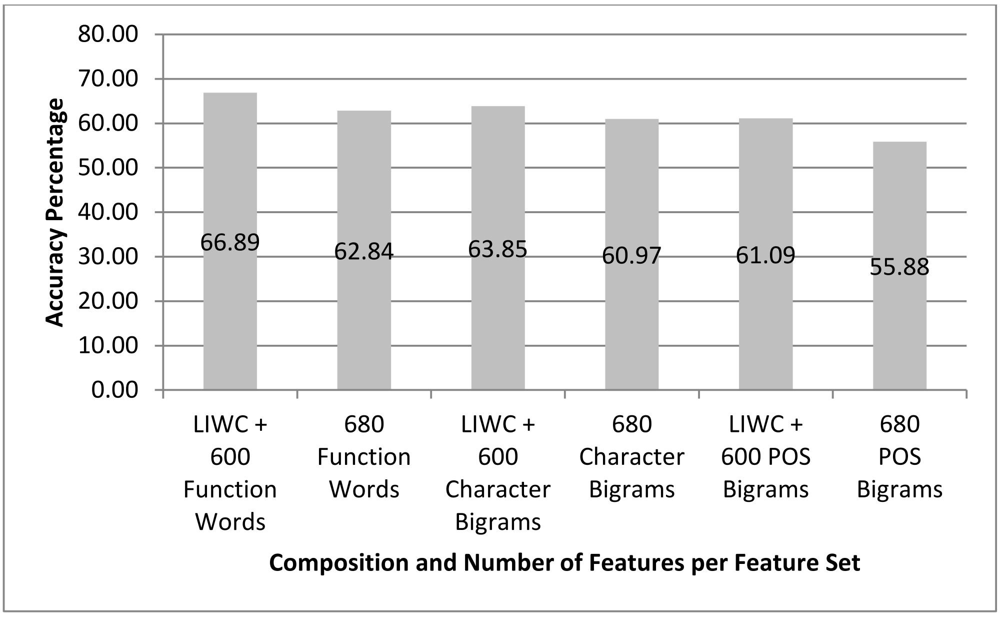

Figure 4-3: Effect of on accuracy for the first language classification of adding LIWC to 600 lexicographic features

Table 4-4 shows that the accuracy of the single type feature sets increases as the number of features increase up to approximately 600 features per feature set. LIWC increased the accuracy of a combination feature set when it was added to 200 features. To see if LIWC could still add a significant amount of insight to a larger feature set, it was added to the 600 sized feature sets and the accuracy was measured against 680 features of the same type. The results are shown in Figure 4-3. LIWC significantly increased the accuracy even when added to a much larger feature set, with p values of 0.0445 (character bigrams), 0.0112 (function words) and 0.003 (POS bigrams.

The final experiment with the full sized documents was to combine the five feature sets used for the first experiments (shown in Figure 4-1) to create a feature set consisting of 390 features (the full 80 features available in the LIWC feature set, the top 80 from each of the function words, character bigrams and POS bigram and the full 70 features present in the POS unigram feature set).

Feature Types Included (✓) or Omitted (×)					Accuracy
LIWC	POS Unigrams	Function Words	Character Bigrams	POS Bigrams	
✗	✓	✓	✓	✓	62.28
✓	✗	✓	✓	✓	64.23
✓	✓	✗	✓	✓	63.56
✓	✓	✓	✗	✓	62.28
✓	✓	✓	✓	✗	62.91
✓	✓	✓	✓	✓	64.97

Table 4-8: Comparative accuracy of the combination feature sets for language classification – feature sets included indicated by ticks.

Each feature set was then removed from the five-way combination, leaving five feature sets with different combinations of four of the five feature types, each consisting of 310 or 320 features, depending on whether the POS unigram feature set was one of the four included. The effect of removing each of the feature sets from the five way combination can be seen in Table 4-7 .

The 390 five-way combination feature set was the most accurate combination, but only marginally more accurate than the four-way combination that excluded the POS unigram feature set. The most profound negative effect on accuracy was caused by removing any one of the LIWC, character bigram or POS bigram feature sets. There was no significant difference between the three most accurate feature sets which all achieved greater than 63% accuracy, or between the three least accurate feature sets which were all less than 63% accurate. There was however, a significant difference between the two most accurate feature sets (greater than 64%) and the three least accurate combinations ($p < 0.0322$).

Number and Type of Features	Accuracy
1000 function words	67.19
LIWC + 600 function words	66.89
800 function words	65.12
LIWC + POS unigrams + function words + character bigram + POS bigram (390 features)	64.97
LIWC + function words + character bigram + POS bigrams (320 features)	64.23
LIWC + 600 + character bigram (680 features)	63.85
LIWC+ POS unigrams + character bigram + POS bigrams (310 features)	63.57
LIWC + POS unigrams + function words + character bigram (310 features)	62.91
680 function word	62.84
POS unigrams + function words+ character bigrams + POS bigrams (310 features)	62.28
LIWC + POS unigrams + function words + POS bigrams (310 features)	62.28
LIWC + 600 POS bigrams (680 features)	61.09
680 character bigrams	60.97
1000 character bigrams	60.72
1000 POS bigrams	57.55
680 POS bigrams	55.88

Table 4-9: Ranking of feature sets consisting of 310 features or more – most accurate to least accurate.

Table 4-8 summarises the accuracy of all feature sets consisting of more than 300 features ranked in order of accuracy. The three most accurate feature sets all include large numbers of function words. As discussed previously, as the number of function word feature expanded, the number of content specific words increased. These words could have identified the topic written about by the first language group cohort rather than any idiosyncrasies of the first language of the students. These results could be disregarded for this reason. The inclusion of LIWC or the character bigram feature set had a greater impact on accuracy than the inclusion POS bigrams or function words. Of the sixteen feature sets listed in Table 4-8, six of the eight most accurate, and none of the four least accurate contain LIWC. Of the eight least accurate feature sets, only two contain LIWC, one in combination with POS bigrams and one of the “leave one out” feature sets that has the character bigrams omitted. If all content words are omitted, the more powerful features would appear to be LIWC followed by the character bigram feature set.

4.2. Results for Shortened Feature Sets with Full Sized Documents

This section will examine the effect of reducing the number of features per feature set. It is hypothesised that the accuracy will fall as the number of features decreases, but that LIWC will decrease less than the lexicographic features. Each LIWC feature is an aggregate of a number of words, unlike character bigrams or function words. POS tags also cover a number of words with each feature, but all of the words covered by each feature are of the same part of speech. The features in LIWC are not necessarily limited in this way either. This could be the reason that it appears to be more powerful than larger feature sets, and that each of its features adds more insight than the lexicographic features.

One thing that was observed, when the large feature sets were ranked using the algorithms from Section 3.4 was that although the different ranking methods gave appreciably different results, some of the same features were in the top ranks. This would imply that not all the features within the top n of any given feature sets are equal. Witten and Frank (2005) suggest one method of sorting features to find the most effective is using a decision tree. The reasoning being that the first feature in the tree is the one that the software uses to split the data most effectively into two classes, and so on down the list of features. The WEKA program also has a number of feature selection functions. Three feature selection algorithms were tested: the J48 tree, the Information Gain and Gain Ratio functions. This gave four lists with different rankings for each of the five feature types used: the original ranking, used for the first part of this Chapter, information gain (IG), gain ratio (GR) and the J48 tree (J48).

The top 80 features from the function words, character bigrams and POS bigrams were used in this section, along with the entire 80 features from LIWC and the full 70 features present from the POS unigrams. These were then ranked using the J48 tree, the information gain and the gain ratio supplied by WEKA (Witten & Frank, 2005). The chi squared method was not included because it gave exactly the same ranking as the information gain in preliminary testing. The top 40 and top 20 features selected using each algorithm, for each feature type were then compared for accuracy. The results are shown in Figure 4-4. As would be expected, there was a significant drop in accuracy for each feature type when the number of features in the feature set was dropped from 70 (POS unigrams) or 80 features with *p* ranging from 0.0001 (function words) to 0.0295 (POS unigrams). The falls were not consistent across each ranking paradigm within each feature type, and the *p* values given are for the comparison of the ranking algorithm that resulted in the highest accuracy for each feature type. The most effective feature ranks were given by the information gain algorithm for the LIWC, function word and character bigram feature sets while the most effective algorithm for the POS unigrams and bigrams was the gain ratio. This was the case for both the 40 feature sets and the 20 feature sets. As can be seen from Figure 4-4, LIWC was again the most effective feature set for sets of 40 features and 20 features.

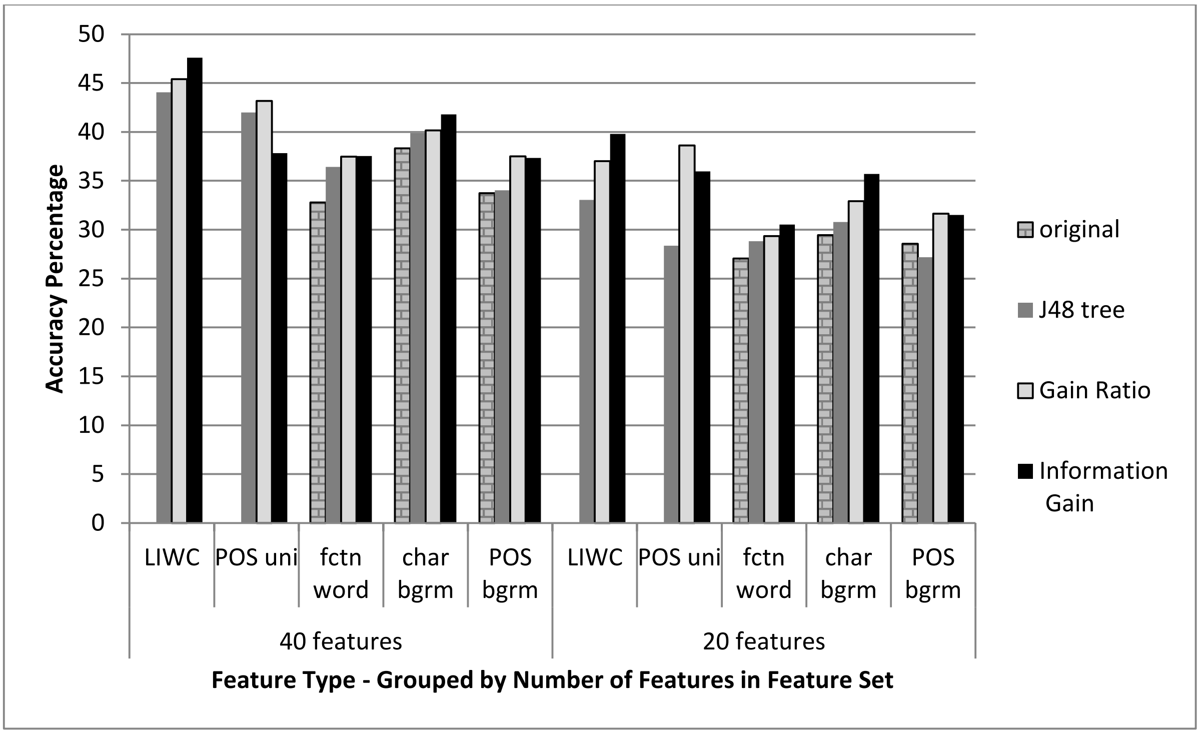

Figure 4-4: Comparison of accuracy for first language classification of top 40 and 20 single type feature sets when ranked by original methods, J48 tree, information gain and gain ratio algorithms.

The experiments in Section 4.1.3 showed that a combination of features were often more effective than larger numbers of single type feature sets. To ascertain if this was the case with reduced feature sets, the 390 feature combined feature set consisting of 70 POS unigrams and 80 each of the LIWC, function word, character bigram and POS bigram feature types were ranked using the same three algorithms: information gain, gain ratio and the J48 three. The top 80, 40, and 20 features from each listing were compared for accuracy. Figure 4-5 shows the results for the combined features set compared to the single type feature sets.

As would be expected, the combination feature set was more effective than any of the single type feature sets at 80 features achieving between 6.79% (LIWC) and 16.09% (POS bigrams) higher accuracy, which was significantly greater ($p < 0.001$). However, as the number of features in the feature sets fell, LIWC became more competitive with the combination feature set. At 40 features in each feature set, LIWC was only 1.92% less accurate than the combination feature set, and while this was still statistically significant ($p = 0.041$) it was not substantial. At 20 features LIWC was only 0.44% less than the combination feature set and not significantly different ($p = 0.3278$). These figures are for the information gain ranking which was significantly more accurate than the other ranking algorithms ($p < 0.039$).

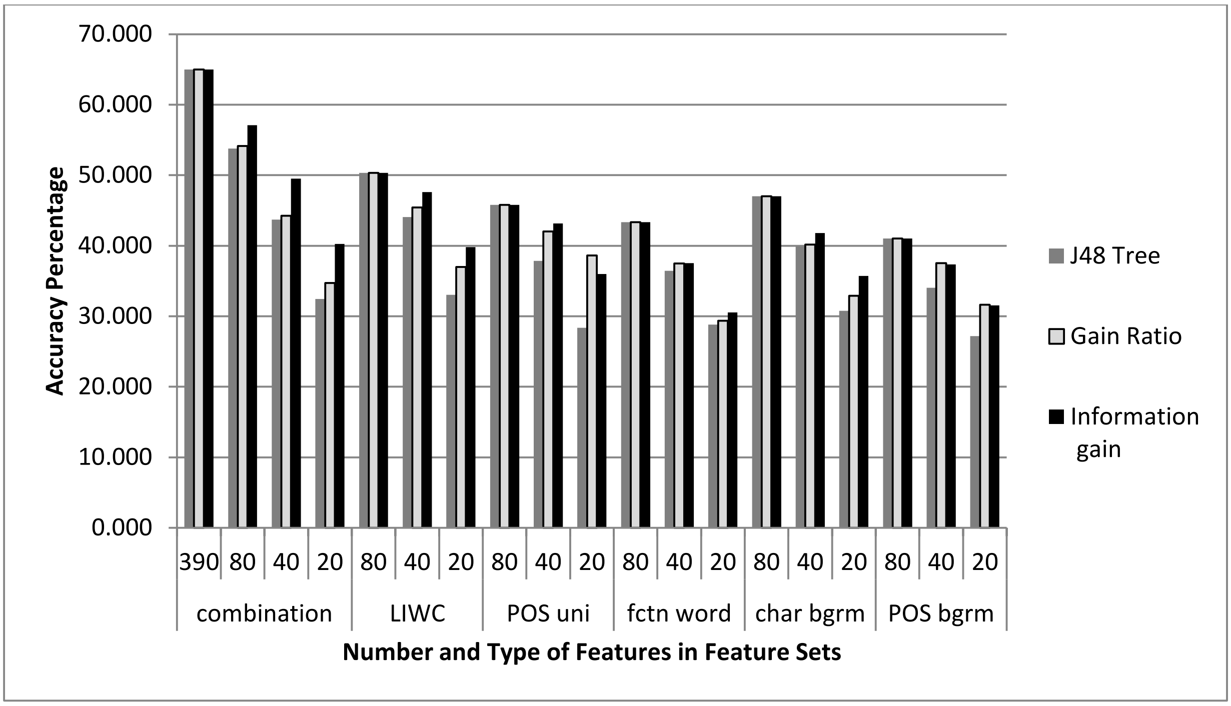

Figure 4-5: Comparison of accuracy for first language classification of top 80, 40 and 20 single and combined feature set when ranked by the J48 tree, gain ratio and information gain algorithms

Classifiers using a combination of features were more accurate than those using the same number of single type lexicographic features at all feature set sizes, and more successful than LIWC at 80 and 40 features. However, this was only the case when the features were ranked using the information gain algorithm. LIWC was marginally, although not statistically significantly, more accurate at 40 and 20 features when the features were ranked using the other two algorithms. The lexicographic features were not more accurate when ranked using any of the three ranking algorithms. The different rankings of the combination feature set were comprised of different numbers of feature types. Table 4-9 shows the numbers of each of the feature type in the combination feature sets under each ranking algorithm.

Ranking Method	Feature Set Size	Feature Type				
		LIWC	POS Unigrams	Function Words	Character Bigrams	POS Bigrams
Information Gain	80	29	15	9	14	13
	40	15	7	3	9	6
	20	10	5	1	3	1
Gain Ratio	80	25	12	8	23	12
	40	12	6	4	13	5
	20	6	2	1	8	3
J48 Tree	80	20	18	16	13	13
	40	13	8	9	7	3
	20	7	4	2	5	2

Table 4-10: Analysis of the number of each feature type in the combination feature sets at different sizes and under different ranking algorithms – first language classification

If there were no difference in the discriminatory power of the five different feature types, it would be expected that the feature types would have approximately equal representation in the combination feature sets, however this is not the case. Chi squared goodness-of-fit tests were conducted on each features set size/ranking algorithm combination. The calculations are shown in Table 4-10.

Ranking Algorithm	Degrees of freedom	n	χ^2	p
Information gain	4	80	14.50	0.0059
	4	40	10.00	0.0404
	4	20	14.00	0.0073
Gain ratio	4	80	14.13	0.0069
	4	40	8.75	0.0677
	4	20	8.50	0.0749
J48 Tree	4	80	2.38	0.6672
	4	40	6.50	0.1648
	4	20	4.50	0.3425

Table 4-11: χ^2 test results showing the statistical significance of imbalance in feature types represented in the reduced feature sets.

For all feature sets sizes in feature sets ranked using the information gain and the gain ratio algorithms, the difference in feature type representation was significant, however the feature sets that were ranked using the J48 tree showed no significant difference in the numbers of each feature type present. The J48 rankings were also the least successful at all feature set sizes. The feature type that consistently has a higher representation in the information gain and gain ratio ranked feature sets is LIWC with more than 30% of the features in these two rankings at all feature set sizes. Chance would be 20%.

The information gain and gain ratio both have disproportionately high numbers of the LIWC feature set in all three feature set sizes, however, the information gain ranking was significantly more accurate in each case (p > 0.03). Chi squared goodness-of-fit test indicated that there was significantly higher proportions of LIWC in the information gain ranked feature sets compared to the gain ratio ranked feature sets (χ^2 (1, n = 80) = 9.9071, p < 0.0001; χ^2 (1, n = 40) = 5.2341, p = 0.0221; χ^2 (1, n = 20) = 4.317, p = 0.0377). Chi squared goodness-of-fit tests conducted to measure the difference between in proportions of the lexicographic features in the combination feature sets indicated that there was no significant differences in the proportions present. This shows that the more successful feature sets had higher proportions of LIWC features present, but the same levels of the lexicographic features.

4.3. Results for Full Sized Feature Sets with Reduced Document Sizes

Computer mediated communication is often carried out in short bursts whereas the classification exercises that have been undertaken up to this point in this study have been on larger documents. Given the absence of suitably labelled corpora of short documents the purposes of this portion of the study it was necessary to construct documents of shortened length. To create these texts, each essay of the corpus was divided into chunks of eight words, and then a random selection of the chunks was made to make up the required document length. Full details of this process are given in Section 3.7.

The experiments undertaken in this section will use single type feature sets consisting of the 80 top features for the function word, character bigram and POS bigram feature sets and the complete feature sets available for the LIWC (80 features) and POS unigrams (70 features) as well as a combination feature set consisting of all the combination of all five of the single type feature sets. The feature sets consisting of various combinations of four feature types did not add any real insight in the full sized document classification exercises discussed in section 4.1.3, and so were omitted from this series of experiments. The documents used include the full sized documents, a set of 500 word documents, sets with the size decreasing in 100 word increments and finally 50 and 24 word documents. As can be seen from Table 4-1, there is a wide range in the minimum essay size across the first language groups, from 500 words (a Turkish essay) down to 180 words (a German essay). However, with the exception of the Tswana language group, the average is well above the 500 word limit, so most of the essays will be affected in the first reduction to 500 words, increasing with each decrement until all essays are affected.

4.3.1. Single Type Feature Sets with Documents Reducing to 24 Words.

As can be seen in Figure 4-6, the reduction in document size had a substantial and uniform effect on the accuracy of all the single feature type classifiers, which became more pronounced as the document size declined, although this may be due to the number of essays affected by the reduction. The average drop in accuracy across all five feature types from full text to 500 word texts was 3.4%, but the decrease in accuracy when the document size fell from 100 words to 50 words was 6.7%, the largest incremental fall. The decrease in accuracy when the documents were further reduced to 24 words was also substantial, at 5.91%. Of the average 31.8% decrease in accuracy, 23.2% of it occurred in the reductions from 300 words down to 24 words per documents. Even so, the LIWC feature set was significantly more accurate than all the other feature types ($p < 0.0005$) at all document sizes. The results for the smallest document size (24 words) were between 18.1% (LIWC) and 10.0% (POS bigrams). These are quite low results even though they are still above chance which is 6.3%.

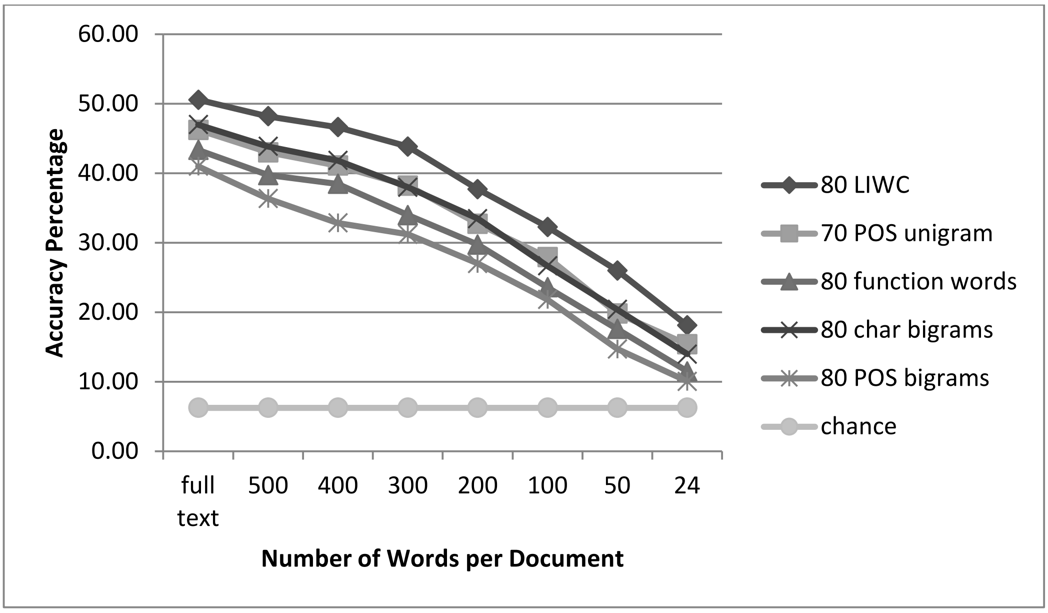

Figure 4-6: Effect on accuracy of single type feature sets as document size falls from full size to 24 words for first language classification

4.3.2. Combination Type Feature Sets with Documents Reducing to 24 Words

The 390 feature combination type feature set consisting of approximately equal numbers of the five feature types was used to classify the eight different sized document corpora. The combination feature set performed better than the single type feature sets on the full sized documents. However as can be seen in Figure 4-7 as the document size reduced, so did the difference in accuracy between the combination feature set and the single type feature sets. When the full sized documents were classified, there was a 14.4% difference in accuracy between the most effective of the single type feature sets (LIWC) and the 390 combination feature set. However the difference reduced at each decrement of the document size until it was only 0.9% for the 24 word documents classification exercise, a difference that is not statistically significant ($p = 0.1383$).

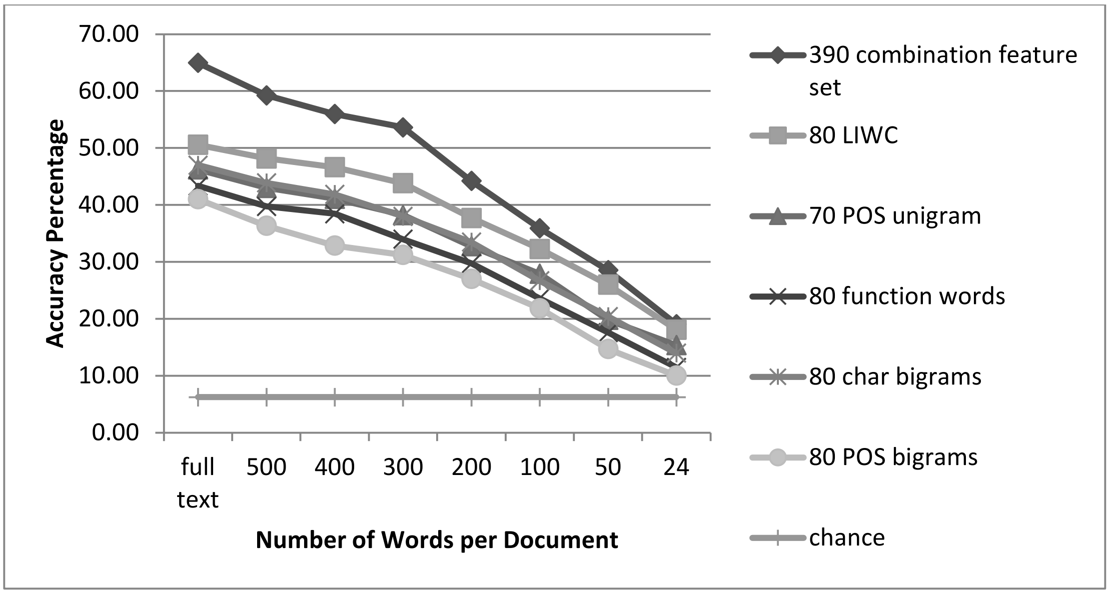

Figure 4-7: Effect on accuracy of first language classification of combination feature set compared to the five single type feature sets as document size falls

4.4. Reduced Feature Set and Reduced Document Size

The “curse of dimensionality” (Chaski, 2008) is the situation where there are more features in the feature set than there are examples of features in the data, which could adversely affect the results. This could be the reason for the observed fall in accuracy of the 390 feature set when the document size was less than 300 words. The average drop in accuracy from full sized documents to the 300 word documents was 3.8% at each decrement in document size. The average fall in accuracy from 300 words to 24 words was 8.6%. To investigate this further, the feature set was incrementally reduced and tested on each of the documents sizes. The information gain ranking was used for the reduction because it proved to be the most effective in the earlier experiments reported in Section 4.2.

