NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AUTHORSHIP DISCOVERY IN BLOGS USING BAYESIAN CLASSIFICATION WITH CORRECTIVE SCALING

by

Grant T. Gehrke

June 2008

Thesis Advisor: Craig H. Martell, Ph.D.
Second Reader: Kevin M. Squire, Ph.D.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*
30–6–2008

2. REPORT TYPE
Master's Thesis

3. DATES COVERED *(From — To)*
2007-07-01—2008-06-20

4. TITLE AND SUBTITLE
Authorship Discovery in Blogs using Bayesian Classification with Corrective Scaling

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Grant T. Gehrke

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Widespread availability of free, public blog platforms has facilitated growth in the amount of individually written electronic text available online. Our research leverages an extremely large blog corpus for a study in authorship discovery, both to evaluate a traditional technique as applied to blogs, as well as to demonstrate the implications of authorship discovery in blogs for intelligence and forensic purposes.

Our study uses a Bayesian classifier with two important extensions. First, we introduce a post-classification corrective scaling technique to mitigate the over-classification of many samples to a few authors. Second, we propose an n-percent-correct threshold metric, whereby we define a "correct" result as one where the true author is within some small subset of the original search space rather than requiring that he or she be the single most probable author. Using this technique, we are able to reduce a search space of 2000 authors to 1% of its original size with 91% accuracy when 1000 bigrams are present, or reduce the search space to 10% of its original size with 94% accuracy when only 500 bigrams are present.

15. SUBJECT TERMS
Authorship Attribution, Authorship Discovery, Natural Language Processing, Machine Learning, Blogs, Bayes, Bayesian Classification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
Unclassified	Unclassified	Unclassified	UU	51	19b. TELEPHONE NUMBER *(include area code)*

THIS PAGE INTENTIONALLY LEFT BLANK

AUTHORSHIP DISCOVERY IN BLOGS USING BAYESIAN CLASSIFICATION WITH CORRECTIVE SCALING

Grant T. Gehrke
Ensign, United States Navy
B.S., United States Naval Academy, 2007

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
June 2008

Author:	Grant T. Gehrke
Approved by:	Craig H. Martell, Ph.D. Thesis Advisor
	Kevin M. Squire, Ph.D. Second Reader
	Peter J. Denning, Ph.D. Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Widespread availability of free, public blog platforms has facilitated growth in the amount of individually written electronic text available online. Our research leverages an extremely large blog corpus for a study in authorship discovery, both to evaluate a traditional technique as applied to blogs, as well as to demonstrate the implications of authorship discovery in blogs for intelligence and forensic purposes.

Our study uses a Bayesian classifier with two important extensions. First, we introduce a post-classification corrective scaling technique to mitigate the over-classification of many samples to a few authors. Second, we propose an n-percent-correct threshold metric, whereby we define a "correct" result as one where the true author is within some small subset of the original search space rather than requiring that he or she be the single most probable author. Using this technique, we are able to reduce a search space of 2000 authors to 1% of its original size with 91% accuracy when 1000 bigrams are present, or reduce the search space to 10% of its original size with 94% accuracy when only 500 bigrams are present.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
INTRODUCTION

1.1 Motivation

In recent years, the blogging phenomenon has dramatically changed how Internet users access and share information. Ongoing, periodic publication of news and opinions for an open audience was once restricted to newspaper and magazine publishers. Even in the digital age, publishing to the Internet was formerly restricted to large businesses and only the most technically savvy. Free availability to many public, easy-to-use blogging host services has lowered that bar such that the only requirements for sharing your writing with the world are a computer with Internet access and something to say.

The influence of individual blogs on reporting of news has made many of their authors into celebrity writers that often drive or even outshine the traditional outlets entirely. Niche blogs allow writers and readers to seek each other out and connect for ongoing commentary on even the most obscure topics. A single individual acts as the writer, editor and publisher, removing the revising and filtering process of traditional publishing and facilitating posting of raw, opinionated, and controversial blogs if the author desires. Further, this individual may or may not choose to reveal his or her true identity.