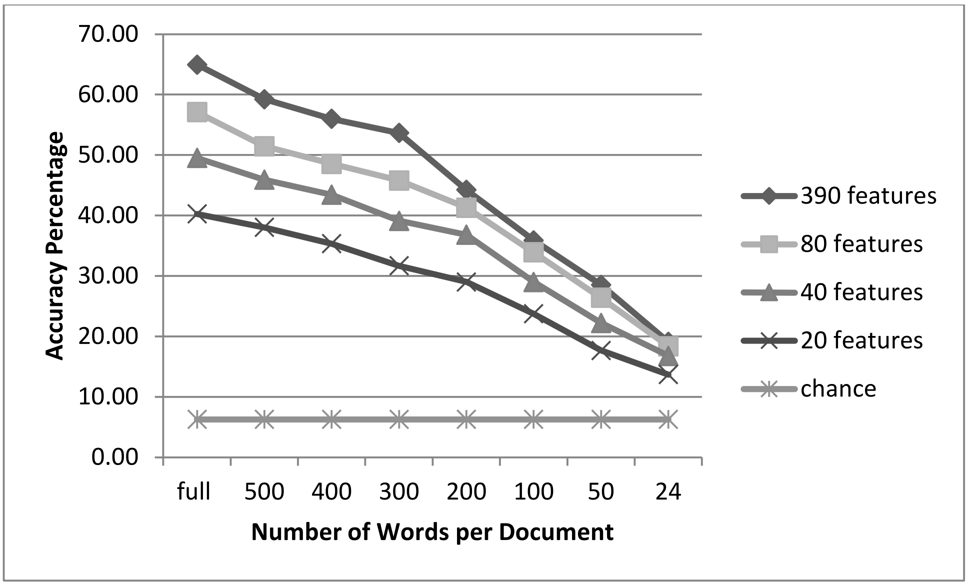

Figure 4-8: Effect on accuracy of first language classification of reducing document size and 390 combination feature set size simultaneously

Figure 4-8 shows the fall in accuracy for each of the four different feature set sizes as the document size falls, compared to chance at 6.25%. As can be seen, the smaller the feature set, the less steep the gradient of the graph. Overall, the feature set consisting of 390 features falls from 64.97% to 19.90% (a drop of 45.90%) while the feature set consisting of 20 features falls from 40.25% to 13.66% (a drop of 25.59%). If the fall in accuracy for each document size over the reducing feature set size is examined, the average accuracy for the full sized documents falls from 64.97% to 40.25% (a drop of 24.72%) but the accuracy for the 24 word documents only drops by 5.41% (from 19.07% down to 13.66%). It would appear that the accuracy for the smaller document is affected less by the reduction in features set size and the accuracy of the smaller feature set is affected less by the reduction in number of words per document. However, the smaller sets are less accurate to start with, so the end result is still that the larger, 390 feature set is more accurate for documents of 24 words, even though it has suffered a larger fall in overall accuracy as the documents have reduced. So in this case reducing the number of features does not remove the "curse of dimensionality" (Chaski, 2008).

4.5. Discussion of First Language Group Classification Exercises

This chapter has presented the results of the application of the LIWC feature set to the problem of first language identification. As LIWC is a tool that has been developed in the psycholinguistic field rather than the computational linguistics field, it was hypothesised that it would be effective, both as a single type feature set because of its linguistic basis, and in combination with other feature sets, because it should be sufficiently diverse from the other feature sets.

As hypothesised, LIWC gave very good results when compared to an equal number of function word, character bigram and POS bigram features. The LIWC feature set gave the highest true positive rate in nine out of the 16 first language groups tested, and the highest average overall accuracy, more than 9% higher than the worst performed feature type, POS bigrams. When the performance of the 80 LIWC features was compared to 200, 400 and 600 of the other three feature types, it performed well. This would indicate that linguistically based features are effective when applied to the problem of first language characterisation.

The character bigram feature set also performed well. When compared with equal numbers of word and POS bigram features, the character bigrams significantly outperformed the POS bigrams at every level, and had significantly better results than the word feature set at 200 features. There was no significant difference between the function words and character bigram feature sets at 400 and 600 features per feature set, but at 800 and 1000 features the function words did achieve a significantly higher accuracy. As discussed previously these results could be skewed by the increasing numbers of words specific to essay topics included in the larger function word feature sets. The results observed for character bigrams supports the hypothesis of Tsur and Rappoport (2007) that character bigrams could indicate a preference for phonemes based on familiarity from a first language. POS bigrams giving the lowest results was unexpected. Common grammatical errors in second language English speakers include omission of words, confusion of tenses and reversing the normal order. POS bigrams would have been expected to indicate this sort of idiosyncrasy. What can be drawn from this result is either that these types of grammatical errors were not present to a large extent in the ICLE corpus, or that a bigram is not long enough to capture them.

As has been indicated in other studies, this study showed that combinations of feature types are better, in most cases, than more features from a single feature type. When the LIWC feature set was combined with 200 and 600 of each of the function word, character bigram and POS bigram feature sets, the combination was significantly more accurate than the same number of the single type feature set alone. That the 80 LIWC features can show a positive effect on the accuracy of a classifier based on as many as 600 other features is an indication of its effectiveness as a feature set for first language characterisation.
To further test the effect of combining LIWC with other features, a combined feature set consisting of similar numbers of all the feature types was tested. The combination consisted

of 80 each of LIWC, function words, character bigrams and POS bigrams, and 70 POS unigrams. Five combinations were produced from this master set, with each feature type removed in turn, giving six feature sets to compare. The removal of either LIWC or character bigrams produced a significant reduction in accuracy, whereas the removal of any of the remaining three feature types did not significantly reduce accuracy. There was no significant difference between the combination feature set that contained all five feature types and the combination that contained the LIWC, function words, character bigrams and POS bigrams.

These results indicate that the LIWC features are sufficiently different that they improve the classification more than simply adding more lexicographic features. When the classification accuracy of each feature set was compared to the accuracy of the same feature set combined with LIWC, the LIWC combination had an increased accuracy rate.

As both the number of features in the feature sets used in the classifier and the number of words in the documents being classified reduced, the accuracy also fell, however the fall was not proportional to the drop in either variable, indicating that there are features in all of the feature types tested that have greater influence on accuracy than others of the same type. It was also shown that very small documents can be classified on the first language of the author with results which although not exceptional, are significantly better than chance.

While the overall accuracy was affected by the addition and removal of feature types, some aspects of the classification remained consistent, especially the comparative accuracy rank of individual first language groups for any given combination of features. Certain first language groups such as Chinese, Tswana and Japanese were consistently well classified, whereas first language groups such as Dutch, Swedish and Finnish were consistently poorly classified.

There could be several reasons for this. All the documents used in this study are from the ICLE corpus, which is from essays written by university students studying English as a Foreign Language (EFL). There are a number of ways to teach EFL, but the main streams are grammar-translation (grammar first) and communicative (expression first). In the first case, the paradigm is for the student to learn the grammar correctly, and it is expected that expression will flow from the understanding they gain. In the second case, the paradigm is for the students to use English to express themselves, and that the understanding of the grammar will come from use and corrections (Connor, 1996; Warschauer & Kem, 2000). Paradigms as different as these could influence the students' use of English and the impact their first language has on their expression in English. The amount of English used in the various communities and the number of other languages that are commonly used/heard by the students could also affect their use of English. For example, the three first languages that are consistently poorly classified (Dutch, Finnish and Swedish) have high levels of English in the community, are bilingual countries and up until recently have used the grammar translation paradigm, changing to the communicative paradigm in the majority of schools relatively recently (Granger, 2001). Finland also used immersion teaching, where the entire school experience is in English. The three first language groups that are consistently well classified (Chinese, Japanese and Tswana) did not use immersion teaching at the time the corpus was compiled. These three first language groups also do not have a high level of English language pervasive in the community, and unlike the European schools, do not have large numbers of native or fluent English speakers to instruct in English language classes. The Chinese essays come from schools in Hong Kong, where the English in the community is heavily influenced by the native Cantonese (Granger, 2001). These factors could have a large impact on the amount of language transfer, a phenomenon discussed by Tsur and Rappoport (2007) and Wong and Dras (2009). With less influence from the authors' first language, there would be less impact on the features that were used for classification.

The large discrepancy between the first language groups within each feature set indicates that for effective classification of all the first language groups used in this study, a multiclass classifier with one set of features may not be the most effective method, and a collection of binary classifiers, with a feature set tailored to each first language could give better results, especially for the first language groups at the lower end of the classification scale. Such a solution would be similar to the Writeprints system (Abbasi & Chen, 2006) where an author's unique writing style is captured in a visualisation and can then be compared to the visualisation of the style of anonymous documents. This, however raises the issue with the number of pairs that need to be classified. In the case of the 16 first language groups in the ICLE corpus, it would be necessary to test 120 pairwise classifiers. In the case of first language group identification, the visualisation could be of the unique language transfer features that effect the English used by native speakers of a given language, and that pattern could be compared to style visualisation of documents to identify the first language of the author, rather than laboriously comparing languages one on one or attempting to find one set of features that will accurately identify multiple first language groups. However, if a 'one size fits all' classifier were to be used, psycholinguistically based features, such as LIWC appear to be more effective than a similar number of lexically based features, both as a standalone feature set and when used in combination with the other feature sets.

Chapter 5 Profiling for Gender of Author

Do males and females use language differently? The most significant hormonal difference between males and females, testosterone, has been linked to aggression, negative moods, improved spatial skills, decreased verbal ability, concerns over status and dominance, and more direct thought and action (Pennebaker et al., 2004). Studies have also shown that many of these factors also impact on both written and spoken language (Newman et al., 2003; Pennebaker et al., 2003; Rude et al., 2004).

The identification of the gender of text based communication is becoming increasingly important. The anonymity of the internet allows less than scrupulous individuals to misrepresent their demographic information, including their gender to gain the trust of unsuspecting internet users and abuse this trust in fraudulent activities. Particularly concerning is research that indicates that internet predators and paedophiles use misrepresentation of gender as a tool when grooming victims (Corney, Anderson, & Mohay, 2002; Dombrowski, LeMasney, Ahia, & Dickson, 2004; Trevathan & Myers, 2012).

There have been a large number of computational linguistics studies attempting to codify the difference in language use between males and females (Argamon et al., 2009; Estival et al., 2007; Peersman et al., 2011; Prasath, 2010) to mention a few. However these studies have all used lexically based features such as counts and ratios of function words, character bigrams and POS n-grams. The results are not comparable to the results in this chapter because the previous studies have used content words and stylistic features as part of the feature sets in the classification. The Blog Authorship Corpus used for these experiments does not contain the original stylistic features from the blogs, so those features cannot be used. The other feature types commonly used are content based features. While there is no doubt that in the main, different genders have different concerns and there for discuss different topics, however if a classifier is to be developed that can distinguish between male and female speech when the topics are constrained, then these features must be eliminated.

This chapter looks at the application of LIWC, a psycholinguistically based feature set, to the problem of gender classification of authors. LIWC is a set of aggregate features that have been created by psycholinguist to assess the mental health of patients by analysing their written and spoken language (Pennebaker et al., 2007). Further discussion of the differences between LIWC and lexically based features is given in Section 2.5.4. Because this is a different paradigm to the commonly used lexicographically based feature sets, it is anticipated that, as in the first language classification detailed in Chapter 4, LIWC will give improved results and add to the accuracy when combined with the lexically based features.

The ICLE corpus used in Chapter 4 is not suitable for gender classification research because it is heavily skewed towards female authors. Over 77% of the corpus is from female authors. Attempting to get a balanced representative of male and female authors would have resulted in a corpus that was too small to accurately classify. Therefore the corpus used for the gender classification experiments detailed in this chapter is a subset of the Blog Authorship Corpus compiled by (Schler et al., 2006). The corpus consists of more than 680,000 posts from over 19,000 individual bloggers gathered from blogger.com in August 2004. The corpus has texts with four classes identified: industry, astrological sign, age group and gender. For the experiments reported in this chapter, all the indicators for classes except the gender markers were removed to avoid creating any bias or noise in the data sets. The gender class had the expected two classes, male and female. The original corpus of over 19,000 texts is too large to handle effectively using the WEKA software, therefore a randomly selected subset of 3996 documents was used. This number was chosen because it is easily divisible into three sections of two classes – three sections for each of the gender groups. It is also approximately the same sized corpus as the ICLE corpus to facilitate any comparisons of results.

The corpus consists of equal numbers of male and female bloggers. However the three age ranges, (teens, twenties and thirties) are heavily skewed towards the younger age groups. A subset of the blog corpus was selected to contain equal numbers of randomly selected files from each of the six age/gender groups (two genders with three age groups each). It is important to have the age groups balance within each gender group to avoid any imbalance within the age groups on the gender classification. The sub corpus was divided into three sections, each containing equal numbers from each of the six age/gender groups. One section was used to perform feature selection and the remaining two were combined and used for testing.

Gender	Number of Files	Total Words	Average File Size	Smallest File	Largest File
Female :	1998	14,549,138	7,281	231	270,177
Male :	1998	14,540,936	7,277	292	420,608
Total	3996	29,090,074	7,279	231	420,608

Table 5-1: Summary of gender files used from the Blog Authorship Corpus.

Table 5-1 gives a breakdown of the corpus with respect to the gender groups. There is a remarkable similarity between the numbers of words for the male and female bloggers. Of the 29 million words, an almost equal number were written by males and females, with only a four words difference in the average. Contrary to anecdotal belief, the longest blog was written by a male, and the shortest was written by a female, both of whom were in the twenties age group.

The Blog Authorship Corpus includes large amounts of non-English text in some files, far more than the ICLE corpus used in the previous chapter. This text was not manually removed because the feature selection methods effectively negate its impact. The function words were selected using the document count methods based on (Grieve, 2007) which ranks the features by the number of documents they appear in. Although there were a large number of files containing non-English words, there were also a large number of languages represented and therefore each foreign word was only in a small sub set of documents and was ranked below the cut-off point for feature selection. The character bigram selection omitted any character that was outside the standard ASCII character set. It therefore skipped many of the foreign words. The feature selection method for the character bigrams was also document count so, as for the function words, any unusual character bigrams were ranked below the cut-off point. The POS tagger used in this research, QTag, and the LIWC feature set handles unknown words in a similar fashion in that they group them all together in a single feature, into the “words not in dictionary” feature for LIWC and tagged as “???” in the POS tagger. These words influence only one of the features present, and as such were not considered to have any substantial impact on the classification accuracy.

The same feature types used in the previous chapter will be again be used in the gender classification experiments: function words, character bigrams, POS bigrams, POS unigrams and LIWC. The last two feature types listed, the LIWC and POS unigrams have small feature sets. The entire 80 LIWC features were used in the experiments. Only 70 of the possible 77 POS unigram features were present in the corpus, so only these 70 were used. The first three feature types listed have very large feature sets and so it was necessary to rank them according to their impact on the classification accuracy. The full feature set for each of these three larger feature types were ranked using the three feature ranking algorithms presented in Section 3.4. There was no significant difference in the accuracy of the three ranking algorithms for any of the feature types, so the document count was used because it is computationally simpler to produce.

The lack of statistical difference between the ranking methods for the gender classification where there was significant difference for the first language classification might be explained by the small number of classes being compared (two) in contrast to the larger number of classes for the first language classification exercise (sixteen).

The remainder of this chapter will follow the same pattern as the previous chapter. First the full sized documents will be classified with various sized single type feature sets and combination feature sets. The single type feature set experiments will show the comparative accuracy of LIWC against similar numbers of the lexicographic features. The combination feature set experiments will allow the examination of the impact of combining LIWC with lexicographic features. The feature sets will then be ranked using three feature selection algorithms found in the WEKA (Witten & Frank, 2005) suite of programs and the top n features will be compared for accuracy. The feature sets will be reduced as far as possible while still achieving a reasonable accuracy. Following that the document sized will be reduced to examine the effect of smaller texts on accuracy. Last of all, the feature set size and the document size will be reduced simultaneously and the effect on accuracy measured. It is hypothesised that the LIWC feature set, being based on psycholinguistics rather than computational linguistics, will give greater accuracy than the other four feature types, and when combined with them, will give a greater increase in accuracy than adding an equal number of the same feature type.

5.1. Comparing and Combining Psycholinguistic and Lexicographic Feature Sets on Full Sized Texts for Gender Classification

To establish a base line for the relative effectiveness of the five feature types being examined, the accuracy of 80 features from each of the function word, character bigram and POS bigram feature sets were compared with the 80 LIWC and 70 POS unigram features. The results are shown in Figure 5-1. All the feature types give an accuracy above chance (50%). However the LIWC feature set is significantly more accurate than any of the lexicographical feature sets ($p < 0.001$). The hypothesis that LIWC would give greater insight into the classification of gender because it is based on the psychological basis of word choice rather than ratios of various lexicographical features appears to be supported.

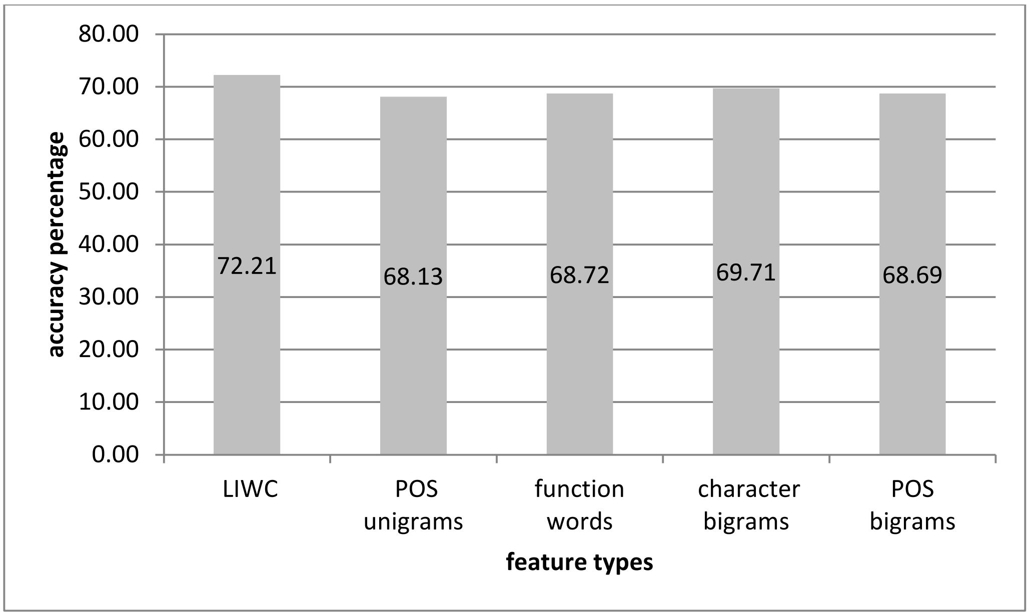

Figure 5-1: Comparison of accuracy for 80 LIWC, 80 function word, 80 character bigram, 80 POS bigram and 70 POS unigram features for gender classification

In the first language classification task, increasing the number of features present in a feature set had a positive effect on the accuracy. To ascertain if this effect would be repeated in the gender classification task, increasing numbers of features in the three larger feature sets (the function words, character bigrams and the POS bigrams) were tested. The results can be seen in Table 5-2. The highest accuracy for each feature type is indicated by bold type and highlighting. The positive effect on accuracy was not repeated in the gender classification task. While there was a significant increase in accuracy from 80 features to 200 features for the function words and character bigrams, there was no significant increase between any of the other feature increments. There was no significant increase between any of the feature increments for the POS bigrams. None of the results were significantly more accurate than the 80 LIWC features.

Feature Type	Number of Features per Features Set					
	80	200	400	600	800	1000
LIWC	72.21					
Function Words	68.72	72.89	72.97	73.39	73.28	**73.48**
Character Bigrams	69.71	72.64	**73.51**	73.00	72.86	72.33
POS Bigrams	68.69	69.65	**70.44**	69.42	69.68	68.78

Table 5-2 : Accuracy percentage of increasing numbers of the same feature type for gender classification

The different pattern of results between the gender and first language classification tasks could be influenced by the different number of classes in the two exercises. The first language classification with sixteen classes could conceivably require a greater number of features to split the classes, whereas the gender classification, with only two classes may not. 200 character bigrams were more effective than the 80 LIWC features, but not significantly so. Many character bigrams are also function words (for example: in, to, of) and some character bigrams relate closely to gender specific topics, such as the character bigram ‘xb’. The only instances of this bigram related to discussion of Xbox gaming, and was mainly found in male blogs. These two factors could explain the success of character bigrams. In Section 4.1.1 different feature sets were more effective in classifying different first language groups, and the average accuracy for each feature type reflected this. For the gender classification, there are only two classes, so the average accuracy is not affected in this way.

Previous studies into authorship attribution have shown that a classifier based on a combination of features is more effective than one based on a single type of feature (de Vel et al., 2001; Koppel, Schler, & Argamon, 2010; Schler et al., 2006; van Halteren, 2004). This was also found to be the case in the first language classification exercises detailed in Chapter 4. Therefore, various combinations of the five feature types were tested to ascertain the effect on accuracy of combining feature types on gender classification. It was hypothesised that the inclusion of LIWC would have a larger effect on the accuracy of the gender classification tasks than the inclusion of additional features from the other feature types. This was based on the results for the first language experiments and the greater accuracy displayed by LIWC when compared to similar numbers of features in the other feature sets.

To examine the effect of adding LIWC to another feature set, the top 200 features from each of the larger feature sets (function words, character bigrams and POS bigrams) were selected and the 80 LIWC features were added to them. Feature sets consisting of the top 280 features of the three larger feature sets were also selected for comparison. The results are shown in Figure 5-2. The inclusion of LIWC increased the accuracy in all cases, significantly so in the case of function words (*p* = 0.0289) and for POS bigrams (*p* = 0.001), but the inclusion of LIWC did not significantly increase the accuracy of the character bigrams (*p* = 0.1787).

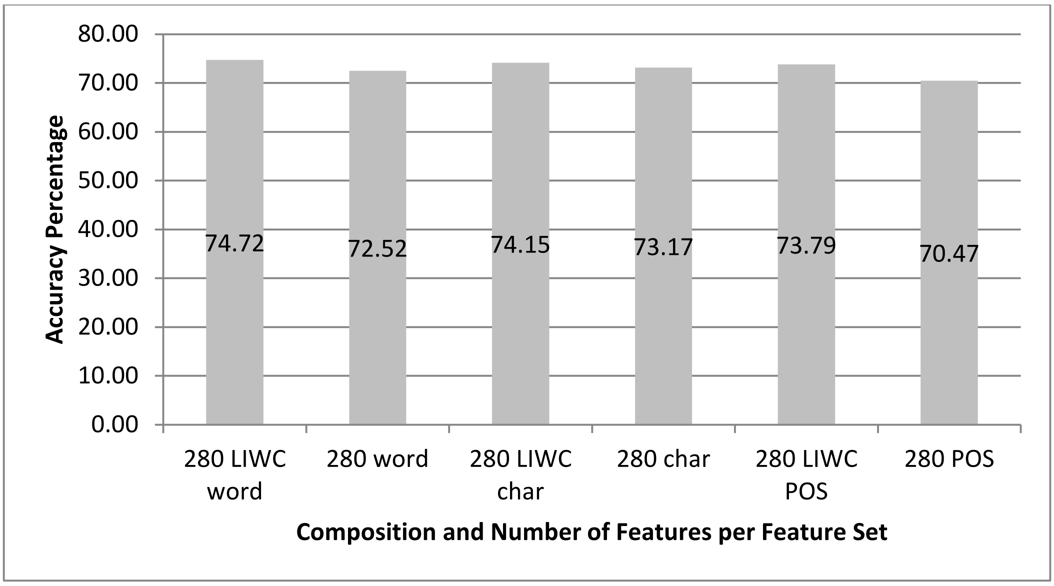

Figure 5-2 : Effect on accuracy of adding LIWC to 200 lexicographic features for gender classification

Adding LIWC to 200 features of the three larger feature types did increase accuracy more than adding 80 more features of the same type. However, previous studies and the results obtained for the first language classification tasks indicated that combinations of feature types, in general, increases accuracy. Therefore, to examine the effect that LIWC has on a combination of feature types, the top 80 features from each of the three larger feature sets, the full 80 features from the LIWC feature set and the 70 POS unigram features present in the corpus were amalgamated to create a combination feature set of 390 features. A series of five feature sets that comprised of four different feature types was created by systematically removing each one of the component feature sets. As can be seen in Table 5-3, the only change to have a significant effect on accuracy was the removal of the LIWC feature set, which significantly reduced the accuracy of the combination feature set (*p* < 0.05). The removal of the other feature types had no significant impact on the accuracy of the combination.

Feature Types Included (✓) or Omitted (✗)					Accuracy
LIWC	POS Unigrams	Function Words	Character Bigrams	POS Bigrams	
✗	✓	✓	✓	✓	72.24
✓	✗	✓	✓	✓	75.23
✓	✓	✗	✓	✓	74.75
✓	✓	✓	✗	✓	74.21
✓	✓	✓	✓	✗	75.03
✓	✓	✓	✓	✓	75.59

Table 5-3: Comparative accuracy of the “Leave One Out” combination feature sets with the five way combination feature set for gender classification

These results show that LIWC being based in psycholinguistics rather than computational linguistics, has a greater impact on the accuracy of the classifier than any of the lexicographically based feature sets, and that the hypothesis, that it gives a greater discriminatory power to the classifier than lexicographic features alone.

5.2. Results for Shortened Features Sets with Full Sized Documents

When the larger feature sets were ranked with the three different algorithms from Section 3.4 it was noted that, even though the accuracy for the top n of each of the feature types under the different rank algorithms was fairly similar, there were very different features in each listing. This would imply that the greatest impact on accuracy is from only a few of the features present. The question is how to find the most effective features within the larger rankings. Witten and Frank (Witten & Frank, 2005) suggest that a useful method to sort features is using a tree to rank features in order of most effective. The reasoning being that the first feature in the tree is the one that the software uses to split the data most effectively into two classes, and so on down the list of features. The WEKA program also has a number of other feature selection functions. Three of these feature selection algorithms were tested: the J48 tree, the information gain and gain ratio functions. This gave four lists with different rankings for each of the five feature types used: the original ranking, information gain, gain ratio and the J48 tree. The feature sets used for the single feature comparisons (results in Figure 5-1) were used as the base feature sets and ranked using these three feature selection algorithms and then the accuracy of the top n features of the resultant four ranked lists, with n decreasing, were compared.

The top 80 features from the function words, character bigrams and POS bigrams were used in this section, along with the entire 80 features from LIWC and the full 70 features present from the POS unigrams From this point on these feature sets will be referred to as the "base" feature sets. These were then ranked using the J48 tree, the information gain and the gain ratio supplied by WEKA (Witten & Frank, 2005). The chi squared method was not included because it gave exactly the same ranking as the information gain in preliminary testing. The accuracy of classifier trained using 50% of the base feature sets were compared. Comparisons were done both between the shortened feature sets created by the different rankings of each feature type and with the full base feature set from each type. The results can be seen in Table 5-4. The highest accuracy for each feature type is highlighted and bolded, the lowest is crosshatched and bolded. There is no value for the original ranking for the LIWC feature set because the original ranking of this feature set is simply the order that the features are produced by the software and this order is unrelated to the effect on gender classification.

Feature Types	Half Base Feature Set (35 or 40 features)				Base Feature Set (70 or 80 Features)
	Gain Ratio	Information Gain	Original Ranking	J48 Tree	
LIWC (40 Features)	70.16	**69.93**		**71.85**	72.21
POS Unigrams (35 Features)	66.50	67.12	**66.27**	**67.37**	68.13
Function Words (40 Features)	**68.16**	**67.96**	68.02	68.02	68.72
Character Bigrams (40 Features)	68.22	**68.44**	68.05	**67.77**	69.71
POS Bigrams (40 Features)	67.09	**67.73**	66.53	**66.10**	68.69

Table 5-4 : Accuracy of top 50% of single type base features sets in order as per the gain ratio, information gain, original and J48 ranking algorithms for gender classification

There was no significant difference between the classifiers based on the top 50% of features in the most accurate ranking and those based on the full feature set for any feature type. In addition, there was also no one ranking method that was the most effective across all five feature types. The J48 tree gave the most efficient ranking for the LIWC and POS unigram features, the information gain ranking was most efficient for the character bigram and POS bigram features, and the function words were most effectively ranked using the gain ratio method. However, unlike the first language classification, there was no significant difference in accuracy between the different ranking methods. This result was very different from those obtained from halving the feature set sizes for the first language classification, where there was a significant drop in accuracy between the 50% and full feature sets.

Feature Types	Ranking Algorithms: 20 Features per Feature Set				Base Feature Set (70 or 80 Features)
	Gain Ratio	Information Gain	Original Ranking	J48 Tree	
LIWC	**68.30**	68.41		**69.48**	72.21
POS Unigrams	**65.31**	**65.71**	65.65	65.40	68.13
Function words	66.75	**67.57**	**65.03**	67.40	68.72
Character Bigrams	65.79	66.86	**65.54**	**67.29**	69.71
POS Bigrams	63.91	**63.76**	**65.90**	64.05	68.69

Table 5-5: Accuracy of top 20 single type features ranked by the gain ratio, information gain, original and J48 ranking algorithms for gender classification

As no statistically significant difference was observed the feature sets were further reduced to 20 features for each feature type/ranking method combination. The results are given in Table 5-5. The highest accuracy for each feature type is highlighted and bolded while the lowest is cross-hatched and bolded. There was again no one method that was the most effective across all feature types. However the LIWC feature sets was the most effective within each ranking. With the number of features reduced to 20 features per feature set, there was a statistically significant reduction in accuracy for all feature type/ ranking method combinations, although, given the size of the feature reduction, it was not a substantial decrease. There was no significant difference in the accuracy for the most and least accurate ranking algorithms within each feature type with the exception of the function word feature type, where the difference of 2.53% between the information gain and original rankings was significant (p = 0.0229). With the largely homogeneous results across the ranking algorithms, it may have been assumed that the features contained within the first 20 features were similar for each ranking. This was not the case. There were between seven and ten features common in the top 20 features across all rankings.

The difference in accuracy between the full base feature sets and the best performed 20-feature feature set for each feature type was a statistically significant but not substantial, reduction in accuracy, especially considering that when the feature sets were reduced to 25% of their original size in the first language classification experiments, accuracy fell by more than 30%. The feature sets were halved again, giving feature sets consisting of 10 features, a reduction of more than 85% of the full sized base feature sets. These were tested for accuracy and the results are in Table 5-6. Even with this small number of features, the average reduction in accuracy between the most effective listing and the full sized base feature sets was only 3.31%. The POS bigrams feature set suffered the most substantial reduction in accuracy of 5.12%. The J48 Tree ranking gave the most accurate results for the LIWC, character bigram and function word feature sets, but the least accurate for the POS unigram feature set. The POS unigram and POS bigram feature sets were ranked most effectively using the Document Count algorithm. The Gain Ratio ranking algorithm had three of the lowest accuracies, those for the LIWC, function word and POS bigram feature sets.

Feature Types	10 features				Base feature set (70 or 80 Features)
	Gain Ratio	Information Gain	Original Ranking	J48 Tree	
LIWC	**68.27**	67.48		**68.92**	72.21
POS Unigrams	65.14	65.34	**65.51**	**65.06**	68.13
Function Words	**64.44**	64.89	64.58	**66.16**	68.72
Character Bigrams	65.71	**65.03**	64.50	**66.72**	69.71
POS Bigrams	**62.86**	63.40	**63.57**	63.20	68.69

Table 5-6: 10 features all feature types ranked with four different method for gender classification. Highest and lowest accuracy indicated by bold highlight and bold crosshatch respectively

For all of the reduced features sets thus far, analysis of variance tests conducted showed that there was no statistical difference between the most and least accurate ranking methods within any feature type. However when the feature sets were further reduced to only five features per feature set, this was not the case for the function word or character bigram feature types.

Feature Types	5 Features				Base Feature Set (70 or 80 Features
	Gain Ratio	Information Gain	Original Ranking	J48 Tree	
LIWC	**66.33**	66.64		**66.69**	72.21
POS Unigrams	**65.06**	64.61	**64.27**	64.55	68.13
Function Words	63.20	63.71	**62.58**	**66.10**	68.72
Character Bigrams	64.67	64.44	**56.64**	**65.88**	69.71
POS Bigrams	**60.75**	61.32	61.91	**63.34**	68.69

Table 5-7: 5 features all feature types ranked with four different methods. Highest and lowest accuracy indicated by bold and highlighted text and bold and crosshatched text respectively

As can be seen in Table 5-7, the original ranking gave very much lower accuracy than any of the other ranking algorithms for these two feature types. The analysis of variance test across the four different ranking algorithms for the function word feature sets gave F(3, 36) = 3.6284, p 0.0219, and across the four ranking algorithms for the character bigram feature sets gave F(3, 36) = 22.0742, $p < 0.0001$.

The greatest drop in accuracy of 13.06% was in original rankings of the character bigram feature set, while the smallest drop in accuracy was in the J48 ranking of the function word feature set. The smallest spread in accuracy was in the LIWC feature set, the largest in the character bigram feature set. When the accuracy of the best ranked feature types were compared, the LIWC feature set was the most accurate, but only significantly more accurate than the POS bigram feature sets (p = 0.0176).

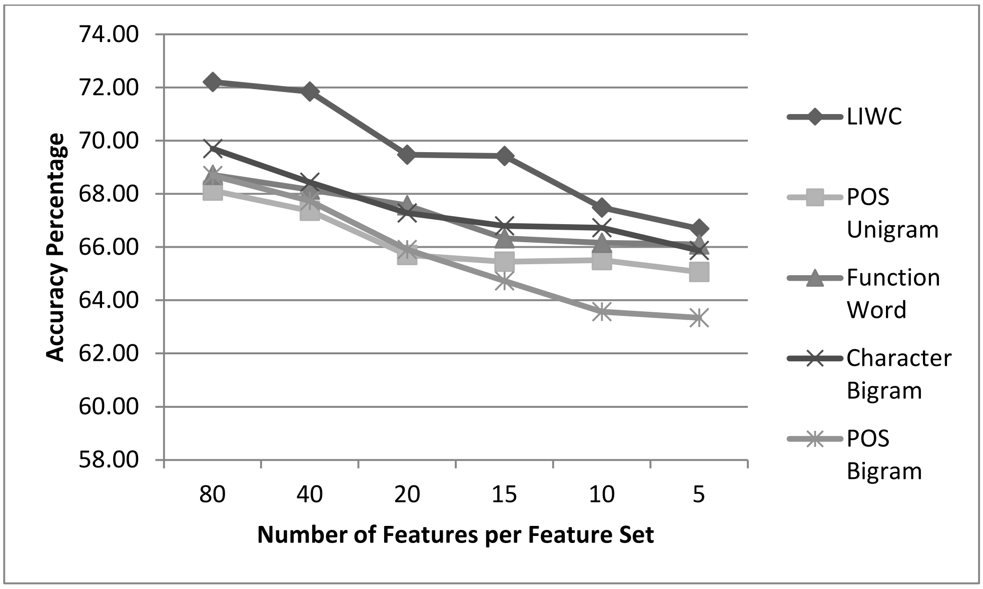

Figure 5-3: Summary of effect on accuracy for gender classification of reducing feature set size from 80 to 5 features

Figure 5-3 summarises the fall in accuracy as the feature sets were reduced. This graph shows the results for the most effective ranking method for each feature type/feature set size combination. It is also worth noting that the x-axis of the graph starts at 58% to highlight the differences in the data. The LIWC feature set was the only feature type that was better ranked by a single ranking method, the J48 tree, at all feature set sizes. LIWC was also consistently more accurate than the other feature types at all feature set sizes, although it also suffered the highest reduction in accuracy at the 5 feature level (5.52%). This would imply that the individual LIWC features have a higher impact on accuracy than the individual features of the other feature types. Each feature in LIWC is an aggregate feature, containing the information for many terms that would be single features in other feature sets. This may give each LIWC feature more discriminatory power than lexicographic features. The POS features also consist of aggregates but they are limited to terms of the same part of speech type, which is not necessarily the case for LIWC.

The comparatively high accuracy using only five features gave rise to the thought that even a single feature may have a high impact on the accuracy of the feature set, especially since the same five features were not present in any of the rankings. Therefore the features that comprised the top five features of each feature set under all ranking algorithms were tested for accuracy on an individual basis.

In the LIWC feature set, there were a total of thirteen features included in the top five features across the three rankings (the original ranking was not applicable to LIWC). The results are shown in the first two columns of Table 5-8. The single most effective feature was that of personal pronouns (I, we, our, etc) which had an accuracy of 64.8% when used individually. LIWC also counts use of first person pronouns in a separate category, which is abbreviated to “i”. This is different to the function word “í” which relates to a single word and not the whole group of first person pronouns (for example “my’, ‘me’, ‘I’, etc) that the LIWC category includes. When the LIWC personal pronoun feature was used individually, it achieved an accuracy of 62.5%, while the pronouns in general feature (including he, she, it, they, etc) gave 62.1% accuracy. The articles feature also gave 61.7% when used individually. Third person pronouns (she/he) also achieved an accuracy of 58.8% when used independently. Family and bio (biological processes including health, body, ingestion and sexual) gave higher than 57%. Of the remaining six features all except the “humans” category were significantly more accurate than chance (*p* < 0.009). Other studies have found that profanity (the category “swear”) gave a good indication of gender in text (Bamman et al., 2014; H. A. Schwartz et al., 2013) however the swear category only gave a result that was 0.9% better than chance in this study, and although this was significantly above chance, it is not substantially so. This could be due to a number of factors including idiosyncrasies of the dictionaries used (the two studies mentioned did not use LIWC) or of the corpus itself. The authors in this corpus may have expressed profanity differently or at a rate that was not high enough to impact on the overall accuracy.

There were ten individual features in the top five features across the four rankings for the POS unigrams, shown in the second set of columns in Table 5-8. The most effective single feature was the determiner feature (DT) followed by the pronoun feature (PP) and the singular noun (NN). There were three punctuation features included: exclamation mark, full stop and the right bracket. The feature “SYM” refers to symbols such as arithmetic symbols, dashes, slashes and back slashes, etc. All of the individual POS unigrams gave accuracy significantly better than chance (*p* < 0.04).

The thirteen top ranked function words that were found in the top five features across the four ranking algorithms are shown in the third set of columns in Table 5-8. In these thirteen there were three determiners (the, a and some), five pronouns (I, me, my, he and them) and four prepositions (of, so, in and to). The pronoun “I” in the function word feature type represents the word “I” only and not the class of first person pronouns as does the LIWC category “i” The remaining feature, the adverb ‘here’ was the only feature that did not give an accuracy significantly greater than chance (p = 0.4435) all the other function word features gave accuracy significantly greater than chance (p < 0.008).

LIWC		POS Unigrams		Function Words		Character Bigrams		POS Bigrams	
Feature	Accuracy	Feature	Accuracy	Feature	Accuracy	Feature	Accuracy	Feature	Accuracy
ppron	64.81	DT	62.08	the	61.15	sp h	61.94	DT NN	60.08
i	62.47	PP	60.75	i	60.89	co	60.47	NN IN	59.88
pronoun	62.13	NN	60.19	of	60.22	sp p	59.91	DT JJ	59.77
article	61.68	NNs	58.33	me	59.74	sp m	59.09	IN DT	59.74
shehe	58.79	IN	57.88	my	59.12	ro	58.08	JJ NN	59.40
family	57.43	exclaim	53.43	so	56.73	s sp	57.77	NNS IN	58.33
bio	57.09	Full stop	52.25	a	56.62	o sp	57.04	IN JJ	57.85
friend	54.17	SYM	51.18	some	56.05	sp c	55.23	PPS NN	55.91
comma	53.29	)	50.79	in	55.74	ay	53.29	PP VBD	55.35
exclaim	53.07	PPX	50.67	he	55.55	sp a	51.55	JJ CC	52.56
anger	51.63			to	53.80	sp d	51.15	NN NN	52.45
swear	50.90			them	50.84	sp b	50.93	NN st	50.37
humans	50.59			here	50.03				

Table 5-8: Summary of accuracy rates of individual features of the top five features from five single type feature sets for gender classification

There were twelve character bigrams in the top five features across the four ranking algorithms, however only two gave more than 60% accuracy when used individually. The character bigrams are shown in the fourth set of columns in Table 5-8. The ‘sp’ refers to a space, so the character bigram ‘sp-h’ indicates a space and then the letter ‘h’. These were a space and ‘h’ (ie words beginning with ‘h’) with 61.9% accuracy and the bigram ‘co’ with 60.47% accuracy. The remainder of the bigrams included words starting with p or m (59% accuracy), the bigram ‘ro’ (58%), and words ending in ‘s’ or ‘o’ (57% accuracy). All the character bigrams in this list gave an accuracy significantly greater than chance when used singly (p < 0.025) except the bigram ‘space b’ (ie words beginning with ‘b’) (p = 0.1249).

There were also twelve POS bigrams included in the top five features across the four ranking algorithms (the furthest right hand columns in Table 5-8). When they were tested individually, three of the top four included DT (determiners): DT_NN (determiner-noun) gave 60.08% accuracy, DT_JJ (determiner-adjective) gave 59.77% and IN_DT (preposition-determiner) gave 59.74%. The second most effective POS bigram when used individually was NN_IN (a preposition and a noun) with an accuracy of 59.88%. All but the least accurate POS bigram gave an accuracy significantly greater than chance (*p* < 0.0006) the least accurate POS bigram (NN_st: a noun followed by a full stop) was not significantly better than chance (*p* = 0.1782). The other parts of speech included in the bigrams listed in Table 5-8 are JJ (adjectives), NNS (plural nouns), PPS (plural pronouns) and CC (conjunctions).

It would appear that the aggregate feature types, such as LIWC and POS unigrams have more effective features than the feature types that rely on single words or phonemes for classification on gender. LIWC outstrips POS unigrams, this is possibly the aggregates are more than simple parts of speech, they have been deliberately compiled to take in the meanings and emotional charge of words as well. POS bigrams were not as useful as POS unigrams. This may be because, although they are also aggregate features, they have too many possible combinations and therefore water down their effectiveness. The features that are the most effective when used singly are pronouns and articles. Previous research has identified these parts of speech as among the more distinguishing features of gender differences in use of language (Newman et al., 2008; Pennebaker et al., 2004).

The reduction in feature set size showed that is it is possible to get an effective classifier with very few features. This would imply that there is a great deal of extraneous/noisy features in the feature sets commonly used and that to strip them down would increase the computational efficiency while not affecting the accuracy a great deal.

5.3. Results for Full Sized Feature Sets with Reduced Document Sizes

The Blog Authorship Corpus contains documents that are a compilation of blogs for each of the more than 19,000 authors. The documents contain, on average 7200 words. However most computer mediated communication is considerably shorter than that. To examine the effectiveness of the feature sets on smaller samples of text, the documents needed to be reduced in size. The method suggested by (Gamon, 2004), of randomly selecting sentences from the larger documents to create a smaller, representative document was not directly applicable to the Blog Authorship Corpus because the punctuation and sentence structure is very casual, and the sentence lengths varied from examples that encompassed an entire paragraph and more, to sentences consisting of only one word. Therefore, the method used in Section 4.3 to reduce the ICLE corpus documents was also used on the Blog Authorship Corpus to obtain shortened documents that remain representative of the authors' styles.

The text for each file was broken up into chunks of eight words and then chunks were randomly selected to make up a document of the desired size. The largest file size was 500 words. The files were reduced in size by 100 word decrements to 100 words and then further reduced to documents consisting of 50 and 24 words. Each document size was created from the original, scrambled document, so that phrases that appeared in one sized document were not necessarily included in documents of different size from the same author. This resulted in seven different sub-corpora, all with files of different sizes, but with the same authors represented in each sup-corpus. There were 234 files in the corpus that had less than 500 words, and a further 14 that had exactly 500 words. For these files, the entire file size was used for the 500 word corpus. There were 64 files with 400 words or less, two of which had less than 300 words, and the smallest file was 231 words. These smaller files had the entire file used for the larger word sizes.

Ten different feature sets were compared for accuracy on each of the sub-corpora. These were the five single type feature sets used for the results displayed in Figure 5-1 and the six combination feature sets used to product the results shown in Table 5-3.

At 500 words per document, 248 of the files had the whole file used and were therefore exactly the same as the files in the full sized corpus (6.2% of the corpus). The results are given in Table 5-9. The accuracy of the five feature combination feature set was 75.6% for the full sized documents. For the 500 word documents, this combination had an accuracy of 70.2%. While this was a statistically significant reduction in accuracy (p = 0.0003), the fall of 5.4% is not substantial when the difference in document size is considered. The least accurate combination feature set was again the one that did not include LIWC with 67.9% accuracy. This was significantly lower than the five way combination (p = 0.0203) and the combination that excluded the POS unigram feature sets (p = 0.0402), but not significantly different than the other combinations. The accuracy the best performed single feature type, LIWC for the full sized documents was 72.2%. For the 500 word documents, the accuracy for LIWC was 68.9%, a drop of 3.5%. This drop was again statistically significant (p = 0.005), but not substantial. LIWC was significantly more accurate than any of the other single type feature sets (p < 0.0003) and not significantly different than the combination feature sets, even though there was only one fifth the number of features in the feature set.

Feature Types Present in Feature Set (500 Words per Document)					Accuracy
LIWC	POS Unigrams	Function Word	Character Bigrams	POS Unigrams	
✓	✓	✓	✓	✓	**70.21**

✗	✓	✓	✓	✓	67.85
✓	✗	✓	✓	✓	69.85
✓	✓	✗	✓	✓	69.17
✓	✓	✓	✗	✓	68.16
✓	✓	✓	✓	✗	69.65
✓	✗	✗	✗	✗	**68.98**
✗	✓	✗	✗	✗	65.00
✗	✗	✓	✗	✗	65.09
✗	✗	✗	✓	✗	65.60
✗	✗	✗	✗	✓	63.77

Table 5-9 : Comparison of accuracy for single type and combination feature sets for 500 words per document for gender classification

For the 400 words per document corpus, there were 43 files that had the entire file used due to small word counts (1.20%). There was no significant difference between the accuracy achieved from the 500 word corpus and that of the 400 word corpus. The results are shown in Table 5-10. Some of the results appear to be higher than those of the larger 500 word documents.

Feature Types Present IN Feature Set (400 Words per Document)					Accuracy
LIWC	POS Unigrams	Function Words	Character Bigrams	POS Bigrams	
✓	✓	✓	✓	✓	70.83
✗	✓	✓	✓	✓	68.21
✓	✗	✓	✓	✓	70.24
✓	✓	✗	✓	✓	70.44
✓	✓	✓	✗	✓	69.28
✓	✓	✓	✓	✗	**71.42**
✓	✗	✗	✗	✗	**69.28**
✗	✓	✗	✗	✗	64.41
✗	✗	✓	✗	✗	63.94
✗	✗	✗	✓	✗	65.79
✗	✗	✗	✗	✓	64.61

Table 5-10: Comparison of accuracy for single type and combination feature sets for 400 words per document for gender classification

Although none of these differences are significant, they may indicate that specific terms or phrases that more easily distinguish gender are present in the smaller documents but not in the larger ones due to the random nature of the document compilation.

At 400 words per document, the most accurate combination feature set was the four way combination consisting of LIWC, POS unigrams, function words and character bigrams. The least accurate feature combination was again the one that excluded LIWC. This combination was significantly lower than any other combination (p < 0.009) with the exception of the combination that excluded the character bigram features. There was no significant difference between the five combination features that included both LIWC and the character bigram feature sets. LIWC was the most accurate single type feature set (p < 0.01) and gave accuracy that was not significantly different from the more accurate combination feature sets. Although there was no significant difference in accuracy between the full text corpus and the 500 words per document corpus, all but two of the feature types showed a marginal increase in accuracy. The only two feature types where accuracy fell were the POS unigrams and the function words.

Features Types Present in Feature Sets (300 Words per Document)					Accuracy
LIWC	POS unigrams	Function word	Character bigrams	POS bigram	
✓	✓	✓	✓	✓	**69.14**
✗	✓	✓	✓	✓	66.16
✓	✗	✓	✓	✓	68.50
✓	✓	✗	✓	✓	68.41
✓	✓	✓	✗	✓	67.62
✓	✓	✓	✓	✗	69.06
✓	✗	✗	✗	✗	**68.86**
✗	✓	✗	✗	✗	63.32
✗	✗	✓	✗	✗	65.03
✗	✗	✗	✓	✗	64.19
✗	✗	✗	✗	✓	63.99

Table 5-11: Comparison of accuracy for single type and combination feature sets for 300 words per document for gender classification

Table 5-11 shows the result for the 300 word corpus. There were only two files that had the entire document used because they contained less than 300 words. The overall pattern is the same as for the 500 and 400 word corpora: the two least accurate combinations are the ones that omit LIWC and character bigrams. They are significantly less accurate than the other four combination feature sets with p< 0.03. LIWC was equally or more accurate than the combination feature sets, although not significantly so either way, but was again significantly more accurate than the other single type feature sets (p < 0.002). The combination feature set that contained all the feature types was significantly less accurate over documents with 300 words than it was over documents of 400 words (p = 0.0395). However the accuracy of LIWC did not fall significantly between these two corpora (p = 0.3569). All feature types with the exception of the function words showed a marginal fall in accuracy of between 2% and 0.5%.

Feature Types Present in Feature Sets (200 Words per Document)					Accuracy
LIWC	POS Unigrams	Function Words	Character Bigrams	POS Bigrams	
✓	✓	✓	✓	✓	65.15
✗	✓	✓	✓	✓	64.11
✓	✗	✓	✓	✓	65.85
✓	✓	✗	✓	✓	65.71
✓	✓	✓	✗	✓	65.77
✓	✓	✓	✓	✗	**66.47**
✓	✗	✗	✗	✗	**66.19**
✗	✓	✗	✗	✗	62.75
✗	✗	✓	✗	✗	63.29
✗	✗	✗	✓	✗	63.01
✗	✗	✗	✗	✓	62.05

Table 5-12: Comparison of accuracy for single type and combination feature sets for 200 words per document for gender classification

All files in the full text corpora contained more than 200 words, so this was the first reduction that affected all files in the corpus. As Table 5-12 shows, although the combination that excluded the POS bigrams was the most accurate combination feature set, there was no significant difference between the accuracy results for the remaining five combination feature sets. There was a significant difference between the most accurate feature set and the least accurate, the one that excluded LIWC ($p = 0.0052$). LIWC was again significantly more accurate than the other single type feature sets ($p < 0.02$), and, while there was no significant difference between LIWC and the five more accurate combination feature sets, it was significantly more accurate than the combination that excluded LIWC ($p = 0.0279$). Both the combination feature sets and LIWC gave significantly lower accuracy over the 200 word corpus compared to the 300 word corpus ($p < 0.01$). The fall in accuracy was greater when the number of words per document was reduced from 300 to 200 words than in the other reductions, with a reduction of 2.2% on average. The combination feature sets suffered the larger fall in accuracy of between 4% (the five way combination) and 1.9% (the combination excluding character bigrams). The single type feature sets also produced lower accuracy by between 2.7% (LIWC) and 0.6% (POS unigrams). Although LIWC did have the largest fall in accuracy, it was still significantly more accurate than the other single type feature sets.