Myriad situations exist in which we may wish to discover the author of some anonymous electronic communication, whether a blog post, a comment on a blog, content on a "wiki," a message board post, chat messages, or an anonymous email. The motivation for discovering the author's identity could range from forensic evidence gathering in criminal proceedings, intelligence analysis, revealing or authenticating a "whistle blower," or simple curiosity. In the absence of other identifying information such as the originating computer's IP address or connection logs, the text itself may be our only method of discovering the true author of an anonymous message. When the list of suspect or potential authors is extremely large, this becomes a daunting task. Application of machine learning techniques, however, could allow the list to be dramatically reduced, ideally to a single individual or a set which is a fraction of the size of the original, making the job of a human investigator much more manageable.

1.2 Organization of Thesis

In Chapter 1 we discuss the motivation for examining authorship attribution of electronic documents, specifically blogs, due to the potential impact of studies on real-world investigations. Chapter 1 also introduces the concept of authorship discovery, in contrast to traditional authorship attribution. In Chapter 2 we first outline the foundations of computational attribution from the earliest studies to modern techniques. Chapter 2 also outlines the characteristic language of blogs and how they compare to other forms of written text.

Chapter 3 presents an experiment in authorship attribution using a blog corpus. First we discuss the corpus preparation, including motivation for using blogs as a testbed. Second, we detail the classification scheme using a Bayesian classifier. Third, we introduce a corrective scaling factor which, applied to the results of a classification, improve results dramatically. Finally, we propose a metric for judging success of classification by reducing the search space to some threshold in scenarios where the search space is extremely large.

Chapter 4 discusses the results of our experiment in Bayesian classification including qualitative discussion of the concept of relaxing the n-percent-correct threshold in real-world problems. Chapter 4 also discusses the possibility of a critical flaw in our approach, which arises from the inclusion of content words, and must be regarded with caution.

Finally, Chapter 5 presents a brief review and proposes several directions in which the study of authorship discovery in blogs can continue to move forward.

CHAPTER 2:
BACKGROUND TOPICS

In this chapter we discuss the foundations for computational authorship attribution. First, a survey of existing techniques for discovering authorship are explored. Second, we explore the reasoning behind examining online blogs and specifically their use as a corpus for authorship studies. Finally, classification using the Naïve Bayes classifier, the primary algorithm used in our research, is explained.

2.1 Authorship Attribution

2.1.1 Research Scope

The task of authorship attribution can be defined as a structured method of determining the individual who generated some sample of text. Specifically, in this thesis we assume that the task is being performed strictly on electronic text files with no markup to help distinguish between authors such as timestamps, originating computer identification, or textual formatting. Tangential fields such as handwriting recognition or computer forensics are, therefore, not discussed. For the purpose of constraining the problem we are also not considering the possibility that a human expert could subjectively determine the author of a sample from its content or style much as a literary expert might, instead restricting our study to computational methods.

The task of authorship attribution is also not strictly aligned with the task of authorship verification. The task of validating with some level of confidence whether a single suspect individual is the true author of a sample will also not be directly explored. In this thesis we address the issue of authorship attribution or what may even be thought of as "authorship discovery."

2.1.2 History of Authorship Attribution

T.C. Mendenhall, in 1887, published what is considered the first scientific study of authorship attribution based on syntactic characteristics of sample texts. In [26], his approach expands on Augustus DeMorgan's suggestion that comparing mean word length in two texts could be an indicator of whether they were written by the same individual. Mendenhall argues that the mean word length is, itself, not discriminating enough, but supposes that comparing a histogram of word lengths, which he calls a "characteristic curve of composition," would more finely resolve the differences between authors. Relating the process to spectral analysis of the light emitted when elements are heated, which is known to precisely identify the element, Mendenhall suggests that an author may generate texts in the same uniquely characteristic manner as a physical specimen would emit light.