Feature Types Present IN Feature Sets (100 Words per Document					Accuracy
LIWC	POS Unigrams	Function Words	Character Bigrams	POS Bigrams	

✓	✓	✓	✓	✓	61.60
✗	✓	✓	✓	✓	61.51
✓	✗	✓	✓	✓	62.61
✓	✓	✗	✓	✓	**62.95**
✓	✓	✓	✗	✓	62.81
✓	✓	✓	✓	✗	61.96
✓	✗	✗	✗	✗	**63.65**
✗	✓	✗	✗	✗	59.68
✗	✗	✓	✗	✗	59.71
✗	✗	✗	✓	✗	60.87
✗	✗	✗	✗	✓	60.81

Table 5-13: Comparison of accuracy for single type and combination feature sets for 100 words per document for gender classification

When the words per document were reduced to 100, LIWC was still significantly more accurate than any of the other single type feature sets (*p* < 0.004). As can be seen in Table 5-13, it was also significantly more accurate than three of the six combination feature sets: the five way combination (*p* = 0.0498) the combination excluding LIWC (*p* = 0.0454) and the combination excluding POS bigrams (*p* = 0.0431). LIWC was more accurate than the remaining three combination feature sets, but not significantly so.

Feature Types Present in Feature Sets (50 words per Document)					Accuracy
LIWC	POS Unigrams	Function Word	Character Bigrams	POS Bigrams	
✓	✓	✓	✓	✓	60.30
✗	✓	✓	✓	✓	58.78
✓	✗	✓	✓	✓	61.06
✓	✓	✗	✓	✓	61.37
✓	✓	✓	✗	✓	60.81
✓	✓	✓	✓	✗	**61.40**
✓	✗	✗	✗	✗	**61.96**
✗	✓	✗	✗	✗	58.08
✗	✗	✓	✗	✗	59.46
✗	✗	✗	✓	✗	57.71
✗	✗	✗	✗	✓	58.87

Table 5-14: Comparison of accuracy for single type and combination feature sets for 50 words per document for gender classification

The reduction in the number of words from 200 to 100 per document also significantly reduced the accuracy of both the combination and single type feature sets, with a drop of 2.9% on average. The largest fall in accuracy was 4.5% in the combination feature set that excludes the POS bigrams, while the smallest fall in accuracy was the POS bigrams only feature set. LIWC suffered a fall in accuracy of 2.5%.

As the number of words per document was reduced, the accuracy of all of the feature sets, both combined and single type, was also reducing, and reducing by larger amounts for each decrement in text size. This trend did not continue when the document sized was reduced to 50 words (Table 5-14). While the accuracy fell for all feature sets, the fall was not significant for the combined feature types (*p* > 0.1678), nor for the function words or POS unigram feature sets. The fall in accuracy was significant for the other three single type feature sets (*p* < 0.02). However the average fall in accuracy across all the feature sets was only 1.7%. LIWC was the most accurate of all the feature sets, significantly more accurate than all the other single type feature sets (*p* < 0.017) and was also significantly more accurate than the combination that excluded LIWC (*p* = 0.0059). LIWC was more accurate than the remaining five combination feature sets, but not significantly so.

Feature Types Present in Feature Sets (24 Words per Document)					Accuracy
LIWC	POS Unigrams	Function Word	Character Bigrams	POS Bigrams	
✓	✓	✓	✓	✓	59.68
✗	✓	✓	✓	✓	57.74
✓	✗	✓	✓	✓	58.59
✓	✓	✗	✓	✓	58.53
✓	✓	✓	✗	✓	59.88
✓	✓	✓	✓	✗	**60.28**
✓	✗	✗	✗	✗	**58.59**
✗	✓	✗	✗	✗	56.22
✗	✗	✓	✗	✗	57.21
✗	✗	✗	✓	✗	55.52
✗	✗	✗	✗	✓	57.04

Table 5-15: 24 Comparison of accuracy for single type and combination feature sets for 24 words per document

Table 5-15 shows the results for the corpus with the smallest document sizes, 24 words per document. The average fall in accuracy when compared to the 50 words per document corpus was 1.9%. This is slightly more than the difference between the 100 and 50 words per document corpora, but still less than the average fall in accuracy between the 200 and 100 words per document corpora.

Figure 5-4 gives a summary of the effect that reducing the number of words per documents had on each of the 70 and 80 sized five single type feature sets and the 390 feature combination feature set. When the corpus consisted of the full sized documents, the LIWC feature set was significantly less accurate than the 390 combination feature set, although with only a 3.3% reduction in accuracy, the difference was not substantial. LIWC was not significantly different to the much larger 390 combination feature set for any other document sizes. As the document sizes reduced, the difference between the six feature sets shown in Figure 5-4 also reduced, however at each document size, the LIWC feature set was significantly more accurate than any of the other single type feature sets with *p* ranging from 0.0176 (for the corpus consisting of 200 word documents) to 0.001 (for the corpus of 400 word documents). The fall in accuracy for each of the feature types shown in Figure 5-4 appears not to be uniform, with some feature types showing that a smaller document size was more accurately classified than a larger one. These differences are not significant and could be explained by the method of creating the reduced documents. The random selection of word chunks was done afresh, from the original sized document for each reduction, therefore the same word chunks had a decreasing likelihood of being included in each decremented document size. Features that have a higher discriminatory power may have been included in the smaller documents but not included from the larger ones due to the random nature of the chunk selection.

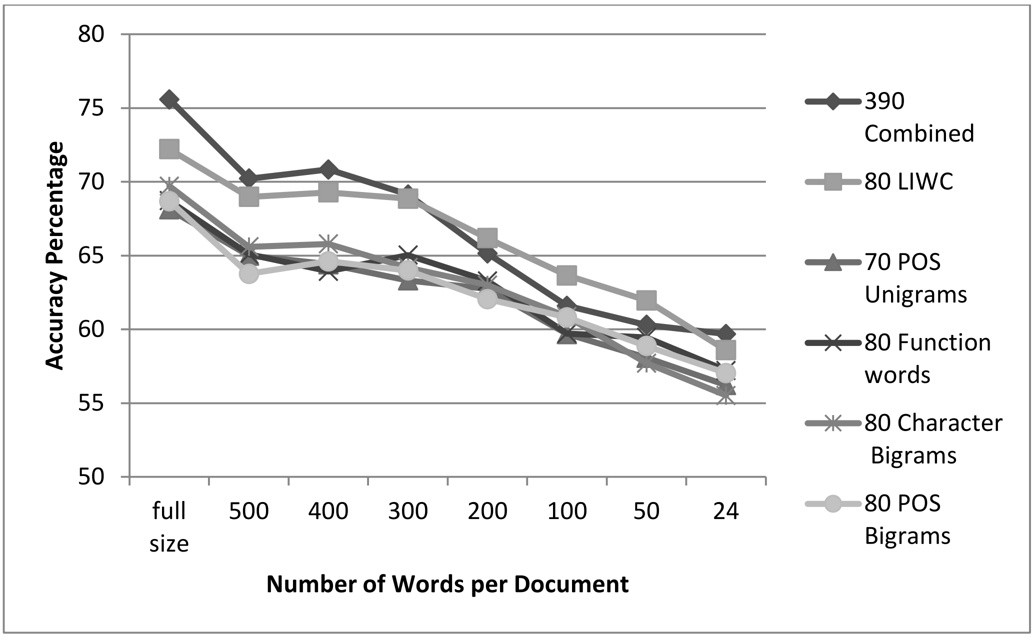

Figure 5-4: Effect on accuracy of combination feature set compared to the five single type feature sets as document size falls for gender classification for gender classification

5.4. Shortened Feature Sets and Document Sizes

The effect of the document size and feature set size decreasing simultaneously was then tested. The two most effective feature sets from Section 5.3 were chosen, the 390 five way combination feature set and the full LIWC feature set. The results from Section 5.4 showed that as the feature sets became smaller, the J48 algorithm was the most effective so that ranking was used in this experiment. Seven different features sets were tested, the full complement of features from both feature sets and sets consisting of the top 80, 40 and 5 features of the combined feature set, and the top 40, and 5 features of the LIWC feature set.

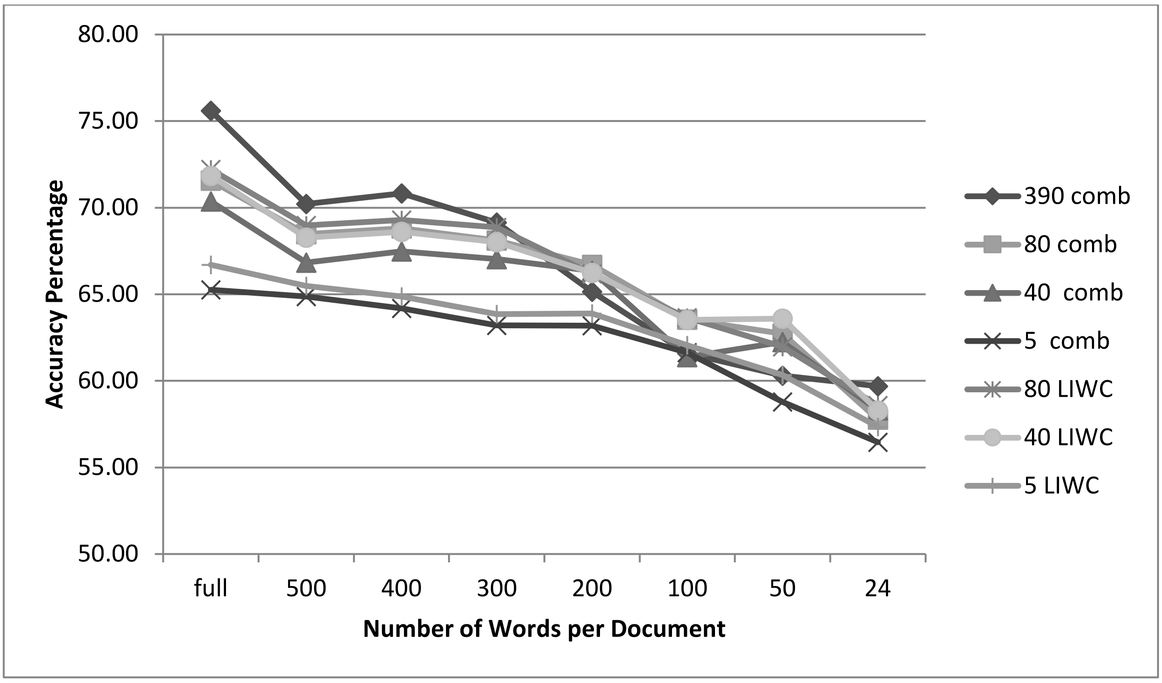

Figure 5-5: reducing feature set size and text size for gender classification.

Figure 5-5 shows the results when both the number of words per document and the number of feature per feature set are reduced. There is a consistent fall in accuracy as the number of words per document decreases for all the feature sets, but as the document size decreases, there is less difference between the feature sets. In the full sized documents there is 10.3% difference in accuracy between the most accurate feature set, the 390 combination feature set, and the least accurate feature set, the 5 combination feature set, where at 24 words per document there is only 3.2% between these two feature sets. For all the feature sets except the two with only five features, there is a sharp drop in accuracy between the full sized documents and the 500 word documents. There appears to be an increase in accuracy from the 500 word documents to the 400 word documents, but this rise is less than one percent and not statistically significant. The same applies to the apparent increase in accuracy between the 100 words documents and the 50 word documents. There is no real difference between the 80 combined feature set, 80 LIWC features and 40 LIWC features across all document sizes. The 40 combined features give almost the same accuracy as these three feature sets for the documents with more than 100 words. The accuracy of the 40 combined features falls more sharply between the 200 and 100 word documents, statistically more so than the other feature sets (*p* = 0.0105). The similarity between the combination feature set and the LIWC feature set at 80, 40 and 5 features can be explained by the proportion of LIWC features in the combinations. In the 80 combination feature set, 26 of them (32.5%) were LIWC. As the number of features fell, the proportion of LIWC features increased, to 42.5% or 17 features in the 40 feature combination and 40% (two features) in the five feature combination. As was noted in Table 5-8, one or two features from each feature set account for much of the accuracy in any feature set. The top LIWC feature, ‘ppron’ (**p**ersonal **pron**oun) was present in all of the combination feature sets.

5.5. Discussion

At the beginning of this chapter, it was asked if men and women use language differently. It was hypothesised that many of the effects linked to the hormone testosterone such as aggression, negative moods, decreased verbal ability, concerns over status and dominance, and more direct thought and action (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012; Newman et al., 2008; Pennebaker et al., 2004), would be prominent in language use and expression and would strongly differentiate between male and female authors.

While the results for all the experiments were significantly above chance, there was not as great a distinction as had been expected. The highest accuracy was between 74% and 75%, achieved by a LIWC in combination of various other feature sets. In comparison with the accuracy for the first language classification experiments, detailed in Chapter 4, where the best result was around 60% over sixteen categories, a higher result was hoped for with only two classes Previous studies have also achieved higher accuracy rates, but they have included features such as names, topics and online behaviour (Argamon et al., 2009; Ludu, 2014; Ugheoke, 2014). The focus of this study was to identify the gender of an author without resorting to features that can easily be counterfeited, such as name or topic information, to create a robust classifier that could detect gender even when it was being deliberately misrepresented by the author. To this end a feature selection method that eliminated or at least vastly reduced the inclusion of content words was used for the function word feature set. While LIWC counts words that are not strictly function words, the words are under general headings and not used as content individually and are not specific to a genre or topic. Therefore there is less impact than using topic specific content words. The character bigram feature set, with only two letters per feature, indicates phonemes rather than words with one notable exception found. The character bigram “xb” was found to be surprisingly common in the teen texts but not as much in the adult texts. Investigation revealed that the bigram is present in the trade name of some game software, “xbox”. This term was used by teens far more than adults. The POS features, the POS bigram and POS unigram feature sets, reduce content words to their part of speech tag. Therefore the character bigrams and POS bigrams and unigrams are, for the most part, unrelated to content. The remaining feature set explored, LIWC creates 80 aggregate features that measure various dimensions of the text, and does not include specific words.

The first research question for this thesis seeks to examine the effectiveness of psycholinguistically based features in comparison to lexicographic features types. To examine if LIWC, the example of a psycholinguistically derived feature set used for this research, was more effective than the other four, lexicographically based feature sets, the LIWC features were tested against increasing numbers of the other, lexicographic feature types. It was found that LIWC was significantly more accurate than similar numbers of lexicographic features, and that the LIWC features were not significantly less accurate than up to 1000 of function word or character bigram features, and were significantly more accurate than 1000 of the POS bigram feature sets when used to classify the gender of an author.

Previous studies and the experimentation in Chapter 4 have shown that a combination of features is more effective for authorship characterisation classification. The second question for this thesis seeks to examine whether LIWC is more useful in a combination than other, lexicographic feature types. Various combinations of feature types and numbers were considered. It was found that adding the 80 LIWC feature sets to 200 of the lexicographic feature sets improved accuracy significantly more than simply increasing the number of the same feature type by that amount. When all the feature types were combined in approximately equal numbers, it was found that the removal of the LIWC feature type significantly reduced the accuracy, while removal of the other feature types did not, and that the feature sets that contained four of the five feature sets including LIWC, were not significantly less accurate than the feature set that contained all five feature types.

Another of the aims of this thesis is to investigate the effect of reducing the number of features used in classification exercises, including that of an authorship gender classification exercise, and as an augmentation of this exercise, identifying the most useful features in any given feature set. A base of the 80 features used in the initial comparisons (70 for POS unigrams) were used as a base and ranked according to three feature selection algorithms: information gain, gain ration and J48 tree. When this exercise was conducted on the ICLE language corpus, there was a significant difference between the three ranking algorithms used, however for the gender classification, while the J48 tree ranking was marginally more accurate, there was little, if any significant difference between them. The J48 tree was significantly less accurate for the first language feature rankings. This could be explained by the fact that there were 16 individual language classes in the ICLE corpus, while there are only two gender classes in the Blog Authorship Corpus. The J48 tree, by its nature, is good at separating two classes.

As the feature sets were reduced there was a surprisingly small reduction in accuracy. Again when this exercise was conducted on the ICLE corpus, there was a marked and significant reduction in accuracy for each reduction in feature set size. There was no significant difference in accuracy between the full feature sets and half sized feature sets for any feature type when ranked using the J48 algorithm, for the gender classification. When the feature sets were reduced to 40 features, one quarter of the base sets, there was a significant, but not substantial drop in accuracy for LIWC, POS unigrams, character bigrams and POS bigrams, but no significant difference for function words. The feature set size was reduce to ten and then to five features, and while there was a significant reduction in accuracy, the results were still well above chance. At all feature set sizes, LIWC was significantly more accurate than the other feature types. The accuracy of classifiers based on only one feature was then examined. It was found that for gender classification, a substantial amount of the classification accuracy could be accounted for by just one feature in each of the feature types. This marked difference from the outcomes of the first language classifications could again be explained by the difference in the number of classes. Gender classification only needs to split the corpus into two parts, male and female. The most effective features were personal pronouns and determiners, and the most effective of these were the representative classes in LIWC. Unlike the other features, the LIWC pronouns and determiners are an amalgamation of many words rather than a single term, giving them more discriminatory power. This agrees with research that shows females are more socially oriented in their use of language, using more pronouns, and males are more task oriented, using more articles. (Argamon et al., 2009; Newman et al., 2008; Pennebaker et al., 2004)

The final aim of this research was to investigate the point at which a document becomes too small to accurately classify. This is an important question because there is a considerable amount of fraud committed by misrepresentation of gender and other demographic details in computer mediated communication (CMC) and other short message systems. As the document lengths were reduced in the Blog Authorship Corpus, there were only small reductions in accuracy, until, at 24 words per document, the accuracy was approximately 10% lower than that of the classification of the full sized documents. The overall patterns remained the same, that LIWC was more accurate than any of the other single feature types, that it also gave similar or better accuracy than the combination feature sets, and that the combination containing LIWC was more effective than a combination excluding LIWC.

From this it can be concluded that males and females do use language differently and that the difference is measurable. The psycholinguistically based LIWC feature set was more effective for gender classification than other, lexicographically based feature sets, and it is sufficiently different that it adds information rather than noise when combined with other feature sets. If the right feature is used, a moderately effective authorship gender classifier can be based on only one feature and documents as small as 24 words can be classified on the gender of the author at a rate considerably higher than chance.

Chapter 6 Profiling for Age Group of Author

In this chapter, the effectiveness of various feature sets for the classification of documents based on the age group of the author will be explored. The misrepresentation of age by internet users is becoming a significant issue for both law enforcement agencies and child protection organisations. Online predators can use the anonymity of the internet to misrepresent their age and/or gender to befriend vulnerable children and adolescents and lure them into inappropriate or dangerous activities (Dombrowski et al., 2004; Eneman, Gillespie, & Bernd, 2010). Underage adolescents can also misrepresent their age to gain inappropriate access to adult sites.

There have been several studies into profiling authors by age group. However, as for the gender profiling results in this study, the previous studies' results are not comparable to the results in this chapter. This is again because the previous studies have used content words and stylistic features as part of the feature sets in the classification. While there is no doubt that different age groups have different concerns and there for discuss different topics, however if a classifier is to be developed that can distinguish between different age groups' speech when the topics are constrained, then these features must be eliminated.

The corpus used for this chapters is the same subset of the Blog Authorship Corpus compiled by (Schler et al., 2006) that was used in Chapter 5. A summary of the age related details of the sub-corpus used is given in Table 5.1. As was discussed in Section 3.1.2, the age groups of the Blog Authorship Corpus are predefined and the actual ages of the participants are not available although Schler (2006) noted that the oldest participant was 47 years of age.

Age Group :	Number of Files	Total Words	Average File Size	Smallest File	Largest File
teens :	1,332	7,134,431	5,356	343	207,357
twenties :	1,332	10,362,585	7,779	231	420,608
thirties :	1,332	11,593,058	8,703	321	339,051
Total	3996	29,090,074	7279	231	420,608

Table 6-1: Summary of age group files used from the Blog Authorship Corpus

The thirties age group were the most verbose, with the highest average number of words per file. The teens age group were the least talkative with both the lowest total and average number of words, while the twenty age group had the greatest spread, with both the longest and shortest files in the selected corpus. The feature sets being tested all use ratios rather than counts of features, so the differences in file size are not expected to impact on the results.

The files were pre-processed to remove all data other than the text information, including dates and breaks between posts. Multiple whitespace characters, such as several new line characters or a row of tabs or spaces were reduced to one whitespace character. Line spacing was not one of the features used, nor paragraph length, so all newline characters were replaced with a single space character. The feature sets used for the age classification exercises were the same feature sets used for the gender classification detailed in Chapter 5

The corpus was divided as for the gender classification experiments, with one third of the corpus was used to rank the features using the methods discussed in Section 3.4 and the remaining two thirds used for testing. As for the gender rankings, there was little or no difference between the accuracies of the various rankings so the document count method was used to rank the three larger feature sets: the function words, character bigram and POS bigrams.

6.1. Comparing and Combining Psycholinguistic and Lexicographic Features on Full Sized Documents for Age Group Classification (Three Age Group Classes)

The process for these experiments followed the patterns established in the previous two chapters. First the accuracy of classifiers using similar numbers of the five feature types were compared for accuracy, followed by increasing numbers of the three larger feature sets, and finally, various combinations of feature types and numbers were tested. The POS unigram feature set consisted of only 70 features because that was all the POS tags that were present in the corpus.

6.1.1. Single Feature Types (Three Age Group Classes)

The initial tests were carried out on 80 (or 70 in the case of POS unigram) features from each feature type and the results compared. When this exercise was conducted for the gender classification, the LIWC feature set was markedly more accurate than the other feature types, and it was anticipated that LIWC would again prove to be a very effective classification tool.

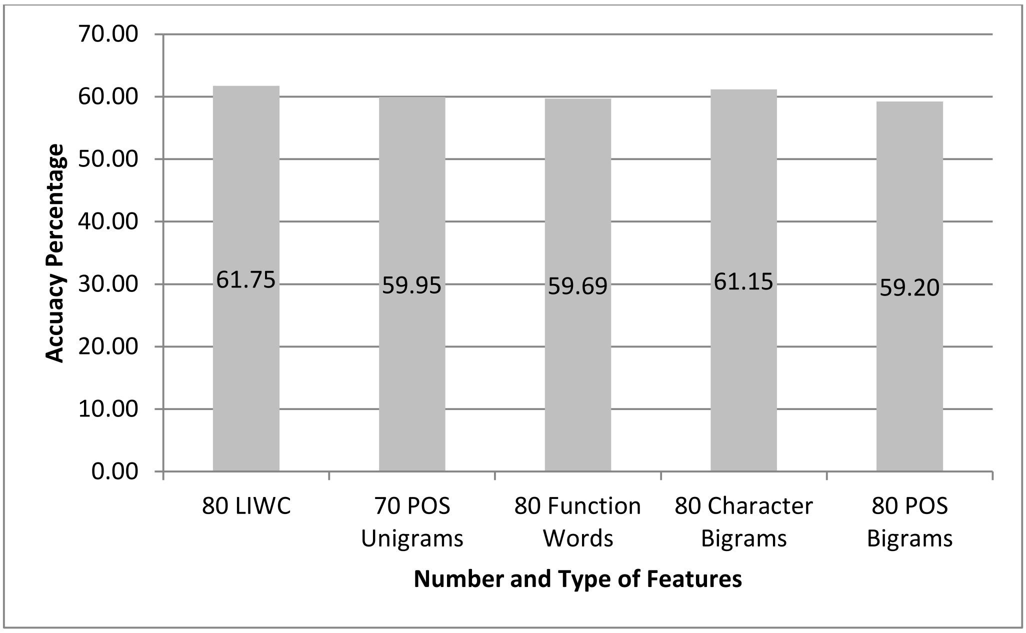

Figure 6-1: Comparison of accuracy of 80 LIWC, 80 function word, 80 character bigram, 80 POS bigram and 70 POS unigram features for three age groups: teens, twenties and thirties.

The results of these experiments are shown in Figure 6-1. As can be seen, the results were much less conclusive than for the gender classifications where LIWC was significantly more accurate than any of the other feature types. While the LIWC results are significantly more accurate than the least successful feature set, POS bigrams (*p* = 0.0357) they are not significantly different to any of the other three feature sets.

Feature Type	Actual Age Group	Predicted Age Group		
		teens	twenties	thirties
LIWC	teens	**72.5%**	19.0%	8.4%
	twenties	15.4%	**45.2%**	39.4%
	thirties	5.8%	26.5%	**67.6%**
POS Unigrams	teens	**71.5%**	18.8%	9.7%
	twenties	19.0%	**45.3%**	35.7%
	thirties	8.9%	28.0%	**63.1%**
Function Words	teens	**71.0%**	18.2%	10.7%
	twenties	18.4%	**44.7%**	36.9%
	thirties	7.4%	29.3%	**63.3%**
Character Bigrams	teens	**70.8%**	18.9%	10.2%
	twenties	18.1%	**47.7%**	34.1%
	thirties	7.8%	27.4%	**64.9%**
POS Bigrams	teens	**71.1%**	18.7%	10.1%
	twenties	18.7%	**40.8%**	40.5%
	thirties	6.4%	27.9%	**65.6%**

Table 6-2 Percentage values for confusion matrices for similar numbers of the five feature type across three age group classes (teens, twenties and thirties)

The results were so different from the results in Section 5.1, an examination of the confusion matrices was undertaken to attempt to ascertain the cause of the discrepancy. The percentage values from the confusion matrices are shown in Table 6-2. The correct classifications are shown in boldface type and the twenties age group, the class that gave the most errors, is shaded. As can be seen, the largest number of errors in classification occurred between the twenties and thirties age groups, although there is also a high number between teens and twenties. The smallest error rates were between the thirties and teens age groups, but that would be expected since they are at opposite ends of the spectrum. These patterns of misclassification were consistent across all feature types.