Examples of Mendenhall's Curves of Composition from [26]

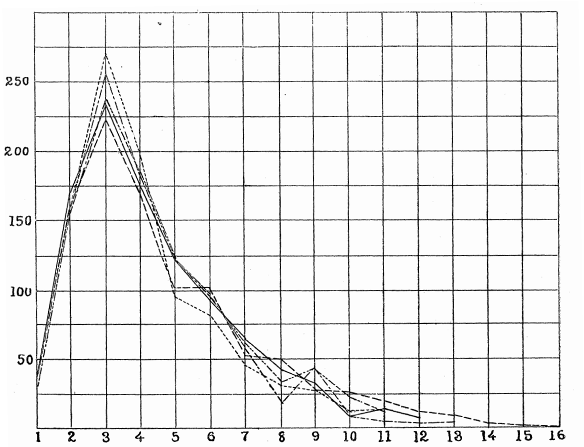


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

Figure 2.1: Histogram demonstrating "consistent" curve between samples from the same author, and in fact the same text. Visible variance is, in Mendenhall's opinion, due to the relatively small sample size of 1000 words.

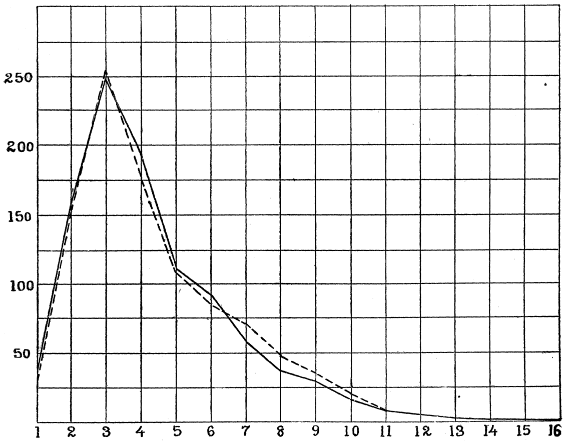


FIG. 7.—TWO GROUPS, OF TEN THOUSAND WORDS EACH, FROM 'OLIVER TWIST,'——; AND FROM 'VANITY FAIR,'— — — —.

Figure 2.2: Histogram representing curves of two different authors. Mendenhall, though, attributes their virtual similarity to "the result of accident" and claims that "it would not be likely to repeat itself."

Mendenhall's results were understandably limited. Generating curves required manual counting of letters in sets of 1000-5000 words at a time from the works of classic authors. His initial paper, as well as a follow-on in 1901, do suggest that, given a large enough sample, characteristic curves emerge which allow discrimination between authors. However, the example curves in figures 2.1 and 2.2 do not make a convincing case for his conclusions. Mendenhall recognizes the benefit of his approach, though, as "purely mechanical in its application," which was a new concept in the field. This is in contrast to the subjective analysis that a literary scholar might perform to describe the differences between the eloquence of Dickens and Thackeray, for example. Further, Mendenhall suggests that the approach could be equally applied to counts of syllables or histograms of word counts per sentence.

Building on Mendenhall's premise that textual statistics can be used as an authorial fingerprint, subsequent researchers have sought to use various additional measures, both in the same manner as Mendenhall's original experiments as well as using new methods of analysis.

- G. U. Yule, in 1939, counted lengths of an author's sentences, concluding that "sentence-length *is* a characteristic of an author's style," but that the judgement of authorship must be "a personal one," given the evidence of sentence length distributions [35]. In the two specific cases Yule presents, he does make conclusive judgments about the authorship of disputed texts, demonstrating, for example, that Thomas á Kempis' mean sentence length of 17.9 matched that of *Imatatio Christi* (mean of 16.2) more closely than Jean Charlier de Gerson, the once believed author, at 23.4.
- Similarly, Conrad Mascol evaluated the New Testament Epistles using a measure of sentences per printed page [25], determining that Paul had not written some of the books which scholars believed he had.
- Wilhelm Fucks discriminated between authors using the average number of syllables per word and average distance between equal-syllabled words [8]. Fucks, too, concluded that a study such as his reveals a "possibility of a quantitative classification which is very simple to realize," but recognizes that his measures delineated samples largely on the language, level of prose, and progressive changes in style through historical periods rather than being strictly indicative of authorship.
- In [7], R. Forsyth, D. Holmes and E. Tse revisit syllable length measures to demonstrate that the Renaissance scholar Sigonio likely faked his supposedly complete version of Cicero's *Consolatio*, which had previously existed only in fragments, concluding that portions use language more characteristic of the Renaissance than classical times.

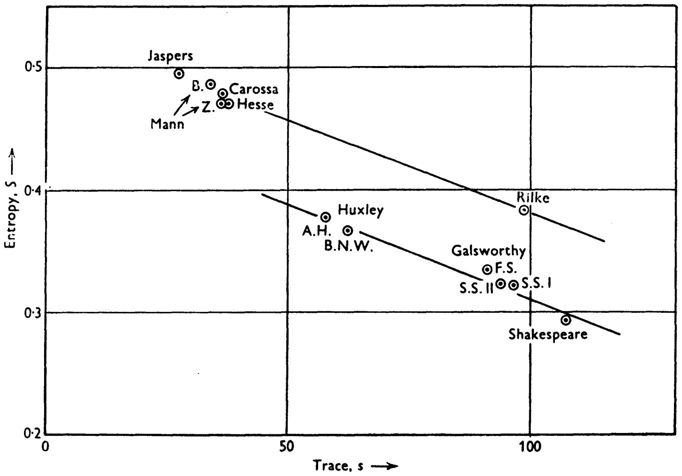


Figure 2.3: W. Fucks' Diagram from [8] relating frequencies of n-syllable words to the distance between words of the same number of syllables. Fit lines indicate German versus English language texts and position on line is indicative of mixture of prose and verse styles of writing.

Further Stylometric methods of quantifying style

Extending beyond word and sentence length histograms, several other textual measures have been proposed and used for authorship attribution problems. In [13], Holmes asserts:

> One of the fundamental notions in computational stylistics is the measurement of what is termed the "richness" or "diversity" of an author's vocabulary. The basic assumption is that the writer has available a certain stock of words, some of which he/she may favour more than others... If, furthermore, we can find a single measure which is a function of all the vocabulary frequencies and which adequately characterizes the sample frequency distribution we may then use that measure for comparative purposes.

Among the most widely used measures in this category is the type-token ratio, a representation of the number of unique word types, V, divided by the counted length of the text sample, N[1]. In plain terms, this measure represents the breadth of the author's vocabulary used in the sample of interest. Unfortunately, the type-token ratio has limited use in authorship studies. In particular, type-token ratio is unstable with the size of the document and it may be highly dependent on other factors such as the style of writing. Type-token ratio does, however, lend itself as an easily understood starting point for understanding the quantification of an author's "style."

Additional stylometric measures include:

- Word Frequency Distributions
  One implication of the well-known Zipf's Law for text samples is that the vast majority of word types in a text are used infrequently, with most of a text sample being comprised of only a small set of types, describing a "frequency distribution of words in human languages." [24] Supposing that this distribution may vary slightly between individual writers, it may be used to compare authors. In particular, counts of *hapax legomena*, word types that are used only once, and *hapax dislegomena*, word types that are used only twice, have been proposed as measures for authorship attribution but have been found to be lacking on their own. [14]

[1] A word *type* encompasses all occurrences of that word in a text whereas a word *token* is a single occurrence of a word or other marker in the text such that multiple occurrences of the same word are each separate *tokens* but are all of the same *type*.