Even though the comparison of the five single type feature sets gave inconclusive results, increasing numbers of the three large feature sets, (function words, character bigrams and POS bigrams) were tested to explore the effect of larger, single type feature sets. The results are shown in Table 6-3 . The table also includes the LIWC results for the purposes of comparison. The highest results for each feature type are indicated in bold highlighted text. When this comparison was undertaken in the gender classification, the LIWC feature set was surprisingly close to the larger feature sets, considering there were between 5 and 10 times as many features present.

Feature Type	Number of Features per Features Set					
	80	200	400	600	800	1000
LIWC	61.75					
Function Words	59.69	64.15	64.53	65.20	**67.27**	66.33
Character Bigrams	61.15	63.17	**65.32**	65.28	63.67	63.21
POS Bigrams	59.20	59.83	60.06	62.20	**65.69**	62.05

Table 6-3: Accuracy for increasing numbers of single type feature sets for classification of author age group over three age group classes (teens, twenties and thirties)

As seen in the previous chapters, the function words were very effective when used to classify document on the first language or the gender of the author, and this efficacy improved as the number of features increased, although that could have been due to the increasing number of content specific words present. It was, therefore, expected that increasing numbers of function words would also increasingly differentiate between the three age groups. However as can be seen in the second line of Table 6-3 , the results did not bear this out. The accuracy increases with each increase in the number of features, albeit not always significantly, until reaching a peak at 800 words, after which it declines. A one-way between-groups analysis of variance indicated that there was a significant difference between the six different sized function word features sets: *F* (5, 54) = 6.63 *p* < 0.01). The effect size, calculated using eta squared, was 0.38 indicating that the difference in mean scores between the groups was quite large, however, post-hoc comparisons, using the Tukey HSD test revealed that the difference between most of the feature sets with the ones adjacent (ie the difference between 200 and 400 features) were not significant on their own. For further explanation of the Tukey HDS test see (Pallant, 2011). The only exception was the increase from 80 features (*M* = 59.69, *SD* = 3.16) to 200 features (*M* = 64.16, *SD* = 3.18) (*p* = 0.036). As shown in Table 6-4, the teens age group was again the most effectively classified, with the twenties giving the highest number of errors between it and the thirties age group and to a lesser, but still disproportionally large extent, between it and the teens age group.

80 Function Words				800 Function Words			
	Teens	Twenties	Thirties		Teens	Twenties	Thirties
Teens	71.1%	18.2%	10.7%	Teens	74.9%	18.8%	6.3%
Twenties	18.4%	44.7%	36.9%	Twenties	14.2%	58.3%	27.5%
Thirties	7.4%	29.3%	63.3%	Thirties	4.9%	29.3%	65.8%

Table 6-4:Percentage values for confusion matrices for 80 and 800 function words across three age group classes (teens, twenties and thirties)

Chi-squared goodness-of-fit tests done for the individual feature set sizes again showed a significant difference between the accuracy for the three age groups across the entire range of feature set sizes. The results for the chi squared tests for the examples given for the lowest and highest accuracy (shown in Table 6-4) were: 80 function words: χ^2 (2, n = 1590) = 54.56, p < 0.001, 800 function words: χ^2 (2, n = 1792) = 18.13, p < 0.001.

Character bigrams also were very successful when used for first language and gender classification. However, the results for the age group classification using character bigrams were very similar to those produced using the function word feature set, but, as can be seen in the third line in Table 6-3, with the peak in accuracy occurring at 400 features rather than at 800. A one-way between-groups analysis of variance test showed that there was a significant difference between the six character bigram feature set sizes, but with less confidence *F* (5, 54) -= 2.7, *p* = 0.03. The effect size, calculated using eta squared was also lower, 0.199, although still indicating a large difference between means. Post-hoc comparisons using the Tukey HSD test showed that there was no significant increase between any of the feature sets with their adjacent sets.

80 Character Bigrams				400 Character Bigrams			
	Teens	Twenties	Thirties		Teens	Twenties	Thirties
Teens	70.8%	18.9%	10.2%	Teens	73.0%	20.2%	6.9%
Twenties	18.1%	47.7%	34.1%	Twenties	13.7%	56.4%	29.8%
Thirties	7.8%	27.4%	64.9%	Thirties	5.5%	27.9%	66.5%

Table 6-5: Percentage values for confusion matrices for 80 and 400 character bigrams across three age group classes (teens, twenties and thirties)

Table 6-5 give the confusion matrices for 80 features (the least accurate for the character bigram sets) and 400 features (the most accurate for the character bigram sets). Chi-squared goodness-of-fit tests conducted on the individual feature set sizes again revealed that, although the differences from the expected accuracy rates were not as large as for the function word feature sets, the differences between the three age groups were significant across all feature set sizes within the character bigram feature sets. Chi squared results for the least accurate feature set (80 features) and the most accurate feature set (400 features) are: 80 features: χ^2 (2, n = 1629) = 41.71, p < 0.001; 400 features: χ^2 (2, n = 1740) = 18.94, *p* < 0.001.

The results for the POS bigrams are in the fourth and final line of Table 6-3. They show effectively the same results as the function word and character bigram experiments, but with a sharper peak in accuracy occurring at 800 features. A one-way between groups analysis of variance showed that there was a significant difference between the means across the group of six POS bigram feature set sizes *F*(5, 54) = 9.91, *p* < 0.01. Post-hoc comparisons using the Tukey HSD test again showed that there was no statistical difference between most of the feature set sizes and the ones adjacent, with the exception of the 800 POS bigram feature set (*M* = 65.69, *SD* = 2.06), which showed a significant difference from the 600 POS bigram feature set (*M* = 62.2, *SD* = 2.03, *p* = 0.32) and the 1000 POS bigram feature set (M = 62.5, *SD* = 2.24, *p* = 0.23).

80 POS Bigram				800 POS Bigram			
	Teens	Twenties	Thirties		Teens	Twenties	Thirties
Teens	71.2%	18.7%	10.1%	Teens	77.7%	16.8%	5.9%
Twenties	18.7%	40.8%	40.5%	Twenties	14.6%	55.1%	30.3%
Thirties	6.42=%	27.93=%	65.6%	Thirties	4.4%	31.0%	64.6%

Table 6-6: Percentage values for confusion matrices for 80 and 800 POS bigram features across three age group classes (teens, twenties and thirties)

The confusion matrices for the six POS bigram feature sets conformed to the pattern established by all the previous classification experiments conducted in that the teens age group was the most accurately classified, closely followed by the thirties age group, with the twenties being comparatively poorly classified. The least accurate POS bigram set was again the 80 feature set and the most accurate POS bigram set was the one consisting of 800 features.

The confusion matrices for the least accurate POS bigram set (80 features) and the most accurate POS bigram set (800 features) are given in Table 6-6. Chi squared goodness-of-fit tests conducted on each individual sized feature set again revealed that the differences between the age groups were significant across all sizes of the POS bigram feature sets. The results for the feature sets shown in Table 5 are: 80 POS bigrams: χ^2 (2, n = 1577) = 78.72, *p* < 0.001; 800 POS bigrams: χ^2 (2, n = 1750) = 33.82, p < 0.001.

6.1.2. Combination Feature Sets (Three Age Group Classes)

A combination of feature types was more effective than any single type feature for the first language classification exercises, and it was considered that this might also be the case for the age group classification. In the first language classification, the removal of the LIWC feature set or the character bigram feature set had the largest impact on accuracy. The same experiment was conducted on the age group corpus. 80 of each of the LIWC, function word, character bigram and POS bigram feature sets, and the full 69 present of the POS unigram feature set were combined, and then each feature set was removed in turn. The results are given in Table 5.7. A one-way between groups analysis of variance test confirmed that there was no significant difference between the means of the five feature sets (*F* (4, 45) = 2.808, *p* = 0.895.

Feature Types Included (✓) or Omitted (✗)					Accuracy
LIWC	POS Unigrams	Function Word	Character Bigrams	POS Bigrams	
✗	✓	✓	✓	✓	64.75
✓	✗	✓	✓	✓	64.26
✓	✓	✗	✓	✓	64.91
✓	✓	✓	✗	✓	64.60
✓	✓	✓	✓	✗	**65.69**
✓	✓	✓	✓	✓	**65.43**

Table 6-7: Comparative accuracy of the "leave one out" feature sets combination feature sets and the five way combination feature set for age group classification across three age groups (teens, twenties and thirties)

The confusion matrices for the five experiments followed the same pattern as all the other age group classification exercises undertaken. The teens age group were consistently well classified, with more than 75% of the 888 teen texts correctly categorised across all feature sets. The thirties age group was less accurately classified with between 66% and 68% of the texts correctly classed. The twenties age group was again very poorly classified, with between 50% and 54% of the 888 files correctly assigned. The confusion matrices for the most accurate, (the LIWC, POS unigram, function word and character bigram), and least accurate (the LIWC, function word, character bigram and POS bigram) feature sets are given in Table 6-8.

LIWC, POS Unigrams, Function Words and Character Bigram Feature Combination (most accurate)				LIWC, Function Word, Character Bigram and POS Bigram Feature Combination (least accurate)			
	Teens	Twenties	Thirties		teens	twenties	thirties
Teens	75.3%	18.8%	5.9%	Teens	75.7%	18.5%	5.9%
Twenties	13.6%	53.8%	32.5%	Twenties	14.5%	50.4%	35.0%
Thirties	5.2%	26.9%	67.9%	Thirties	5.4%	27.9%	66.7%

Table 6-8: Percentage values for confusion matrices for the 320 feature “leave one out” feature sets with the highest and lowest accuracy for age group classification (three age classes: teens, twenties and thirties)

A chi squared goodness-of-fit test for the most accurate combination (LIWC, POS unigrams, function words and character bigrams) indicated that there was a significant difference between the number of correct classifications for each of the age groups: χ^2 (2, n = 1750) = 32.26, $p < 0.001$. A chi squared goodness-of-fit test conducted for the lest accurate combination (LIWC, function words, character bigrams and POS bigrams) also indicated that there was a significant difference in the number of texts classified correctly between the three age groups: χ^2 (2, n = 1712) = 45.16, $p < 0.001$. Chi squared goodness-of-fit tests conducted for the other “three leave one” out combination feature sets all gave similar results, that there is a significant difference in the number of correct files between the age groups within each of the feature sets.

6.1.3. Discussion of Results Across Three Age Groups (Teens, Twenties and Thirties)

The results for both the single type and combination feature sets were disappointing in that they resulted in a much lower accuracy percentage than the accuracy rates achieved for the first language and gender profiling exercises. The results for all of the feature sets, regardless of number or composition were reasonably homogenous, with changes to composition and/or size having only a small effect on accuracy. The confusion matrices were remarkably consistent showing that the teens age group was far more accurately classified than the other two age groups and the thirties age group was more accurately classified than the twenties age group, which was consistently very poorly categorised. The chi squared goodness-of-fit tests conducted individually on the confusion matrices showed that there were significant differences between the proportions of correctly classified instances for each age group and the proportions that could reasonably be expected.

The observed results are possibly explained by recent discoveries in the neurological field. Neurological studies using functional Magnetic Resonance Imaging (fMRI) scans have given new insights into how the brain grows and matures. The fMRI scan allows researchers to observe brain activity in real time. These studies have shown that the brain does not fully mature until the third decade of life (ie between the ages of 21 to 30) and that the maturation process 'flows' across the brain from back to front, and centrally to outer lobes (Casey, Galvan, & Hare, 2005; Luna, Garver, Urban, Lazar, & Sweeney, 2004; Zukerman & Purcell, 2011). While the parts of the brain that handle speech such as Broca's area (which assists in speech production and grammar) and Wernicke's area (which is used in the understanding of speech) are in the temporal and central parts of the brain, the area of the brain which controls the 'executive' functions of the brain (such as impulse control and conversational narrative control) are in the frontal lobes which are the last to mature (Brauer, Anwander, & Friederici, 2011; Isaacowitz & Riediger, 2011; Luna, Padmanabhan, & O'Hearn, 2010; Nelson & Guyer, 2011; Steinberg, 2005). The time the brain takes to mature, as shown by these studies, could also explain the poor performance of the twenties age group in the language classification tasks. The twenties age group, consisting of individuals aged from 23 to 27 is right in the middle of the transition zone from adolescent frontal lobes to mature, adult frontal lobes.

As has been discussed in Section 3.1.2, the Blog Authorship Corpus has the age groups pre-defined and the actual age of individual participants is not available. Therefore, to examine the effect of age group boundaries, the corpus was divided in to the three age groups, and then two combination age groups were created, a "junior" age group consisting of teens and twenties files, and an "adult" age group consisting of twenties and thirties files. The combination age groups were created by randomly selecting 444 files from each of the constituent age groups so the files would not be skewed towards either component age group and number of files would be consistent with the original age groups. Five new sub-corpora were then created, each consisting of two of the age groups.

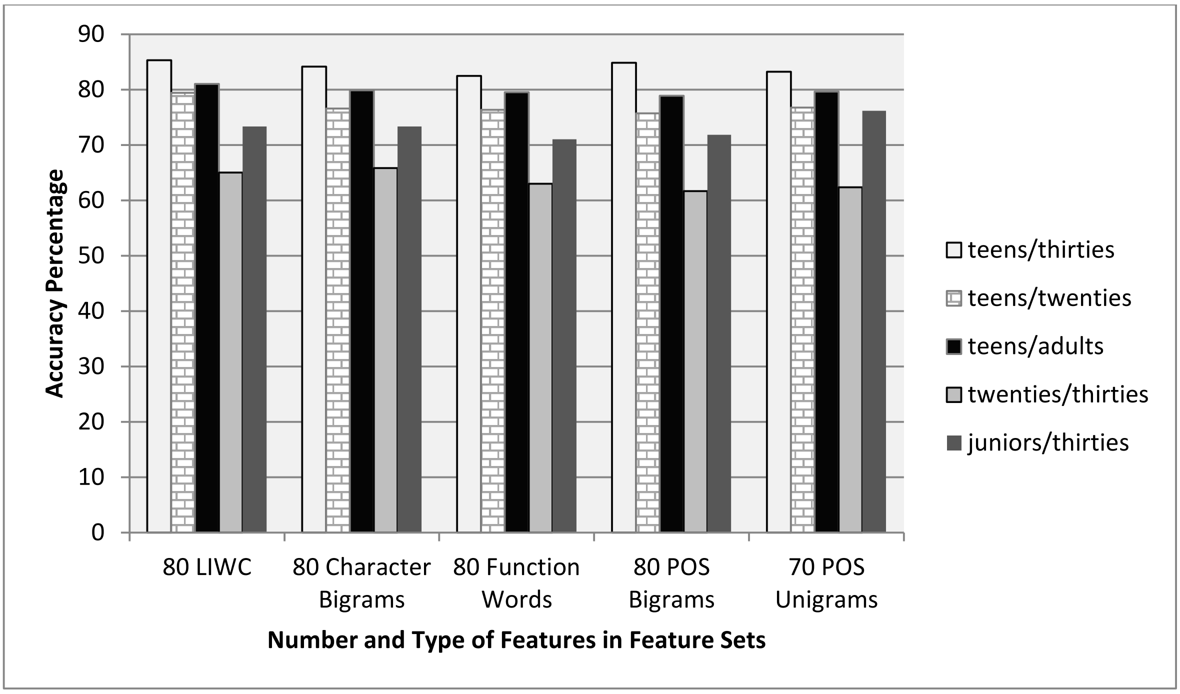

Figure 6-2: Comparison of accuracy for pairs of five age groups using the five base feature sets

Figure 6-2 shows the accuracy results for the five age group pairs classified with each of the five feature types. What was immediately apparent, although perhaps to be expected, was that the teen/thirties classification exercise yielded far higher accuracy than the other age group pairs. The removal of the twenties age group would have created a large gap between the eldest example of the teens age group and the youngest example of the thirties individuals, and if it is possible to discriminate text on the age of the author, this gap would be expected to make the classification much more clear cut. What was less expected was that the teens/twenties age group pairing also gave high accuracy, along with the teen/adult pairing, while the junior/thirties pairing and as would have been expected, the twenties/thirties pairing showed very poor classification results across all five feature types. In fact the twenties/thirties pairing only achieved between 2.4% and 4.6% higher accuracy than the classification exercise that was conducted across all three age groups. One-way between groups analysis of variance conducted on each of the five feature sets showed that there was a significant difference between the means of the five age group pairings within each feature set.

These results indicate that there was very little difference between the twenties and thirties age groups use of language within the blog corpus, with the greatest difference being between the teens and thirties, with the next highest differences being in the teens/adults sub corpora. Chi squared goodness-of-fit tests done for the classification exercise undertaken with each of the five feature sets on the teens/adults corpus showed no significant difference in the number of accurately classified text for each age group. The confusion matrices for the most (LIWC) and least (POS bigram) accurate feature sets for the teen/adult corpus are given in Table 6-9. For the LIWC feature set χ^2 (1, n = 1439) = 1.81, $p < 0.179$, for the POS bigram feature set, χ^2 (1, n = 1401) = 222, $p < 0.128$.

80 LIWC			80 POS Bigrams		
	Teens	Adult		Teens	Adult
Teens	78.1%	21.8%	Teens	75.7%	24.3%
Adult	16.1%	83.9%	Adult	17.9%	82.0%

Table 6-9: Percentage values for confusion matrices for 80 LIWC features and 80 POS bigram features for the teen/adult corpus age group classification

6.2. Results for the Teens/Adults Corpus

The χ^2 showed that the age group that was not being classified as well as would be expected was the twenties age group and neurological studies indicated that this could be a result of the stage of brain maturation in the twenties age group. The results shown in Figure 6-2 showed that the best division of the corpora for age group classification would be the teens/adults division. Therefore, the classification exercises that had been conducted on the teens/twenties/thirties corpus were repeated on the teens/adults corpus to discover if the amalgamation of the two apparently similar age groups (the twenties and thirties age groups) would improve the accuracy and give results closer to that achieved in the gender classification exercises undertaken in Chapter 5. There is also a practical relevance to these age classes in that they are probably the age distinctions that are of most interest for child protection applications (Gupta et al., 2012; Whittle et al., 2013). The experiments based on the two age group corpus are detailed in the next section.

6.2.1. Teens/Adults Corpus – Comparing Psycholinguistic and Lexicographic Features Sets on Full Sized Texts.

Feature sets consisting of 80 each of the LIWC, function word, character bigram and POS bigram features and 70 of the POS unigram features were compared for accuracy. The results are shown in Figure 6-3. The results are much higher than those from the three age groups reported in Section 6.1, although any direct comparison of the accuracy is problematic due to the different number of classes. When the accuracy of the five feature types, over the two age groups were compared, the LIWC feature set was significantly more accurate than the lexicographic feature sets (*p* < 0.0444), although the difference was relatively small.

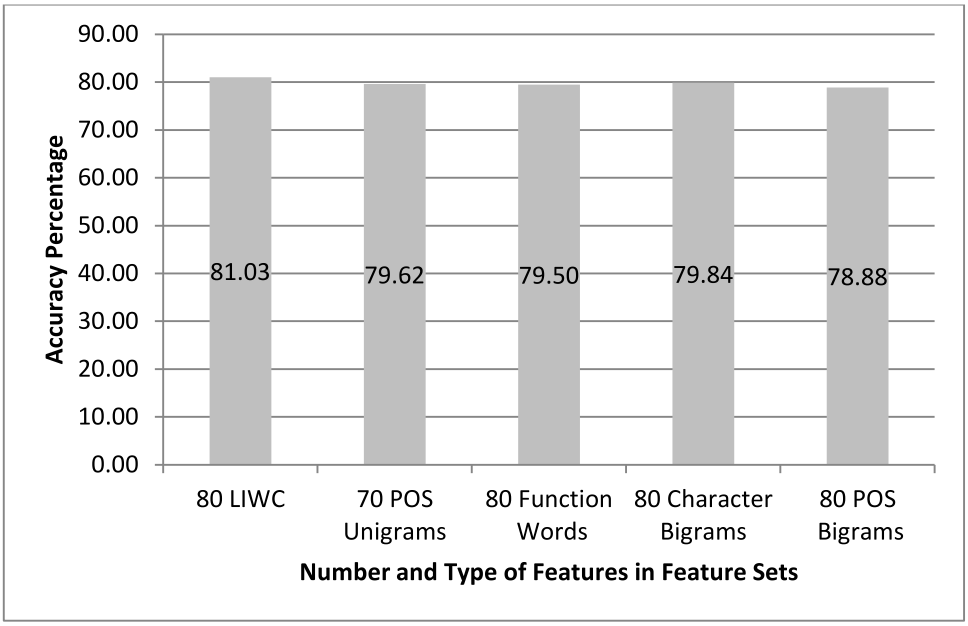

Figure 6-3: Comparison of accuracy for 80 LIWC, 70 POS unigram, 80 function word, 80 character bigram and 80 POS bigram features for classification of teen and adult age groups

To examine the effect of increasing the numbers of features in each feature set, increasing numbers of the large feature types (function words, character bigrams and POS bigrams) were tested on the teens/adults corpus. The results are given in Table 6-10. There was a significant increase in accuracy from 80 to 200 features in all of the feature sets. However there was no significant increase in accuracy between the 200 and 1000 sized feature sets in any of the feature types. None of the results for any of the feature set sizes in any of the feature types were significantly more accurate than the results for LIWC.

Feature Type	Number of Features per Features Set					
	80	200	400	600	800	1000
LIWC	81.03					
Function Words	79.50	82.15	82.71	81.25	82.94	**82.88**
Character Bigrams	79.84	81.36	82.60	**82.26**	81.93	80.85
POS Bigrams	78.88	79.56	80.47	80.30	**80.91**	79.96

Table 6-10: Accuracy for increasing numbers of single type feature sets for classification of author age group over two age group classes (teens and adults)

6.2.2. Teens/Adults Corpus – Combining Psycholinguistic and Lexicographic Features Sets on Full Sized Texts.

The results from previous chapters have shown that, for first language and gender classification, combining LIWC with a larger number of a single type of lexicographic features increased the accuracy more than adding another 80 features of the same type. Experiments were conducted to ascertain if this pattern would be seen in the age group classifications over two age groups.

The LIWC feature set was combined with 200 of each of the larger feature sets (function words, character bigrams and POS bigrams) and compared to 280 features of the same feature type. The results are given in Figure 6-4. The inclusion of LIWC increased the accuracy significantly for character bigrams and POS bigrams (*p* < 0.05) and by 1.35% for the function words, however this is not significant at the *p* = 0.05 level.

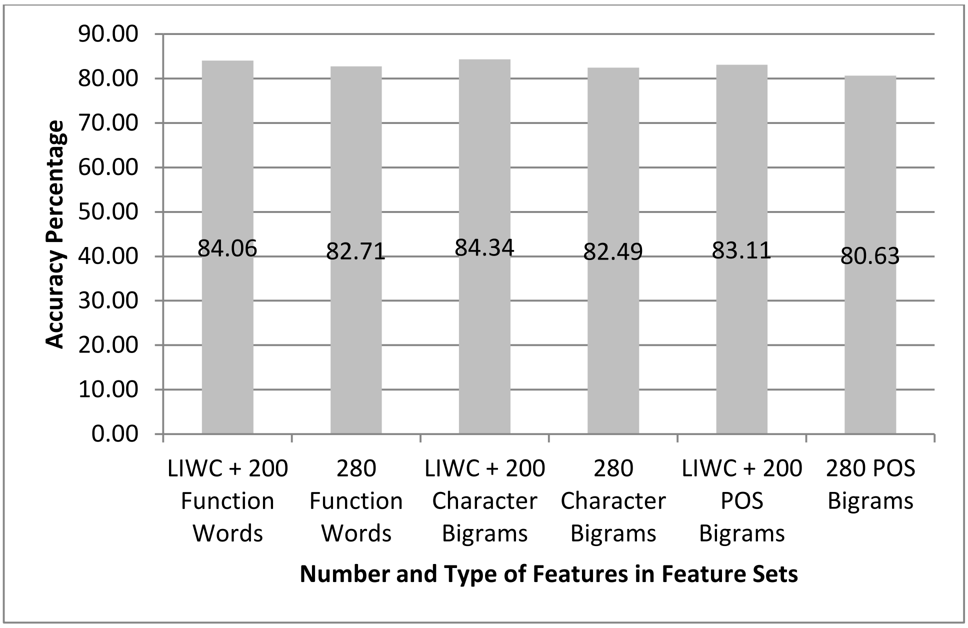

Figure 6-4: Effect on accuracy of adding 80 LIWC features to 200 lexicographic features compared to 280 lexicographic features for classification of two age groups (teens and adults)

The highest accuracy for the larger single feature type feature sets was at approximately 600 features, therefore the effect of adding the 80 LIWC features to 600 of each of the function word, character bigram and POS bigram features was examined. The results are given in Figure 6-5. The addition of LIWC significantly increase accuracy for the function words (*p* = 0.0371) and the POS bigrams (*p* = 0.0036) and there is a 93.5% probability that LIWC also increased the accuracy of the character bigrams (*p* = 0.0644).