- Yule's Characteristic (K)
  Defined as $K = (10^4 \sum_r r^2 V_r - N)/N^2$ where V_r represents the number of words that occur r times in the sample and N represents the total number of tokens in the sample. Yule's Characteristic is based on a Poisson distribution and describes the proportion of words in a sample that are repeated r times, weighted by r^2. [13]

- Simpson's Index (D)

  $$D = \sum_r \frac{r(r-1)V_r}{N(N-1)}, (r = 1, 2, ..) = \sum_{i \in V} \frac{n_i(n_i - 1)}{N(N-1)}$$

  for r occurrences greater than zero or all i types in V.
  Simpson's Index measures the probability that two tokens, drawn randomly from a sample of text, will be of the same type. In particular, it is useful for comparing texts of different lengths. [31] [13]

- Entropy
  Borrowing from the thermodynamic concept of entropy, $S = -k \sum_i p_i \log p_i$, where p_i is the probability of appearance of the ith lemma and k is an arbitrary constant, represents the measure of disorder or randomness in a text sample. [13] [8] [4]

- "S" measure introduced by Golcher [10]
  $S(T, t) = \sum_{m,n} \log(F_T(s'_{m,n}) + 1)/L$ where $F_T(s'_{m,n})$ is the number of occurrences of $s'_{m,n}$ in T. In practical terms, S measures how frequently substrings of characters of all lengths, reminiscent of a power set, are repeated in a text. Golcher's published results perform comparably with other methods, including "correct" classification of all the disputed *Federalist* papers.

- Gunning-Fog Index, Simple Measure of Gobbledygook, Automated Readability Index, Flesch Reading Ease, Flesch-Kincaid Grade Level.
  In [22], Mala borrows several novel linguistic measures, most of which represent the complexity of a text as a level of linguistic sophistication by quantifying syllables per sentence, for example, and uses a 3D visualization technique to product on-screen "objects" which a human subject can quickly and naturally determine to be similar or dissimilar.

Further, a multivariate approach, combining or comparing several different measures will almost certainly lend them even greater discriminating power. [14]

Lexical Approaches to Authorship Attribution

Whereas the above stylometric approaches to authorship attribution seek to generalize a text sample based, in most cases, on statistics of its construction, a somewhat different approach is to examine the distribution of the actual words, or in some cases letters or other graphemes, and their comparative usage between texts. In most cases these lexical techniques do not approach the level of semantic analysis, where the words would have some inherent "meaning" to the classifier, but the words themselves *are* counted and manipulated directly.

In [5], Ellegård took an extremely labor intensive approach to building word frequency distributions for determining authorship in the *Junius Letters*. He manually constructed a "distinctiveness" measure similar to tf-idf[2], where words that appeared frequently (or infrequently) in each of the suspect authors' known works, but which which do not appear frequently in other writers' documents, were highly ranked. Ellegård then manually counted these "plus" and "minus" words in each of the *Junius Letters* for each author, arriving at a similarity score for each author on each document. In the end, Elleågard's conclusion was that Sir Phillip Francis, the suspected author of the letters, was the true writer. His approach was not without its faults, however. In particular, Ellegård *did* include content words in his lists of "plus" and "minus" words. It is now common practice to regard a word with a high tf-idf score as distinctive of the primary *topic* of some given document in a corpus. Because this is what Ellegård was essentially matching, his approach has the potential to more closely align two distinct authors who write about similar topics than one author who writes about disparate topics.

In their landmark 1963 and 1964 studies on the *Federalist* papers, [28] [29], Mosteller and Wallace examine the *Federalist* papers with statistical analysis of word frequencies. According to [28],

> The *Federalist* papers were published anonymously in 1787-1788 by Alexander Hamilton, John Jay, and James Madison to persuade the citizens of the State of New York to ratify the Consititution. Of the 77 essays, 900 to 3500 words in length, that appeared in newspapers, it is generally agreed that Jay wrote five: Nos. 2, 3, 4, 5, and 64, leaving no further problem about Jay's share. Hamilton is identified as the author of 43 papers, Madison of 14. The authorship of 12 papers (Nos. 49-58,

[2] Term Frequency - Inverse Document Frequency is a method of scaling the importance of a term to a document based on how frequently it occurs in the document scaled by how infrequently it occurs in all documents in a corpus.