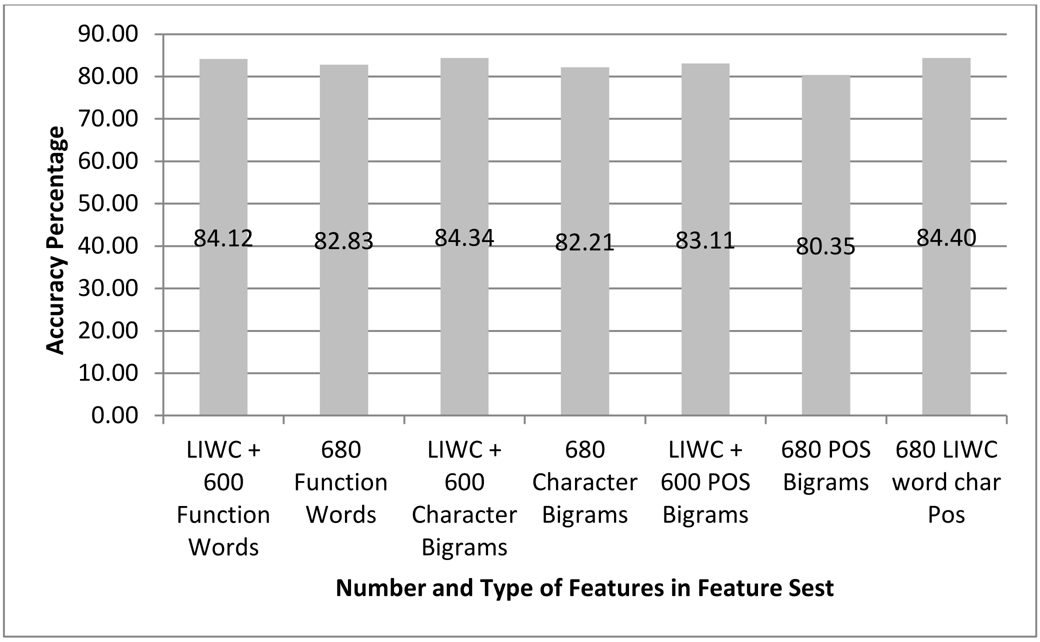

Figure 6-5: Effect on accuracy of adding 80 LIWC features to 600 lexicographic features compared to 680 lexicographic features and a feature sets consisting of 80 LIWC and 200 each of function words, character bigrams and POS bigrams, for classification of two age groups (teens and adults)

Figure 6-5 also shows the results for a 680 feature combination feature sets consisting of 200 each of the function word, character bigrams and POS bigrams and the 80 LIWC features. There was no significant difference between this feature set and the 600 lexicographic features plus LIWC feature sets.

To examine the effect of including LIWC with a combination of lexicographic feature types, a combination feature set was created using approximately equal numbers of the five feature types being examined. A series of feature sets consisting of a combination of four of the feature types was generated by systematically removing each one of the component feature sets, giving five different combinations.

Table 6-11 shows the results for the combination feature sets. As expected, the removal of the LIWC feature set had the most profound impact on accuracy, reducing it significantly ($p = 0.0198$). The removal of the lexicographic feature sets did not have a significant impact on the accuracy of the classifiers. There was no significant difference between the feature set that omitted LIWC and the feature sets that omitted the character bigrams and the POS bigrams ($p > 0.1$). However the feature sets that retained these three feature types (LIWC, character bigrams and POS bigrams) but omitted either the function word or POS unigram feature sets were significantly more accurate than the feature sets that omitted those three feature types ($p < 0.036$).

Feature Types Included (✓) or Omitted (✗)					Accuracy
LIWC	POS Unigrams	Function Word	Character Bigrams	POS Bigrams	
✗	✓	✓	✓	✓	82.43
✓	✗	✓	✓	✓	84.74
✓	✓	✗	✓	✓	84.35
✓	✓	✓	✗	✓	83.67
✓	✓	✓	✓	✗	83.84
✓	✓	✓	✓	✓	84.57

Table 6-11: Comparative accuracy of the "Leave One Out" combination feature sets with the five way combination feature set for classification of two age groups (teen and adult)

6.3. Results for Reduced Feature Sets with Full Sized Documents

While there was no significant difference between the accuracy within feature types when they were ranked using the methods discussed in Section 3.4, each feature type had different features ranked highest under different ranking paradigms. This would imply that, as for the first language and gender classification exercises, that all the features in the top n of any given feature type do not have equal impact on the accuracy of the classification. The top 80 features from the function word, character bigram and POS bigram, and the full LIWC and the POS unigram feature sets were each ranked with the three algorithms in the previous two chapters: the J48 tree, the information gain and the gain ratio. The accuracy for each feature set, under each ranking algorithm, was tested on 40 features (half the base for all the feature sets except POS unigrams) and on 20 features (a quarter of all except POS unigrams). These were compared with the top 40 and 20 features, respectively, from the original rankings of the POS unigram, function word, character bigram and POS bigram feature sets. The original ranking for the LIWC feature set was not used because it is the default ranking that is produced by the software and has no relation to the classification accuracy of the features for age group.

Table 6-12 shows the accuracy achieved by classifiers based on the top 40 features for each feature type under each ranking algorithm. The full sized base feature set results are included for comparison purposes. The highest value in each row is in bold text and highlighted, the lowest value is cross hatched. The results were similar to those in the gender classifications but not in the first language classifications when the feature sets were reduced. This is probably explained by the similar number of classes, two as opposed to sixteen in the first language experiments.

Feature Types	Ranking Algorithms: 40 Features per Feature Set				Base Feature Set (70 or 80 Features)
	Information Gain	Gain Ratio	J48 Tree	Original Ranging	
LIWC	79.14	79.22	**80.77**		81.03
POS Unigrams	78.38	78.40	77.14	**79.02**	79.62
Function Words	**79.08**	78.63	77.79	78.63	79.50
Character Bigrams	77.62	**78.01**	76.74	77.42	79.84
POS Bigrams	78.66	78.55	78.60	**78.94**	78.88

Table 6-12: Accuracy of top 40 of the single type base features sets in order as per the gain ration, information gain, original and J48 ranking algorithms for two age classes (teens and adults)

Analysis of variance tests showed that there was no significant difference between the means of the four 40 features sets and the 80 feature set for the POS unigrams (*F*(4, 45) = 1.6089, *p* = 0.1885), the function words (*F*(4, 45) = 1.2671, *p* = 0.2969) or the POS bigrams (*F*(4, 45) = 0.3353, *p* = 0.8527). For the LIWC feature set, there was a significant difference between the two least accurate rankings (information gain and gain ratio) and the most accurate results (the 40 features ranked by the J48 tree and the full 80 feature set)(*p* < 0.0013), but with a difference of approximately 2%, the drop in accuracy was not substantial given that the feature set size has been halved.

Feature Types	Ranking Algorithms: 20 Features per Feature Set				Base Feature Set (70 or 80 Features)
	Information gain	Gain Ratio	J48 Tree	Original Ranking	
LIWC	78.83	78.77	**80.07**		81.03
POS Unigrams	**77.39**	77.05	76.97	75.51	79.62
Function Words	77.42	**77.53**	77.19	76.94	79.50
Character Bigrams	**76.49**	75.53	74.32	76.38	79.84
POS Bigrams	76.60	75.98	75.87	**77.25**	78.88

Table 6-13: Accuracy of top 20 of the single type base features sets in order as per the gain ration, information gain, original and J48 ranking algorithms for two age classes (teens and adults)

The character bigrams also showed a significant, although not substantial difference between the full base feature set of 80 features and the four feature sets consisting of 40 features (*p* < 0.07), however, an analysis of variance test undertaken on the four half sized character bigram feature sets showed that there was no significant difference within the means (*F*(3,36) = 0.9186, *p* = 0.4416). Comparing the results of the five feature types within each ranking method, there was no significant difference within any of the ranking methods except the J48 tree method. When the features were ranked with the J48 tree, LIWC was significantly better than the other four feature sets (*p* < 0.02).

The feature set sizes were further reduced to the top 20 features under each ranking algorithm. Table 6-13 shows the results, with the most accurate result for each feature type bolded and shaded and the least accurate results cross hatched. Once again, while the difference between the quarter sized feature and the base feature set was statistically significant, it was not substantial, with reductions of between 1.42% (LIWC) and 4.87% (character bigrams). When the best performed ranking for each of the feature types were compared, the LIWC feature set significantly outperformed the others (p < 0.001). There was, however, no one method that gave a better ranking for all five feature types.

Because there was no substantial reduction in accuracy, the feature sets were further halved, to 10 features per feature set. The results, shown in Table 6-14, again show a very slight reduction in accuracy, even though the feature sets contain only one eight of the features contained in the base feature sets. The fall in accuracy was between 4.05% (character bigrams) and 1.67% (LIWC).

Feature Types	Ranking Algorithms: 10 Features per Feature Set				Base Feature Set (70 or 80 Features)
	Information Gain	Gain Ratio	J48 Tree	Original Ranking	
LIWC	77.79	77.48	**79.36**		81.03
POS Unigrams	75.62	75.73	74.38	**76.49**	79.62
Function Words	75.53	**75.90**	75.00	75.00	79.50
Character Bigrams	**75.79**	75.70	70.21	75.45	79.84
POS Bigrams	75.00	75.39	74.46	**76.18**	78.88

Table 6-14: Accuracy of top 10 of the single type base features sets in order as per the gain ration, information gain, original and J48 ranking algorithms for two age classes (teens and adults

Of the 19 different ranking/feature type cells in Table 6-14, only three have an accuracy of less than 75%. These occur for the POS unigrams, the character bigrams and the POS bigrams when each of them is ranked using the J48 algorithm. LIWC consistently shows the smallest drop in accuracy across all ranking algorithms, however the most effective for LIWC is the J48 algorithm.

Since reducing the feature set size to ten features still did not result in a substantial drop in accuracy, the top five features from each of the five feature sets, ranked with the four different algorithms were tested. Table 6-15 shows the results. When the five feature types were ranked using the different algorithms, there was a great deal of agreement with the top five features. The information gain and gain ratio gave the same top five features for the LIWC, character bigram and POS bigram feature types, with only one feature different in the function word and POS unigram feature types. The J48 tree ranking had the same top two features for the LIWC feature set, two out of the top five features were the same for the function word feature set, three out of five for the POS bigram and POS unigram feature sets and four out of five features were the same for the character bigram feature set. The original ranking algorithm had two out of the top five the same as the information gain for the POS unigram feature set and three out of five the same for the function word and POS bigram features sets, but no overlap for the character bigram feature set.

Feature Types	Ranking Algorithms: 5 Features per Feature Set				Base Feature Set (70 or 80 Features)
	Information Gain	Gain Ratio	J48 tree	Original Ranking	
LIWC	76.52	76.52	**77.00**		81.03
POS Unigrams	75.70	75.65	73.37	**76.04**	79.62
Function Words	73.39	73.39	**74.72**	73.96	79.50
Character Bigrams	**74.86**	**74.86**	62.87	73.48	79.84
POS Bigrams	74.66	74.66	73.82	**75.87**	78.88

Table 6-15: : Accuracy of top 5 of the single type base features sets in order as per the gain ration, information gain, original and J48 ranking algorithms for two age classes (teens and adults

At five features per feature set, only the LIWC feature set achieved more than 75% accuracy for all ranking algorithms. The best ranking algorithm for the both POS unigrams and POS bigrams was the original ranking algorithm, which also achieved more than 75% accuracy. The fall in accuracy when the most effective algorithm was compared with the base feature set ranged from 5.02% (character bigrams) to 3.58% (POS unigrams). The LIWC feature set again had the highest accuracy, although not significantly higher than the POS unigrams (p = 0.2176) or the POS bigrams (p = 0.1097), LIWC was significantly more accurate than the function word and character bigram feature sets (p < 0.019). Whereas with the 10 feature sets, there was a marked difference in the fall in accuracy between the most and least accurate feature type, when the features were reduced to only five features per set, the reduction in accuracy was more uniform, although still very insignificant when the percentage drop in feature set size is considered.

Figure 6-6 gives a summary of the reduction in accuracy for all feature types as the number of features per feature set reduces. The data used for each feature size/feature type combination is the highest accuracy for that combination. E.g.: the J48 tree ranking method gave the best accuracy for LIWC at 5 features so that is the data used, however the information gain ranking method gave the most accurate results for the function word features, so that is the data point represented. As can be seen from Figure 6-6, LIWC gives significantly more accurate results for each reduction in feature set size with the exception of the five feature set, where it is more accurate but not substantially so. The other feature types are not significantly different from each other at each level.

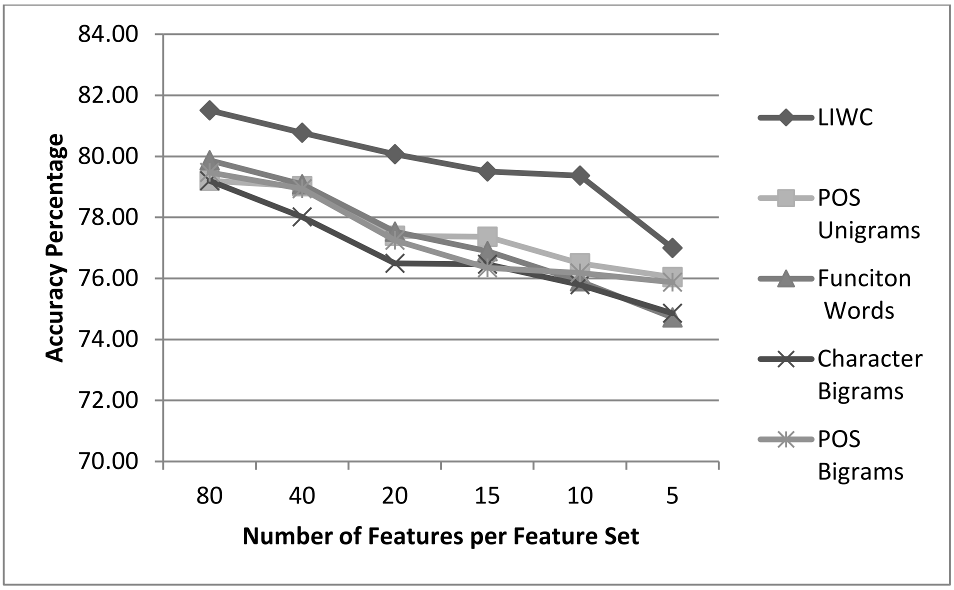

Figure 6-6: Summary of the effect of reducing feature set sizes from 80 features to five features per feature set for age group classification on two classes (teens and adults)

The results detailed in this section appear to imply that, similar to the gender classifications, just a few features have a very strong impact on the accuracy of a given feature set. To test this, each of the top five features from each ranking algorithm for each feature set were tested individually. This meant that each classifier had just one feature.

LIWC		POS Unigram		POS Bigrams		Function Words		Character bigrams	
Feature	%	Feature	%	Feature	%	Feature	%	Feature	%
assent	72.83%	DT	72.22%	NN_IN	72.41%	the	68.41%	ti	66.98%
article	71.73%	IN	71.54%	IN_DT	72.16%	of	68.02%	in	66.47%
preps	70.92%	UH	68.64%	DT_NN	70.95%	in	67.68%	on	65.20%
sixltr	68.13%	VBN	68.13%	DT_JJ	70.33%	go	66.47%	so	65.20%
affect	66.77%	NNS	66.44%	NNS_IN	70.24%	so	66.24%	sp g	64.89%
filler	62.53%	excl	60.47%	unk_NN	63.10%	a	65.01%	sp o	64.22%
insight	54.79%	st	59.88%	NNS_st	60.22%	like	63.15%	ar	63.68%
comma	54.42%	unk	59.43%	NN_st	55.21%	then	60.64%	sp-a	57.91%
		NN	58.98%			been	55.46%	sp-b	55.88%
		comma	56.81%			to	54.70%	sp-c	55.49%
		PP	56.28%					sp-f	55.12%
								sp-d	54%

Table 6-16: Summary of accuracy rates of individual features of the top five features from each ranking algorithm for the five single type feature sets for age group classification on two classes (teens and adults)

The results for the single feature classifiers are given in Table 6-16. LIWC and the POS bigram feature sets had the least amount of variation, with only eight features in the top five position across all four ranking algorithms, the character bigrams had the most variation with twelve features in the top positions, POS unigrams and function words had eleven and ten features in the top positions across the four algorithms, respectively.

There is a commonality across the different feature types in the top five features. LIWC had three features that gave more than 70% accuracy, ‘assent’, ‘article’ and ‘preps’. The ‘assent’ category contains 30 words such as ‘agree’, ‘ok’, ‘yes’. The ‘article’ category only contains three words: ‘a’, ‘an’ and ‘the’. The ‘preps’ category refers to prepositions and contains 43 words such as ‘on’, ‘to’, and ‘from’. The ‘filler’ category has nine entries that are conversational fillers, such as ‘blah’, ‘you know’, ‘I mean’. The ‘insight’ category is part of the larger psychological process category and includes 195 words such as ‘think’, ‘know’ and ‘consider’. The ‘comma’ category is part of the punctuation category and counts the number of commas in the text. The POS bigrams feature set had five individual features that scored more than 70% accuracy. All of these features contain prepositions (PP), determiners (DT) or nouns (NN for singular nouns and NNS for plural nouns). POS unigrams had two features that gave over 70% accuracy; the POS unigrams ‘DT’ and ‘IN’. The next most accurate POS tag ‘UH’ indicates an interjection (words such as ‘uh’, ‘yeah’, ‘um’). The label ‘unk’ indicates unknown words in the corpus. These would include slang, misspelled words and foreign words. These tags only accounted for 59.43% of the classification in the POS unigrams, but did score 63.10% for the POS bigrams, however in the POS bigrams, the unknown word is teamed with a noun, so it is more likely it is an English unknown word rather than a foreign word. The ‘comma’ has the same meaning as for the LIWC. The function words did not have a single feature that gave more than 70% accuracy, however the most accurate word was ‘the’, which would be tagged as a determiner by QTag and included in the article category by LIWC, with the remaining features being prepositions or auxiliary verbs. The single feature character bigrams also all gave less than 70% accuracy. The most accurate was the bigram ‘ti’ followed by ‘in’ and ‘on’. The bigrams ‘in’ and ‘on’ are in themselves prepositions, but, of course are included in many other words.

The single feature classifiers that achieved above 70 % accuracy are all the type of feature that would be expected to differentiate between male and female communication (Newman et al., 2008; Pennebaker et al., 2004) rather than teen and adult. These features include those that indicate agreement (such as the LIWC ‘assent’) and features that indicate a more objective use of language (the POS ‘DT’ and the LIWC ‘article’). However Pennebaker and Stone (2003) found that speech patterns moved towards more ‘male’ features as the speakers’ age increased tending toward less subjective and more objective use of language. They also found that older speakers used more sophisticated language constructs including more complex sentences and longer words. The LIWC feature ‘sixltr’ which indicates the number of words with more than six letters present in the text, would measure this and also gave a very high accuracy when used on its own.

The feature set that consisted of a combination of features from all five feature types was the most accurate feature set when the full number of features was present. The features in this feature set were also ranked using the three ranking algorithms used for the single type feature sets. The highest twenty, ten and five features from each ranking algorithm were compared for accuracy. When the individual feature types were reduced, there was no single ranking algorithm that was the most effective for all feature types at any feature set size. However, as can be seen in Figure 6-7, when all five feature types were combined and ranked using the three ranking algorithms, the J48 tree was the most effective in all three feature set sizes tested, although this difference was only significant in the feature sets consisting of ten features ($p = 0.0109$). The largest drop in accuracy as the feature set size reduced of 7.55% was in the features ranked by the gain ratio algorithm, the least drop in accuracy of 6.6% was in the features ranked by the J48 tree. The difference in accuracy between the features sets containing twenty features and the feature sets containing five features is only significant for the J48 listings ($p = 0.0158$). Although this appears to be an amazingly small reduction in accuracy considering the difference in feature set sizes, the data presented in Table 6-16 indicates that a great deal of the classification results can be accounted for by a very few individual features.

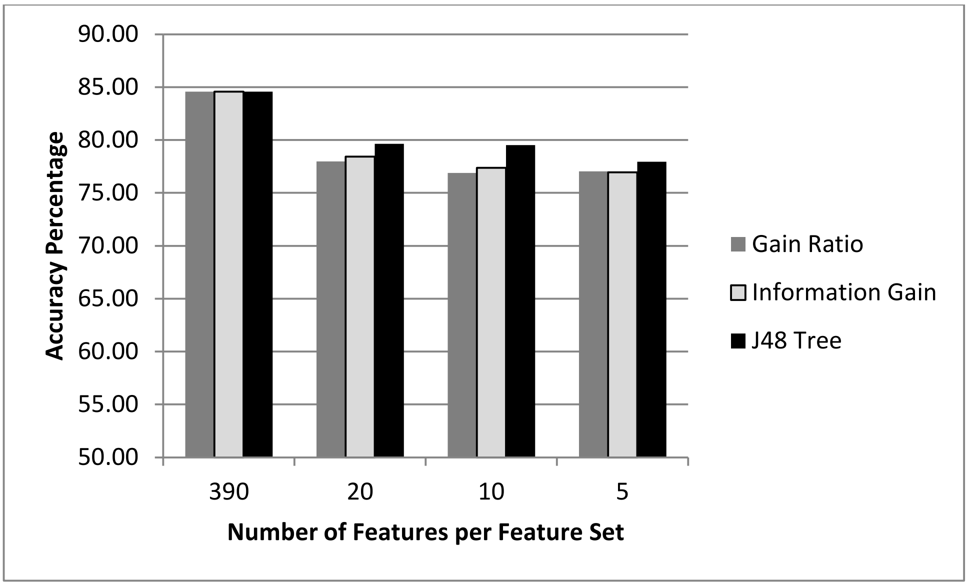

Figure 6-7: Effect on accuracy of reducing the number of features in the five way combination feature type for classification on two age classes (teens and adults)

Table 6-17 gives an analysis of the individual feature types present in each of the combination feature sets. The most effective ranking, that resulting from the J48 algorithm contains a higher number of LIWC features than the other two listings at all feature sets sizes.

Ranking Method	Feature Set Size	Feature Type				
		LIWC	POS Unigrams	Function Words	Character Bigrams	POS Bigrams
Information Gain	20	4	5	4	1	6
	10	3	3	0	0	4
	5	1	2	0	0	2
Gain Ratio	20	4	4	3	2	7
	10	2	3	0	0	5
	5	2	2	0	0	1
J48 Tree	20	7	4	5	1	3
	10	4	2	3	0	1
	5	3	1	0	0	1

Table 6-17: Analysis of the number of each feature type in the combination feature sets at different sizes and under different ranking algorithms – age group classification for two classes (teens and adults)

Most of the features used in the top five combination feature sets for all rankings are present in Table 6-16 which gives the individual accuracy of all the features in the top five ranks of all ranking algorithms. The LIWC feature ‘assent’ was the only feature that was present in the top five features in all ranking algorithms. The feature that is not present in that table is the LIWC feature “anger” which is included in the top five features for the J48 ranking of the combination feature set. This feature was ranked eighth in the J48 ranking for the individual features, but ranked very low by all the other ranking algorithms. When this feature is used on its own in a classifier for age group with the two classes of teens and adults, it achieves 59.9% accuracy, not as much as many of the other individual feature classifiers, but still well above chance.

6.4. Results for Shortened Text for Age Group Classification

The reduced sized documents that were created for the gender classification experiments on shortened texts in Section 5.3 were also used for the experiments in this section. The feature sets need to be tested on shorter documents because many computer mediated communications, for example twitter, are considerably shorter than the blog documents that have been classified in the earlier selections of this chapter. In the case of twitter posts, messages are restricted to not more than 140 characters which on average equates to 28 words.

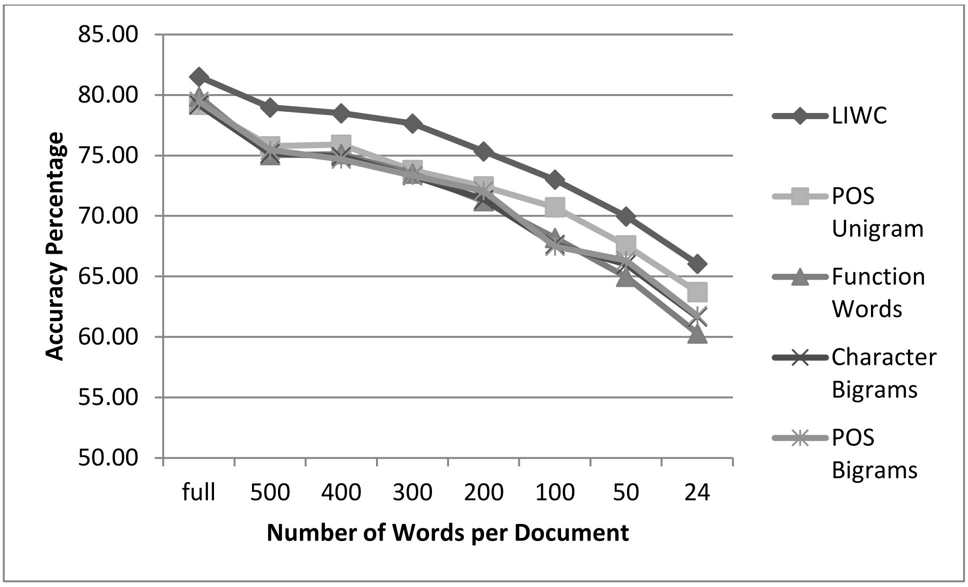

Figure 6-8: Effect of reducing document size on accuracy for single type features sets for classification of age group with two classes (teens and adults)

Classification exercises were undertaken on documents ranging in size from the full sizes available within the corpus down to documents of only 24 words. As in Section 5.2, the document were reduced in size to 500 words and then by increments of 100 words until the document size was 100 words, and then halved to 50 words per document and then reduced again, to 24 words per document. The feature sets used were the feature sets used as a base for the feature size reductions in Section 6.3, 80 each of the LIWC, function word, character bigram and POS bigram feature sets and the 70 present for the POS unigram feature set.

As was the case for the gender classification experiments, the accuracy fell as the document size decreased. The LIWC feature set was significantly more accurate at all document sizes (*p* < 0.009). LIWC also suffered the smallest fall in accuracy of 15.5% from full sized documents to 24 word documents, however this was only marginally less than the fall in accuracy for the POS unigrams, where the accuracy fell by 15.5%. The largest fall in accuracy was in function words, with a fall of 19.6%. The robustness and similar accuracy reduction of LIWC and POS unigrams could be explained by the fact that they are both aggregate features. Each feature measures the incidence of a type that can include many words, rather than a single word or part of a word. The POS bigrams are also an aggregate feature, meaning that each individual feature consists of a number of words, but the combination of the two bigrams would lead to each POS feature being involved in many separate features, and possibly compromising their classification strength.

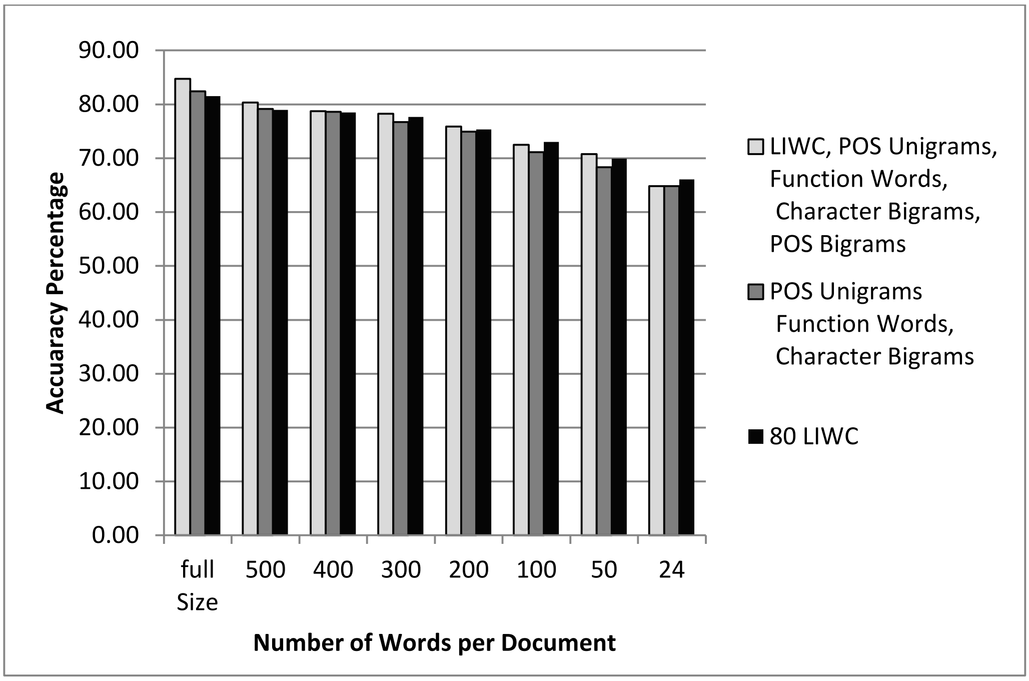

Figure 6-9: combination feature sets and LIWC with reducing document sizes for classification of two age groups (teens and adults)

Feature sets consisting of a combination of feature types have been shown to be more effective than features containing only one feature type in this Chapter. Two of the most effective feature combinations were selected for testing on the reduced size documents.

The two features sets chosen were the feature set that contained examples of all five feature types, and the feature set that contained four feature types not including LIWC. The reason being that when each feature set was removed to create five combination feature sets consisting of four feature types, the removal of the LIWC feature set had the most profound negative impact on accuracy, while removing any of the lexicographic feature sets had a statistically insignificant effect on the outcome of the classifiers. These two feature sets represent the two extremes in accuracy for the combination feature types. The results are given in Figure 6-9. LIWC is included, as the most successful single type feature set, for comparison purposes. At full sized documents the five way combination feature set is statistically more accurate than the single type LIWC feature set (*p* = 0.0008) but not substantially so with the difference being only 3.2%. The four way combination feature set, the one that excludes LIWC is approximately midway between the two other results and is statistically less accurate than the five way combination (*p* = 0.0198) but is not statistically different from the LIWC feature set (*p* = 0.1604). As the document size decreases, the difference between the five way combination and the LIWC feature sets reduces until, at 300 words there is no significant difference between the three feature sets. Although there is no significant difference between the feature sets for the lower document sizes, LIWC is increasingly more accurate until at 24 word documents, LIWC is 1.2% more accurate and with a *p* value of 0.1392, it is over 80% certain that the result is not due to chance.

6.5. Shortened Feature Set and Document Sizes

The effect of simultaneously decreasing both the document size and the feature set size was then tested. The two most effective feature sets from Section 5.3 were chosen, the 390 five way combination feature set and the full LIWC feature set. The results from Section 5.2 showed that as the feature sets became smaller, the J48 algorithm was the most effective so that ranking was used in this experiment. Nine different feature sets were tested, sets consisting of the top 80, 40, 20, 10 and 5 features of the combined feature set, and sets consisting of the top 40, 20 10 and 5 features of the LIWC feature sets. The results for all the feature sets were remarkably similar.

Figure 6-10 shows the results for the full sized combination and LIWC feature sets, the top 80 features for the combination feature set, and the top 40 features and the top 5 features for both base feature sets. When this exercise was undertaken on the first language classification on the ICLE corpus, the smaller feature sets/document sizes showed a smaller overall drop in accuracy than the larger feature sets/document sizes. For the gender classification, there was no observable difference in the reduction in accuracy of the larger verses the smaller feature sets. For the age group classification there was also essentially a uniform decrease in accuracy as the document sizes decreased from full sized to 24 words, of between 19.73% (390 combination features) and 15.5% (80 LIWC features). There was also a uniform decrease in accuracy as the size of the feature set decreased from 7.6% (full sized documents) to 5.5% (100 word documents).

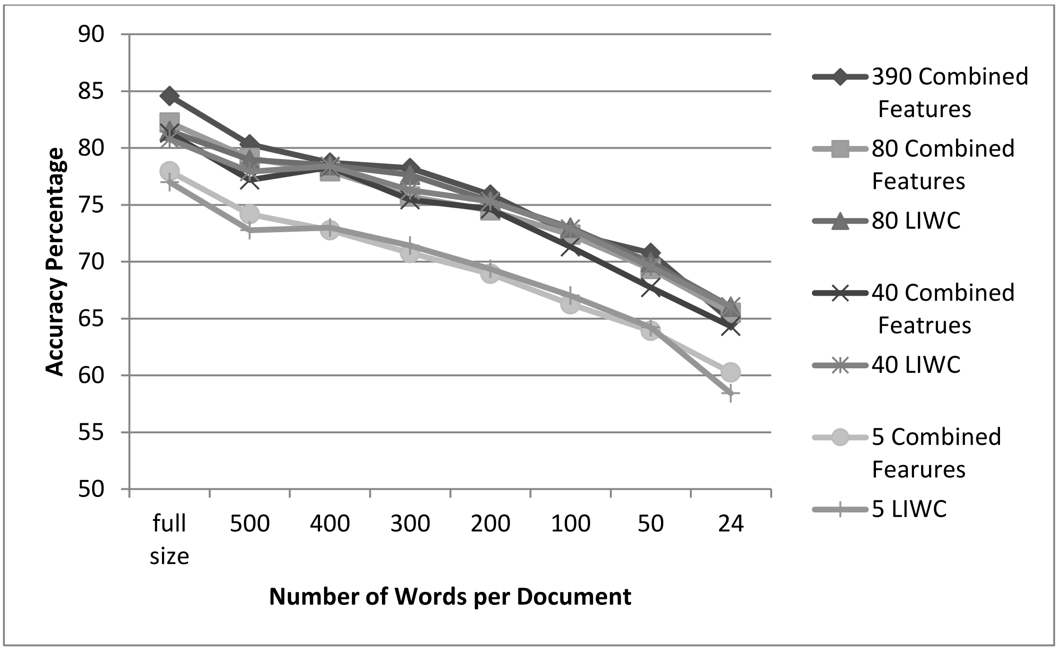

Figure 6-10 accuracy as number of features and document size decreases

There was statistically no difference between the five feature sets that are grouped together at the top of the graph, (ie the 390 combination feature set, both 80 sized feature sets and both 40 sized feature sets). There was also no statistical difference between the two 5 sized feature sets. An examination of the features present in both of the five sized feature sets, shown in Table 6-18 shows that two of the five features in the combination feature set are also present in the LIWC feature set (assent and comma).

The most accurate feature in the LIWC only feature set was the “assent” feature for the larger three document sizes, and the “article” feature for the remaining five document sizes. In the combination feature set, the most accurate single feature was the LIWC feature “assent” for all but the two smallest document sizes, and the POS bigram consisting of a singular noun and a preposition was the most accurate for the two smallest document sizes. The “anger:” feature of LIWC was interesting. Colloquially, it is more socially acceptable for men to express anger than women, and almost expected that teens will display stronger extremes of emotion more readily than adults. However when this feature was tested alone, it gave considerably less accurate results than the best features, especially in the smaller document sizes.

Type of feature		Number of words per document							
		full size	500 words	400 words	300 words	200 words	100 words	50 words	24 words
LIWC only	article_L	71.73%	68.27	68.19	66.67	65.32	61.54	58.5	57.21
	insight_L	54.79%	53.35	51.77	52.76	52.82	51.83	51.01	50.87
	assent_L	72.83%	68.69	69.57	65.99	64.64	60.84	57.8	53.46
	filler_L	62.53%	58.05	56.56	57.35	55.43	52.79	51.35	49.85
	comma_L	54.42%	52.67	50.93	52.59	50.76	50.87	50.23	49.61
+Combination feature set	assent_L	72.83%	68.69	69.57	65.99	64.64	60.84	57.8	53.46
	anger_L	59.94%	58.39	56.84	56.25	56.02	55.24	52.28	51.01
	comma_L	54.42%	52.67	50.93	52.59	50.76	50.87	50.23	49.61
	???_U	59.43%	59.29	58.9	58.28	59.4	58.61	56.98	55.72
	NN_IN_P	72.41%	67.79	65.91	64.58	62.81	57.74	59.01	54.87

Table 6-18: Results for classification by individual features on reduced sized documents for age group classification over two classes (teens and adults)

6.6. Discussion

The classification of first language exercises performed on the ICLE corpus (Chapter 4) and the gender classification experiments (Chapter 5) showed that LIWC improved the accuracy of text characterisation. The same effect had been expected for the age group classification. However when the subset of the Blog Authorship Corpus being studied was classified using the three original age classes (teens, twenties and thirties) the accuracy much lower than had been expected. The results in Chapter 4 also showed that combining different types of features (with or without LIWC) also improved the accuracy in first language text characterisation exercises. However, when the original three age classes were used for the Blog Authorship Corpus, the combination and standalone feature sets gave equally poor results.

The classification exercises were repeated on the teen/adult corpus. As for the gender classification, the LIWC feature sets was significantly more accurate than similar numbers of the lexicographic feature sets. When LIWC was added to large numbers of other features, it consistently raised the accuracy by a significant amount and when similar numbers of the five feature sets were combined, the removal of LIWC resulted in a significant reduction in accuracy where the removal of any one of the other feature sets did not.

The assent feature measures the number of times agreement words such as 'yes',' ok', 'mmhmm' are used. It covers 18 different words and fillers. In the full sized documents, over 77% of the corpus had some type of assent word counted, however 88.2% of teens had a value for the assent feature, while only 67.4% of the adults did. The majority of the larger values for assent were for teens rather than adults. The distribution of the article feature in the 24 word documents is not as clear cut. There were again just over 77% of the corpus that had a value for the article features, 72.0% of teens and 82.2% of adults. However there were no clear groupings of teens or adults in the higher or lower values for the feature. At 24 words there were 145 documents that had a value for the assent feature, but no value for the article feature and 2429 documents that had a value for the article feature, but no value for the assent feature. There were 668 that had no value for either feature, 18.8% of the corpus.

Creating a corpus by randomly selecting chunks of words is an artificial method of acquiring short texts. In a genuine conversation of short utterances, many words would be repeated, there would be utterances lacking verbs, or nouns, and an absence of grammar that would appear in even casual writing such as a blog. A further test for these features would be to acquire transcripts of real world chats and see if they are as effective in classifying teens and adults age groups as with the amended blog corpus. While real world chats may have been more faithful to real world applications, there is currently no corpora of this kind available and to create one with the correct parameters would be prohibitively expensive in time, money and computer power.

Chapter 7 Conclusion

This thesis has examined the effectiveness of a psycholinguistically based feature set, LIWC, for authorship characterisation on three demographic attributes of an author, their first language, gender and age group. The effectiveness was compared with four of the lexicographic features that have been previously used in authorship characterisation; function words, character bigrams, POS unigrams and POS bigrams. The remainder of this chapter will reiterate each of the research questions stated in Chapter 1, and discuss the relevant research outcomes. It will also examine the limitations of the study and discuss application of the results of this study and possible future research directions.

7.1. Research Questions and Contributions

7.1.1. Are psycholinguistically based features more effective than lexicographical features for authorship characterisation?

The experiments detailed in this thesis support the hypothesis that the psycholinguistically based feature set tested (LIWC) was more effective than the lexicographic features with which it was compared for authorship characterisation. The experiments to examine this were undertaken using three demographic characteristics of the authors: first language, gender and age group. Initially the accuracy of approximately equal numbers of each of the five feature sets was compared.

The corpus being classified for the first language experiments was the ICLE corpus which consists of sixteen first language groups. The LIWC feature set achieved a significantly higher average accuracy for this demographic characteristic than the lexicographic feature sets

The gender classification experiments were conducted on the Authorship Blog Corpus. LIWC was again significantly more accurate than each of the feature sets consisting of a similar number of the lexicographic features.

The age group classification experiments were also conducted on the Authorship Blog Corpus. The corpus in its original configuration, has three age group classes (teens, twenties and thirties) however investigation revealed that dividing the corpus in to two age groups, teens and adults gave a stronger separation of the classes. LIWC achieved significantly more accurate results in the two class age group classification than any of the similar sized lexicographic feature sets.

7.1.2. Is the theoretical basis for psycholinguistic features sufficiently different from that of lexicographical features that the combination of psycholinguistic features with lexicographical features will be significantly more effective in authorship characterisation than equal amounts of lexicographical features alone?

The experimentation undertaken for this thesis also supports the hypothesis that the LIWC feature set is sufficiently different to the lexicographic feature sets that its addition to a lexicographic feature set increases accuracy significantly more than the addition of a similar number of additional features of the same feature type. It was also found that in a feature set consisting of multiple feature set types, the removal of LIWC had a significant negative impact on the accuracy, at least equal to the impact of the removal of any other feature set type being tested.

LIWC was added to feature sets consisting of 200 single type lexicographic features and the accuracy compared to 280 features of the same type. The six feature sets were used to classify the ICLE corpus into the first language of the author. The accuracy for the feature set that included LIWC was significantly higher than for the corresponding single type feature set in all cases. This experiment was repeated using base feature sets of 600 features. The addition of LIWC significantly increased the accuracy of the function word and POS bigram feature sets when compared to the accuracy of the relevant feature set consisting of 680 single type features. The accuracy also increased for the feature set consisting of LIWC and 600 character bigrams, but the increase was not significant at the 0.05 level.

LIWC was added to feature sets consisting of 200 single type features and the classification accuracy compared with the relevant feature set consisting of 280 single type features for gender classification. The feature sets containing LIWC were again significantly more accurate than the same number of the base feature type. Larger base feature sets were not used in these experiments as the inclusion of larger numbers of the base feature sets had no real impact on accuracy.

Similar experiments were conducted for age group classification. As for the single type feature classifications, the two age classes were used. LIWC was added to 200 single type features and the accuracy compared with 280 features from the base feature set. The inclusion of LIWC increased accuracy significantly more than increasing the base feature set by 80 features. LIWC was then added to 600 single type features and the accuracy compared with 680 single type features from the same feature type. Again, the addition of the LIWC features increase the accuracy significantly more than merely increasing the number of the base feature set by the same amount of features.

Combination feature sets are, in general more accurate than using similar numbers of features from a single feature type. To fully explore the impact of LIWC in combination feature sets, feature sets consisting of combinations of approximately equal numbers of all five feature sets. Then feature sets were created by systematically removing one of the feature types, leaving five different combinations of four feature types.

For the first language classification, the removal of LIWC had a significant negative impact on the accuracy of the classifier, as did the removal of the character bigram feature set. The feature set that had the least impact on the accuracy of the first language classifier was the POS unigram feature set.

When this experiment was repeated for the gender and age group classifications, the removal of the LIWC feature set lead to the largest fall in accuracy. This fall in accuracy was significantly greater than the removal of any other feature set for both classification experiments.

7.1.3. Do the number and/or type of features used in an authorship characterisation classification model have an effect on the success and accuracy of that model?

The question as to whether increasing numbers of the same feature type have a corresponding increase accuracy was also explored in this thesis. The answer appears to be dependent on the demographic characteristic being classified, and could be related to the number of classes present in the characteristic.

For the first language classification, the accuracy of the character bigram feature set increased with each incremental increase up to 600 features, and then remained constant up to 1000 features. The POS bigram feature set also continued to increase as the number of features increased, however the overall accuracy for this feature set was comparatively low. Increasing the number of function words did appear to increase the classification accuracy. However, inspection of the features revealed that there were an increasing number of topic specific words included in the list, which could have identified the topic discussed by a language group rather than idiosyncrasies of the first language itself.

The gender classification experiments did not show the same effect. Increasing any of the three feature sets had no significant effect after the threshold of 400 features was reached. The LIWC feature set was either significantly more accurate or not significantly less accurate than the larger sized feature sets for the gender classification.

The age group classification, again over two classes, showed no significant increase in accuracy after the threshold of 200 features for all three feature sets use. The 80 LIWC features again gave an accuracy that was not significantly different from any of the larger feature set sizes.

The fact that the feature sets became more accurate as the number of features they contained also increased for the first language experiments but not for the gender and age group experiments may be related to the number of classes in each demographic characteristic. The first language classification was over sixteen classes, while the gender and age group were only over two.

7.1.4. Is there an effective lower limit to the number of features that can be used for a classification model for authorship characterisation?

As feature sets become larger, they also become computationally unwieldy and prone to creating noise in the classification rather than insight. Irrelevant attributes also create an negative effect on the classification process so it is common to eliminate all but the most relevant attributes (Witten & Frank, 2005). The question of how much feature sets can be reduced, while still maintaining a reasonable level of accuracy was also investigated in this thesis. The feature sets used in these experiments were the five single type feature sets consisting of 70 (POS unigrams) or 80 features and the 390 sized feature set that consisted of these five single type feature sets combined. The feature sets were ranked by three of the feature selection algorithms that are supplied with the WEKA suite, the information gain, the gain ratio and the J48 tree algorithms. The answer to the question again appears to be dependent on the number of classes within the classification, however for two classes the effective lower limit is exceptionally low, and if the correct feature is used, one feature can achieve a reasonably accurate result.

Information gain was consistently the most effective ranking algorithm for the first language classification across all the feature sets examined. The 390 feature set was reduced to the top 80 features, giving it the same or similar size to the other feature sets. The combination feature set was the most effective of the 80 sized feature sets. However as the feature set sized was reduced, the LIWC feature set became more competitive, until at 20 features, there was no significant difference between the combination feature set and LIWC. The lexicographic single type feature sets were significantly less accurate than LIWC at all feature set sizes. An inspection of the feature types making up the combination feature set showed that information gain, the more effective ranking, had a disproportionate number of LIWC features present. The accuracy fell with each decrement of the feature set size, but not by a proportional amount, and at 20 features, one quarter the size of the base feature sets, the accuracy was still above 39%.

When the feature set sizes were reduced for the gender classification, there was no significant difference in the accuracy of the half sized and full sized feature sets. There was also no one ranking algorithm that was consistently more effective across the feature set types and sizes. The feature sets were reduced to one quarter of the full sized feature set, to 20 features and, while there was a significant fall in accuracy, the fall was not substantial, being less than 3% for all feature types. The feature sets were further reduced to ten and then to five features per feature set. At each reduction, there was a significant reduction in accuracy, but not substantial, until at five features, LIWC, the most effective feature set at all feature set sizes, achieved less than 6% lower accuracy than the full sized 80 features. Further investigation revealed that almost 65% of the accuracy could be accounted for by one LIWC feature, that of personal pronouns. In fact there were a number of features, across all the feature types that achieved accuracy greater than 60% when used for classification individually.

Reducing the feature set sizes for the age group classification gave similar results to that of the gender classification. There was no one ranking algorithm that was consistently superior across all feature set types and sizes. There was also negligible reductions in accuracy as the number of features included in each feature set fell, although LIWC was the most accurate feature set at all feature set sizes. The fall in accuracy from the full sized feature sets to feature sets consisting of only five features was again less than 6%. Individual features were tested as classifiers, and, again similar to the gender classification experiments, several features achieved a greater than 70% accuracy. The most accurate individual features were the LIWC feature “assent”, the POS unigram feature “DT” (determiner) and the POS bigram feature NN_IN (noun followed by preposition). There were no function words or character bigram features that achieved more than 70% accuracy as individual features.

7.1.5. Is there an effective lower limit to the number of words in a text that can be classified using authorship characterisation techniques?

The classification and identification of authors of short texts is becoming more and more relevant to the area of text categorisation with the increasing use of shorter forms of electronic communication such as SMS and other Computer Mediated Communications (CMCs) such as twitter, emails and chat rooms. Again the answer as to the effective lower limit for text size would appear to be dependent on the number and type of classes being classified.

As the number of words per document fell for the first language classification, the accuracy also decreased, until at documents of 24 words, the classification accuracy, although much higher than chance, was still lower than 20%. As the document size decreased, the LIWC feature set became comparatively more accurate, until at 100 words there was no significant difference between it and the 390 sized combination feature set.

For the gender classification experiments, a similar pattern was observed. The accuracy fell as the document size reduced, but although the reduction was significant, it was not substantial, and the accuracy for the LIWC feature set was less than 14% different between the full sized documents and the 24 word documents. The LIWC feature set was not significantly less accurate than the combination feature sets at any document size.

Again, for the reducing document size experiments, the age group classification showed similar results to the gender classifications. There was a significant reduction in accuracy for each decrement in document size, but not a substantial one. The LIWC feature set became comparatively more accurate as the document sizes reduced, with no significant difference between it and the combination feature set in documents of 400 words or less.

When the feature set size and the document size was reduced simultaneously for each of the demographic characteristics being examined, there was a slightly steeper fall in accuracy, but not substantially so.

7.1.6. Other Contributions

Torney, R., Vamplew, P., & Yearwood, J. (2012). Using psycholinguistic features for profiling first language of authors. *Journal of the American Society for Information Science and Technology, 63*(4) doi:10.1002/asi.22627

7.2. Applications of Results

The internet has brought many benefits, not the least of which is more open and free access to information and communication. However this new, open access has brought with it the problems of keeping vulnerable individuals, such as children and minors, safe from both inappropriate sites and dangerous individuals.

There are many sites on the internet, that while perfectly legal, are not appropriate for underage individuals. A study found that up to 75% of teenagers aged between 16 and 17 years had been exposed to pornography accidentally (Shirali-Shahreza & Shirali-Shahreza, 2008). While there are adult content filtering programs, that try to identify adult content and block access, they remain client side methods. Most adult sites also have warnings that they contain explicit materials and the user has to confirm that they are over the legal age, but a child can do this and gain access to the materials. In the real world, an individual must prove they are an adult by producing some form of identification to gain access to restricted goods or services, such as a driver's licence (Shirali-Shahreza & Shirali-Shahreza, 2008).

Underage individuals wandering onto an inappropriate site is one problem, but a more pressing one is individuals actively targeting children and teenagers in supposedly safe sites to induce them into dangerous and inappropriate acts. Paedophiles groom victims in chat rooms, often by posing as a peer of the victim, that is the same age group and gender, and befriend them (Berson, 2003; Gupta et al., 2012). Currently, these predators are tracked and caught by law enforcement personnel posing as minors in the chat rooms, however there are far too many instances of grooming for them to be able to make an impact.

If a filter could be designed with features that were able to identify the age group of a person before they could access restricted materials on line, or identify the age and gender of an individual misrepresenting them, this would advance the protection of minors from both explicit content and online predation. In this study, such features were discovered that gave an accuracy of 81.03% with one feature set (LIWC) and 84.7% with a combined feature type feature set.

Individuals also misrepresent other information, such as their country of origin to perpetrate scams such as romance scams (Rege, 2009; Wolak, Finkelhou, Mitchell, & Ybarra, 2008) where the perpetrator fleeces the victim of large amounts of money with bogus promises and situations. Again a filter that could flag misrepresentation of first language could aid in the identification and prevention of such criminal activity. In this study, LIWC proved to be adept at identifying the first language of an author of an essay from sixteen classes with an average accuracy of 50.13%, and up to 79% for specific languages. The accuracy increased as numbers of other features were included.

7.3. Limitations

The experiments conducted for this research only used one corpus for each of the demographic characteristics examined. While all feasible actions were taken to prevent it, it is possible that the results obtained have been influenced by idiosyncrasies within the corpora being studied. The corpus used for the first language classification in particular consists of essays written by students of English. Each language group consists of one or more cohorts who were instructed to write on a particular topic, this making it more likely that any content specific words could identify the language group. However, at the tme of this study, the ICLE was the only corpus available and since it has been compiled under strict conditions specifically for the purpose of studying non-native English speakers English language skills it was considered appropriate.

The age range of the corpus used for the age group classification was limited to teens to late thirties. The age and gender of this corpus is also self-reported by the authors. However, given that the corpus consists of over 19,000 individual authors, it is unlikely that any misrepresentation of demographic data would be pervasive enough to skew the results. This corpus has also been used in a number of previous studies into authorship attribution and characterisation. Previous studies into the impact the age of the author has on language use have used a corpus with far wider age ranges (Pennebaker & Stone, 2003). However, the corpora used in those studies were not available for this research.