> 62, and 63) is in dispute between Hamilton and Madison; finally, there are also three joint papers, Nos. 18, 19, and 20, where the issue is the extent of each man's contribution.

Early manual examination suggested that the use of certain words such as 'upon,' or preference for 'while' versus 'whilst,' were strong discriminators between Madison and Hamilton. Extending this concept, Mosteller and Wallace constructed a set of 30 words, comprised of function words such as 'by,' 'of,' and 'to,' as well as "well-liked" words such as 'commonly,' 'vigor,' and 'particularly,' which were determined not to convey topical meaning and not to vary with context. Examples of words *not* counted in the study were 'war,' 'executive,' and 'legislature' despite the fact that they appeared very frequently, a standard often used for determining function words. Counting the frequencies of these words for each author and fitting to a Poisson or negative binomial distribution (the difference is "not of major importance" [28]) allows a model of prior probabilities to be built.

Turning to the disputed texts, Mosteller and Wallace used Bayes' Theorem to balance the prior probabilities of each individual's potential authorship with the posterior odds that each text was written by the individual given its word frequencies. In their example, if x is one sample from a discrete set of possible observations, p_i is the prior probability of hypothesis i and $f_i(x), i = 1, 2$ is the conditional probability of observing x given that hypothesis i is true, then

$$P(Hypothesis\ 1 \mid x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}.$$

Mosteller and Wallace make judgments in the paper based on the "odds" of one hypothesis being true over the other, with hypothesis 1 being that Hamilton was the author of the paper in question and hypothesis 2 that Madison was the author. Final odds are defined as the initial odds multiplied by the likelihood ratio, or,

$$Odds(1, 2 \mid x) = \frac{P(Hypothesis\ 1 \mid x)}{P(Hypothesis\ 2 \mid x)} = \frac{p_1 f_1(x)}{p_2 f_2(x)} = \left(\frac{p_1}{p_2}\right)\left(\frac{f_1(x)}{f_2(x)}\right).$$

Further, the likelihood ratio for multiple words is the product of the likelihood ratios for each word individually and, to make the numbers manageable, the odds can also be computed as a log-likelihood. In [28], the problem of choosing initial odds is explained away through the assumption that any appreciable number of observed words with strong likelihood ratios will quickly overwhelm any variation in the initial odds. In a problem such as the disputed *Federalist* papers, the initial odds for Hamilton versus Madison may as well be 1, or a 50-50 chance that either individual was the author.

The result of Mosteller and Wallace's study confirms what historians believed about the *Federalist* papers, that Madison had written all twelve of the disputed documents. Additionally, they raise several issues relevant to the study of statistical authorship attribution in general, such as the utility of function words as discriminating features and the observation that prior distributions, which may have otherwise required human intervention through scholarly study of a disputed text, are of negligible importance.

2.2 Lexical Characteristics of Blogs

It is quite clear to anyone who reads blogs that they are a unique form of written communication. Looking strictly at their language use, the subtle differences between blogs and other forms of writing begin to emerge. In [27], Mishne provides a thorough overview of language use in blogs and the difference between blogs and other forms of text. In particular, he identifies top indicative words from distributions for web, usenet and blog genres, noting that "blogs have a distinctive personal feel," but contain "words related to personal surroundings [. . .] and references to current events," supporting the intuition that their language model is a combination of personal correspondence and news reporting.

Mishne also examines several measures of lexical difference between blogs and other corpora such as the Kullback-Liebler divergence, perplexity, and three "readability" measures. KL divergence expresses how different two probability distributions, p & q, are and is defined as their relative entropy, [24]

$$D(p \mid\mid q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{2.1}$$

Using a measure of KL divergence, blogs are most similar to "personal letters" (with a score of 0.25) and most divergent from "scientific articles" (with a score of 1.06). Perhaps surprisingly, blogs are significantly different from "newspapers" (with a score of 0.48) and the web at large (with a score of 0.75).