The studies into the classification efficacy of shorter texts used artificially shortened documents. The results were promising, however the experiments should be repeated on genuine short texts to test the accuracy of the feature sets in a real world situation. Other studies into the classification of short texts have used Twitter feeds as their corpus. However the age group and gender of the Twitter author is not tagged in these corpora, and has to be manually assessed and marked up by the researcher. This was not considered an effective method for the probable veracity of the end results.

7.4. Future Work

7.4.1. Research Directions

This study examined the effectiveness of combinations of function words, character bigrams, POS bigrams, and LIWC features for the identification of the first language, gender and age group of an author of English texts. Of particular interest would be examining corpora of short communications to ascertain the effectiveness of the features on genuine short communications rather than artificially shortened texts. Future research may concentrate on the development of suitable short text corpora that have reasonable accuracy in the allocation of age group and gender. Development of a corpus consisting of a wider age range may also be considered for further research in this area.

Future studies will also examine the effectiveness of the features sets examined on other corpora to ascertain if the results are consistent across data sets. An extension of the research into first language could include seeking a set of features that would identify the speaker of a particular first language rather than attempting to sort a corpus into multiple first language groups.

The shortened text experiments lead to feature vectors with exceptionally sparse data points. Further research could be conducted into methods of working with and improving the results for such feature vectors.

7.4.2. Applications of the Research

The results detailed in this thesis have shown that, with a surprisingly small feature set, it is possible to accurately determine the age and/or gender of an author. This could be of immense value in the detection of on-line predators that frequently misrepresent their age or gender or both to gain the trust of their victims. These results could also be of use in the restriction of sites or content to users in an inappropriate age range, either to prevent underage access to adult sites, or to prevent predators trawling on youth specific chat rooms.

References

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE Volume 20*(Issue 5), 67-75. doi:10.1109/MIS.2005.81

Abbasi, A., & Chen, H. (2006). Visualizing authorship for identification. In S. Mehrotra, & et al (Eds.), *ISI 2006, LNCS 3975* (1st ed., pp. 60-71). Berlin Heidelberg: Springer-Verlag.

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems, Volume 26*(Number 2, Article 7) doi:10.1145/1344411.1344413

Abu-Jbara, A., Jha, R., Morley, E., & Radev, D. (2013). <br />Experimental results on the native language identification shared task. *<br />Proceedings of the Eighth Workshop on Innovative use of NLP for Building Educational Applications,*

Adams, S. H. (1996). Statement analysis. *FBI Law Enforcement Bulletin, 65*(10)

Alison, L., Smith, M. D., & Morgan, K. (2003). Interpreting the accuracy of offender profiles. *Psychology, Crime and Law, 9(2)*(2), 185-195.

Andreou, G., Karapetsas, A., & Galantomos, I. (2008). Modern greek language: Acquisition of morphology and syntax by non-native speakers<br />. *The Reading Matrix, 8*(1), 35-42.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). Lexical predictors of personality type. *Literary and Linguistic Computing, 17*(4), 401-412.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday, 2011*(5th April), 4th May 2011. doi:http://dx.doi.org/10.5210/fm.v12i9.2003

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, Volume 52*(No 2), 119-123. doi:10.1145/1461928.1461959

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. *Proceedings of the 2005 ACH/ALLC Conference,* Victoria, BC, Canada. 1-3.

Argamon-Engelson, S., Koppel, M., & Avneri, G. (1998). Style-based text categorization: What newspaper am I reading? *Proceedings of the AAAI Workshop of Learning for Text Categorization,* , 1-4.

Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. *6th JADT Des Joronees Internationalres D'Analys Statistique Des Donnes Textuelles,* 29-37.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135-160.

Berson, I. R. (2003). Grooming cybervictims. *Journal of School Violence, 2*(1), 5-18. doi:10.1300/J202v02n01_02

Bhargava, M., Mehndiratta, P., & Asawa, K. (2013). Stylometric analysis for authorship attribution on twitter. *Big data analytics* (pp. 37-47) Springer.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A CORPUS OF NON-NATIVE ENGLISH. *ETS Research Report Series, 2013*(2), i-15.

Bogdanova, D., Rosso, P., & Solorio, T. (2013). Exploring high-level features for detecting cyberpedophilia. *Computer Speech and Language, 28*, 108-120. doi:http://dx.doi.org/10.1016/j.csl.2013.04.007

Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive Psychology, 43*(1), 1-22.

Brauer, J., Anwander, A., & Friederici, A. D. (2011). Neuroanatomical prerequisites for language functions in the maturing brain. *Cerebral Cortex, 21*(2), 459-466.

Canter, D. (2004). Offender profiling and investigative psychology. *Journal of Investigative Psychology and Offender Profiling, 1*, 1-15.

Casey, B. J., Galvan, A., & Hare, T. A. (2005). Changes in cerebral functional organization during cognitive development. *Current Opinion in Neurobiology, 15*, 239-244. doi:DOI 10.1016/j.conb.2005.03.012

Chaski, C. E. (2008). *Text and pretext on the internet: Recognizing problematic communications*. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/ALIAS_ISTTF_Submission.pdf (10March09):

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

Christensson, P. (2006). Definition of cybercrime. Retrieved from http://www.techterms.com/definition/cybercrime

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu.Rev.Psychol., 55*, 591-621.

Clarke, R. V. (1994). *Crime prevention studies volume 2* Willow Tree Press. doi:ISBN 1-881798-01-1.

Cockburn, N. B. (1998). *The bacon shakespeare question: The baconian theory made sane.* Guildford and Kings Lynn: Biddles Limited.

Cohn, M., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding september 11. *American Psychological Society, 15*(10), 687-693.

Collins English Dictionary - Complete and Unabridged. (2003). **fraud. (n.d.)** . Retrieved from http://www.thefreedictionary.com/fraud

Connor, U. (1996). Contrastive rhetoric: Cross-cultural aspects of second-language writing. In M. H. Long, & J. C. Richards (Eds.), (1st ed., pp. 5-11). Cambridge, United Kingdom: Cambridge University Press.

Corney, M. (2003). *Analysing E-mail text authorship for forensic purposes*

Corney, M., Anderson, A., & Mohay, G. (2002). Gender-preferential text mining of E-mail discourse. *18th Annual Computer Security Applications Conference, Las Vegas, NV.*, Las Vegas, NV.

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner english: The nus corpus of learner english. *Proceedings of the Eighth Workshop on Innovative use of NLP for Building Educational Applications,* 22-31.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. *Proceedings of the 21st International Conference on World Wide Web,* 699-708.

Daudaravicius, V. (2013). VTEX system description for the NLI 2013 shared task. *Proceedings of the Eighth Workshop on Innovative use of NLP for Building Educational Applications,* 89-95.

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining E-mail content for author identification forensics. *ACM SIGMOD Record, Volume 30*(Issue 4 (December)), 55-64.

Dombrowski, S. C., LeMasney, J. W., Ahia, C. E., & Dickson, S. A. (2004). Protecting children from online sexual predators: Technological, psychoeducational, and legal considerations. *Professional Psychology: Research and Practice, 35*(1), 65-73.

Douglas, J. E., Ressler, R. K., Burgess, A. W., & Hartman, C. R. (1986). Criminal profiling from crime scene analysis. *Behavioral Sciences and the Law, 4*(4), 101-421. doi:10.1002/bsl.2370040405

Edwards, L. (2012). Digital detecting: Tracking and identifying cyberaggression.

Eneman, M., Gillespie, A. A., & Bernd, C. S. (2010). Technology and sexual abuse: A critical review of an internet grooming case. *ICIS 2010 Proceedings,* Paper 144.

Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007). Author profiling for english emails. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics,* University of Melbourne, Australia. 263-272.

Farrell, G., Phillips, C., & Pease, K. (1995). Like taking candy: Why does repeat victimization occur. *British Journal of Criminology, 35*(3), 384-399.

Finegan, E., Blair, D., & Collins, P. (2000). *Language: Its structure and use* (2nd ed.) Thompson.

Fink, C., Kopecky, J., & Morawski, M. (2012). Inferring gender from the content of tweets: A region specific example. *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media,* Trinity College, Dublin, Ireland. 459-462.

Fradkin, D., & Muchnik, I. (2006). Support vector machines for classification. In G. Abello, & G. Cormode (Eds.), *Discrete methods in epidemiology, volume 70 of DIMACS series in discrete mathematics* (pp. 13-20) AMS, Providence, RI, USA, 2006. 6, 6.1.

Gamon, M. (2004). Linguistic correlates of style: Authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics,* 611.

Gardner, W. (2005). Just one Click—Sexual abuse of children and young people through the internet and mobile phone technology by tink palmer with lisa stacey, barnardo's, ilford, 2004. 37pp. ISBN 0-902046-99-9 (pbk),£ 5. *Child Abuse Review, 14*(6), 448-449.

Geeraertz, D., & Piersman, Y. (2011). Zones, facets, and prototype-based metonymy. In A. Barcelona, R. Benczes & Ruis de Mendoza Ibanez, F. (Eds.), *Defining metonymy in* *cognitive linguistics: Towards a consensus view* (pp. 89-102). Hillsdale, New Jersey, USA: John Benjamins Publishing Company.

Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., . . . Stein, B. (2013). Recent trends in digital text forensics and its evaluation. *Information access evaluation. multilinguality, multimodality, and visualization* (pp. 282-302) Springer.

Granger, S. (2001). *International corpus of learner english: The ICLE project*

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251-270.

Gupta, A., Kumaraguru, P., & Sureka, A. (2012). Characterizing pedophile conversations on the internet using online grooming. *arXiv Preprint arXiv:1208.4324,*

Hyman, P. (2013). Cybercrime: It's serious, but exactly how serious? *Communications of the ACM, 56*(3), 18-20. doi:10.1145/2428556.2428563

Isaacowitz, D. M., & Riediger, M. (2011). When age matters: Developmental perspectives on “Cognition and emotion”. *Cognition & Emotion, 25*(6), 957-967. doi:10.1080/02699931.2011.561575

Isard, A., Brockmann, C., & Oberlander, J. (2006). Individuality and alignment in generated dialogues. *Proceedings of the Fourth International Natural Language Generation Conference,* Sydney, Australia. 25-32.

Johnson, D., Malhotra, V., & Vamplew, P. (2006). More effective web search using bigrams and trigrams. *Webology, 3*(4)

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval, Volume 1*(number 3), 233-334. doi:10.1561/1500000005

Jurafsky, D., & Martin, J. H. (2000). In Horton M. (Ed.), *Speech and language processing*. Upper Saddle River, New Jersey, USA: Prentice-Hall Inc.

Kestemont, M. (2014). Function words in authorship attribution from black magic to theory? *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) at 14th Conference of the European Chapter of the Association for Computational Linguistics,* Gothenburg, Sweden. 59-66.

King, L., & Dickinson, M. (2013). Shallow semantic analysis of interactive learner sentences. *NAACL/HLT 2013,* , 11-20.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology, 60*(1), 9-26.

Koppel, M., Schler, J., & Argamon, S. (2010). Authorship attribution in the wild. *Language Resources and Evaluation, 45*(1), 83-94. doi:10.1007/s10579-009-9111-2

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship attribution with thousands of candidate authors. Paper presented at the *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, , August*(August) 659-660.

Koppel, M., Schler, J., & Zigdon, K. (2005). Determining an author's native language by mining a text for errors. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* Chicago, Illinois. 624-628.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management, 44*, 1448-1466.

Lagazio, M., Sherif, N., & Cushman, M. (2014). A multi-level approach to understanding the impact of cyber crime on the financial sector. *Computers & Security, 45*, 58-74. doi:10.1016/j.cose.2014.05.006

Layton, R., Watters, P., & Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second,* 1-8.

Lea, S. E. G., Fischer, P., & Evans, K. M. (2009). The economic psychology of scams. Paper presented at the *Joint Conference of the International Association for Research in Economic Psychology and the Society for the Advancement of Behavioural Economics,* Halifax, Nova Scotia, Canada.

Li, J., Zheng, R., & Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM, Volume 49*(Issue 4), 76-82. doi:http://doi.acm.org/10.1145/1121949.1121951

Lippi-Green, R. (1997). The myth of non-accent. *English with an accent: Language, ideology, and discrimination in the united states* (pp. 41-52) Routledge.

Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature selection: An ever evolving frontier in data mining. *Fourth Workshop on Feature Selection in Data Mining,* 4-13.

Ludu, P. S. (2014). *<br />Inferring gender of a twitter user using celebrities it follows.* Unpublished manuscript.

Luna, B., Garver, K. D., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child Development, 75 Number 5*(September/October), 1357-1372.

Luna, B., Padmanabhan, A., & O'Hearn, K. (2010). What has fMRI told us about the development of cognitive control through adolescence? *Brain and Cognition, 72*, 101-113.

Lunceford, B. (2009). Building hacker collective identity one text phile at a time: Reading phrack. *Media History Monographs, 11*(2), 1-26.

Luyckx, K., & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing, 26*(1), 35-55.

Ma, L., Torney, R., Watters, P., & Brown, S. (2009). Automatically generating classifier for phishing email prediction. *10th International Symposium on Pervasive Systems, Algorithms, and Networks*, Kaohsiung, Taiwan. 779-783.

Mala, T., & Geetha, T. V. (2007). Visualizing author attribution using blobby objects. *Proceedings of the Computer Graphics, Imaging and Visualisation **(CGIV 2007) - Volume 00***, Bangkok, Thailand. (August)

Malyutov, M. B. (2005). Authorship attribution of texts: A review. *General theory of information transfer and combinatorics* (Volume 4123/2006 ed., pp. 362-380). Berlin / Heidelberg: Springer. doi:10.1007/11889342

Mann, P. S. :. (2010). *Introductory statistics* (7th ed.). USA: John Wiley and Sons Inc.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students’ daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*(4), 857-870. doi:10.1037/0022-3514.84.4.857

Meina, M., Brodzinska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., & Wilk, M. (2013). Ensemble-based classification for author profiling using various features. *The Notebook for 2013 PAN at the Conference and Labs of the Evaluation Forum (CLEF),*

Mihalcea, R., & Nastase, V. (2012). Word epoch disambiguation: Finding how words change over time. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2,* 259-263.

Mosteller, F., & Wallace, D. L. (1963). Inferences in an authorship problem. *Journal of the American Statistical Association, 58,* 302-309.

Nelson, E. E., & Guyer, A. E. (2011). The development of the ventral prefrontal cortex and social flexibility. *Developmental Cognitive Neuroscience, 1*, 233-245. doi:doi:10.1016/j.dcn.2011.01.002

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*, 211-236.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personal Social Psychology Bulletin, 29*, 665-675. doi:10.1177/0146167203029005010

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). “How old do you think I am?”: A study of language and age in twitter. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Cambridge, Massachusetts, USA.

Ortega, L. (2009). *Understanding second language aquisition*. Oxford, Uk: Hodder Education.

Pallant, J. (2011). *SPSS survival manual* (4th Edition ed.). Crows Nest, NSW, Australia: Allan and Unwin.

Paradis, C. (2004). Where does metonymy stop? senses, facets, and active zones. *Metaphor and Symbol, 19*(4), 245-264.

Paradis, C. (2011). Metonymization: A key mechanism in semantic change. In A. Barcelona, R. Benczes & Ruis de Mendoza Ibanez, F. (Eds.), *Defining metonymy in cognitive linquistics: Towards a consensus view* (pp. 69-88). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Peersman, C., Daelemans, W., & Van Vaerenberg, L. (2011). Predicting age and gender in online social networks. *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, New York, USA. doi:10.1145/2065023.2065035

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin Texas: liwc.net:

Pennebaker, J. W., Groom, C. J., Loew, D., Dabbs, J. M., & Abbasi, A. (2004). Testosterone as a social inhibitor: Two case studies of the effect of testosterone treatment on language. *Journal of Abnormal Psychology, 113*(1), 172-175.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, a. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review Psychological, 54*, 547-577.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*(2), 291-301. doi:DOI: 10.1037/0022-3514.85.2.291

Platt, J. C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines.* (No. Technical Report MSR-TR-98-14 Microsoft Research).

Potha, N., & Stamatatos, E. (2014). A profile-based method for authorship verification. In A. Likas, K. Blekas & D. Kalles (Eds.), *Artificial intelligence: Methods and applications lecture notes in computer science volume 8445* (pp. 313-326). Switzerland: Springer International Publishing.

Prasath, R. R. (2010). Learning age and gender using co-occurrence of non-dictionary words from stylistic variations. *Rough sets and current trends in computing* (pp. 544-550)

Raymond, E. (2003). The jargon file. Retrieved from http://catb.org/jargon/html/B/bug.html

Reed, J., D. (1987). The shakespeare mystery: Some ado about who was or was not shakespeare. *Smithsonian, September*, 13 Februrary 2013.

Rege, A. (2009). What's love got to do with it? exploring online dating scams and identity fraud. *International Journal of Cyber Criminology, 3*(2), 494-512.

Rohrdantz, C., Haulti, A., Mayer, T., Butt, M., Keim, D. A., & Plank, F. (2011). Towards tracking semantic change by visual analytics. *Proceedings of the 49th Annual Meeting of* *the Association for Computational Linguistics: Short Papers,* Portland, Oregon, USA, June 2011. 305-310.

Rosenthal, S., & McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oregon, USA. (June) 762-772.

Rude, S. S., Gortner, E., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognitive and Emotion, 18*(8), 1121-1133.

Rudman, J. (2010). The state of non-traditional authorship attribution studies – 2010: Some problems and solutions. *Digital Humanities, 31*(4), 351-365.

Saevanee, H., Clarke, N., & Furnell, S. (2011). Sms linguistic profiling authentication on mobile device. *Network and System Security (NSS), 2011 5th International Conference On*, 224-228.

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Preactical statitics for user research*. Waltham MA USA: Morgan Kaufmann.

Sazonova, V., & Matwin, S. (2014). Combining binary classifiers for a multiclass problem with differential privacy. *Transactions on Data Privacy, 7*(1), 51-70.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 6*, 199-205.

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *Plos One, 8*(9), e73791.

Schwartz, R., Tsur, O., Rappoport, A., & Koppel, M. (2013). Authorship attribution of micro-messages. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Washington USA. 1880-1891.

Shannon, C. (1948). A mathematical theory of communication, bell system technical journal 27: 379-423 and 623–656. *Mathematical Reviews (MathSciNet): MR10, 133e,*

Shirali-Shahreza, S., & Shirali-Shahreza, M. (2008). Identifying child users: Is it possible? *SICE Annual Conference, 2008,* 3241-3244.

Shull, A. (2014). *Global cybercrime: The interplay of politics and law.* (Internet Governance Papers No. 8). Canada: Center for International Governance Intervention.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (1999). Automatic authorship attribution. *Proceedings of EACL' 99,* 158-164.

Steinberg, L. (2005). Cognitive and affective development in adolescence. *TRENDS in Cognitive Sciences, 9 No. 2*(February), 69-74.

Sun, X. (2008). *Why gender matters in CMC: Gender differences in remote trust and performance with initial social activities* (Doctor of Philosophy).

Tetreault, J., Blanchard, D., & Cahill, A. (2013). <br />A report on the first native language identification shared task. *<br />Proceedings of the Eighth Workshop on Innovative use of NLP for Building Educational Applications,* 48-57.

Torney, R., Vamplew, P., & Yearwood, J. (2012). Using psycholinguistic features for profiling first language of authors. *Journal of the American Society for Information Science and Technology, 63*(4) doi:10.1002/asi.22627

Trevathan, J., & Myers, T. (2012). Anti-social networking? *World Academy of Science, Engineering and Technology, 72*, 127-135.

Tsur, O., & Rappoport, A. (2007). Using classifier features for studying the effect of native language on the choice of written second language words. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition,* 9-16.

Tufis, D., & Mason, O. (1998). Tagging romanian texts: A case study for qtag, a language independent probabilistic tagger. *Proceedings of the First International Conference on Language Resources and Evaluation (LREC), , 1* 589-596.

Ugheoke, T. O. (2014). *Detecting the gender of a tweet sender* (Master of Science).

Vajjala, S., & Loo, K. (2013). Role of morpho-syntactic features in estonian proficiency classification. *NAACL/HLT 2013,* , 63-72.

Van Halteren, H. (1999). Weighted probability distribution voting, an introduction. *CLIN,*

van Halteren, H. (2004). Linguistic profiling for author recognition and verification. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* , 199-206.

van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linquistics, Volume 12*(No 1), 65-77. doi:10.1080/09296170500055350 van Halteren, H., Tweedie, F., & Baayen, H. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*(3), 121-132. doi:10.1093/llc/11.3.121

Van Horne, P. (2013). From the bottom up – profiling like sherlock holmes. *The CPJournal,* doi:http://www.cp-journal.com/2013/05/from-the-bottom-up-profiling-like-sherlock-holmes/

Warschauer, M., & Kem, R. (Eds.). (2000). *Network-based language teaching: Concepts and practice* (3rd ed.). New York, USA: Cambridge University Press.

Watters, P. (2002). Discriminating english word senses using cluster analysis. *Journal of Quantitative Linguistics, 9*(1), 77-86.

Whittle, H., Hamilton-Giachritsis, C., Beech, A., & Collings, G. (2013). A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior, 18*(1), 62-70.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques with JAVA implementations* (2nd ed.). San Francisco, California, USA: Elsevier.

Wolak, J., Finkelhou, D., Mitchell, K. J., & Ybarra, M. L. (2008). Online "predators" and their victims: Myths, realities and implications for prevention and treatment. *American Psychologist, 63*, 111-128.

Wong, S. J., & Dras, M. (2009). Contrastive analysis and native language identification. *Procedings of the Australasian Language Technology Association Workshop 2009,* , 53-61.

Yang, L., Ma, A., & Cao, Y. (2013). Lexical negative transfer analysis and pedagogic suggestions of native language in chinese EFL writing. *2013 Conference on Education Technology and Management Science (ICETMS 2013),*

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society of Information Science and Technology, February*, 378-393.

Zukerman, W., & Purcell, A. (2011). Brain pruning continues into early adulthood. *New Scientist, 211*(2826) doi:02624079, 8/20/2011

Appendix A

Category	Abbrev	Examples	Words in Category
Linguistic Processes			
Word count	wc		n/a
words/sentence	wps		n/a
Dictionary words	dic		n/a
Words>6 letters	sixltr		n/a
All Punctuation	AllPct		n/a
Period	Period		n/a
Comma	Comma		n/a
Colon	Colon		n/a
Semi Colon	SemiC		n/a
Question Mark	QMark		n/a
Exclamation mark	Exclam		n/a
Dash	Dash		n/a
Quote	Quote		n/a
Apostrophe	Apostro		n/a
Parentheses	Parenth		n/a
Other Punctuation	OtherP		n/a
Total function words	funct		464
Total pronouns	pronoun	I, them, itself	116
Personal pronouns	ppron	I, them, her	70
1st person singular	i	I, me mine	12
1st person plural	we	We, us, our	12
2nd person	you	You, your, thou	20
3rd person singular	shehe	She, her, him	17
3rd person plural	they	They, their, they'd	10
Impersonal pronouns	ipron	It, it's, those	46
Articles	article	A, an, the	3
[Common verbs]	verb	Walk, went, see	383
Auxiliary verbs	auxverb	Am, will, have	144
Past tense	past	Went, ran, had	145
Present tense	present	Is, does, hear	169
Future tense	future	Will, gonna	48
Adverbs	adverb	Very, really, quickly	69
Prepositions	prep	To, with above	60
Conjunctions	conj	And, but, wheras	28
Negations	negate	No, not never	57
Quantifiers	quant	Few, many much	89
Numbers	number	Second, thousand	34
Swear words	swear	Damn, piss, fuck	53
Psychological processes			

Social processes	social	Mate, talk, they, child	455
Family	family	Daughter, husband, aunt	64
Friends	friend	Buddy, friend, neighbour	37
Humans	human	Adult, baby, boy	61
Affective processes	affect	Happy, cried, abandon	915
Positive emotion	posemo	Love, nice, sweet	406
Negative emotion	negemo	Hurt, ugly, nasty	499
Anxiety	anx	Worried, fearful, nervous	91
Anger	anger	Hate, kill, annoyed	184
Sadness	sad	Crying, grief, sad	101
Cognitive process	cogmech	cause, know, ought	730
Insight	insight	think, know, consider	195
Causation	cause	because, effect, hence	108
Discrepancy	discrep	should, would, could	76
Tentative	tentat	maybe, perhaps, guess	155
Certainty	certain	always, never	83
Inhibition	inhib	block, constrain, stop	111
Inclusive	incl	And, with, include	18
Exclusive	excl	But, without, exclude	17
Perceptual processes	percept	Observing, heard, feeling	273
See	see	View, saw, seen	72
Hear	hear	Listen, hearing	51
Feel	feel	Feels, touch	75
Biological processes	bio	Eat, blood, pain	567
Body	body	Cheek, hands, spit	180
Health	health	Clinic, flu, pill	236
Sexual	sexual	Horny, love, incest	96
Ingestion	ingest	Dish, eat, pizza	111
Relativity	relativ	Area, bend, exit, stop	638
Motion	motion	Arrive, car, go	168
Space	space	Down, in, thin	220
Time	time	End, until, season	239
Personal concerns			
Work	work	Job, majors, xerox	327
Achievement	achieve	Earn, hero, win	186
Leisure	leisure	Cook, chat movie,	229
Home	home	Appartment, kitchen, family	93
Money	money	Audit, cash, owe	173
Religion	relig	Altar, church, mosque	159
Death	death	Bury, coffin, kill	62
Spoken categories			
Assent	assent	Agree, OK, yes	30
Nonfluencies	nonflu	Er, hm umm	8
Fillers	filler	Blah, imean, youknow	9