Turning to perplexity, defined for the probability distribution of a large sample of text from the genre, P, as $2^{H(P)}$ where $H(P)$ is the entropy of the distribution [24], blogs have relatively high scores, averaging 301. For comparison, newspapers have a reported score of 355, essays are scored at 295, fiction is 245, and personal letters are 55 [27]. Mishne concludes:

> The relatively high perplexity of blog language, compared with other genres to which it is similar in the type of vocabulary, indicates less regularity in use of language: sentences are more free-form and informal, and adhere less to strict rules.

Finally, in terms of readability, a measure tied to the familiar concept of "grade-levels," Mishne scores blogs relatively low, ranking them 9.9, 7.0 and 8.9 on the three scales examined (Gunning-Fog[3], Flesch-Kincaid[4], and Simple Measure of Gobbledygook[5], respectively). These scores are higher than fiction, but lower than both "school" and "university" essays as well as newspapers. Mishne attributes much of this to the actual age of most bloggers, which is in the teens, and their subsequently shorter sentence and word lengths.

In general, Mishne concludes, in concert with other researches who have studied the lexical characteristics of blogs, that they are most similar to school essays. They clearly have similarity on some levels with the language usage in news outlets and fictional writing, but must be considered a separate genre with regard to the standard language model used in the blogosphere. The conclusion that blogs do not, however, generally conform to a single, standard language model, is encouraging for authorship studies, where an individual may not feel as compelled to shoehorn their own style into the formalized rules mandated by other forms for writing.

[3] $GFI = 0.4\left(\left(\frac{words}{sentence}\right) + 100\left(\frac{complex\ words}{words}\right)\right)$ [22]

[4] $FGL = 0.39\left(\frac{total\ words}{total\ sentence}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59$ [22]

[5] $SMOG = \sqrt{total\ complex\ words\left(\frac{30}{total\ sentences}\right)} + 3$ [22]

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:
Experiment in Bayesian Classification

3.1 Use of Blogs as Authorship Attribution Corpus

3.1.1 The Personally Revealing Nature of Blog Writing

The content of blogs is typically very personal to the writer. All will express some unique viewpoint on the topic of interest, with some going so far as to write almost exclusively about their personal lives. In recent years, these ‘diary’ type blogs outnumber those of the earlier ‘filter’ and ‘notebook’ blogs. Some authors even “define themselves through their blog.” [27] [33] If we regard the goal of authorship studies as building accurate models of an author’s particular internal language model, then the availability of a corpus with this level of access into the mind of the author is a great asset to the study of authorship attribution.

In [13], Holmes suggests, for example, that sentence length measures are only applicable when the author’s sentence division intent and use of punctuation are preserved. In traditional authorship studies, where a document may have been edited prior to publishing, had its punctuation usage standardized, or been translated between languages, this concern forces researchers to approach these measures with reservation. Typical blogs, on the other hand, are almost invariably the work of a single author and are not subject to the same level of editorial scrutiny necessary for traditional print media.

3.1.2 The Technical Suitability of a Blog Corpus for Authorship

Compared to the text subjects of traditional authorship attribution studies, blogs are relatively easy to collect. Though the prevalence of resources such as Project Gutenberg[1] has made access to classic literature much more reasonable than in past eras, where researchers spent years manually counting words off a printed page, blogs exist in a natively electronic format that is readily accessible to anyone who wishes to access it.

In particular, they allow us to sample many times more authors than could be reasonably examined through study of published literature or student essays generated for a particular study, the traditional corpora for authorship studies.

[1] Project Gutenberg available at http://www.gutenberg.org. Accessed 30 June, 2008.

Blogs are inherently time sensitive, with the date and time being crucial to each post's relevance. Though we didn't examine the chronological aspects of an author's writing to blogs, this information is available directly from the web host, making blogs an ideal corpus for researchers examining the progression of an author's writing style over time, for example.

3.2 Corpus Preparation

The authorship corpus we are using was developed by J. Schler, M. Koppel, S. Argamon and J. Pennebaker [17] and contains writings from nearly 20,000 authors on blogger.com. The corpus contains a single XML file for each author in the corpus, with each file containing all posts by that author accessible at the time of download (August 2004) annotated with the date and time of posting. All formatting in the original HTML blog has been removed, leaving only plain text. Additionally, the original researchers removed all URL links, replacing them with the token 'urllink.'

In the interest of processing time, the larger corpus was limited to at most 2000 authors for each experiment. To establish training and testing sets, at least 10% of each author's posts were set aside. Each blog's posts were first shuffled to remove any chronological influence. Next, a size threshold for each author's training set was chosen at 10% of the size, in words, of all posts in the original file combined. Whole posts were then removed from the original file and placed in a new testing file until the test sample's size, in words, met or exceeded the 10% threshold. The remaining posts were designated for training and written to a new file.

Training and test sets were both regarded as bag-of-words[2] models. For this reason, we can treat the concatenation of all posts in an author's training set as a single document, and likewise with the test document. For each classification experiment, the test document size was further limited to 100, 250, 500, 750, and 1000 bigrams in order to test the improvement in accuracy as the size of the document in question increases, a practical consideration for scenarios when we may wish to classify a diminutive text. The unit of classification, therefore, will be on the level of a partial document, comprised of concatenated posts and truncated to the test length.

[2] A *bag-of-words* model is one where "all the structure and linear ordering of words within the context is ignored." [24]. This assumption is naïve in that the frequency of a word type's occurrence certainly depends on its context, but evidence suggests that results are not severely impacted in many scenarios.

3.3 Model Building for Each Author

3.3.1 Construction of Models for Training Data

To cope with the large size of the set of suspect authors, a model for each was constructed from the training data and saved prior to classification.

Sentence Chunking

Each post was divided into sentences so that we could retain nominal position information, particularly what word were most likely to occur as the first word of a sentence or as the last. Boundaries were detected by the Punkt tokenizer, described in [15]. The Punkt algorithm is based on simple division rules for punctuation, but it is initially "trained" on large samples of text so that it can learn the nuances of when punctuation does or does not actually indicate the end of a sentence. For example, Punkt will learn that 'Dr.' occurs frequently as an abbreviation but its period does not *necessarily* mark the end of a sentence. In the presence of other information to indicate that the author intended to conclude their sentence with the abbreviation 'Dr.,' for example, the system *will* divide the sentence there. Though we did not train Punkt on annotated blog data, instead using standard english training data, the algorithm performed very well across the blog corpus. The lack of strict formalities in blog writing, however, makes the notion of dividing into traditional sentence inherently difficult. Consistency between training and testing data should mitigate this.

Bigram Tokenization

For classification we focused on bigram word frequencies. Word n-grams are groupings of n words appearing next to each other in the text. *Uni*grams are, therefore, n-grams with n=1, or single word tokens, and *bi*grams are n-grams with n=2. Use of bigrams does allow us to retain some notion of sequential information without the problem of sparsity when larger groupings are used. Bigrams are determined with a simple sliding window such that each bigram is the space-separated string "$w_i w_{i+1}$" for $i = \{1, 2, ..., n-1\} where \mathrm{w}_i$ is the word at position i in the text sample and n is the length, in words, of the text. Additionally, a new token, '<S>' is inserted to retain start-of-sentence and end-of-sentence position information. As a result,

- The first bigram in a sentence is '<S> firstword'
- The last bigram in a sentence is 'lastword <S>'

Frequency Distributions of Bigrams

Each model consists of a frequency distribution, using the NLTK Frequency Distribution object [21], keyed on bigrams as space-separated strings. NLTK, the Natural Language Toolkit[3], is a module for the Python[4] programming language containing frequently used functions for computational linguistics and natural language processing tasks. Applied to bigram token samples, the FreqDist object records samples and provides the frequency of any particular type as a fraction of all observed samples.

Therefore, for each author a and bigram b,

$$P(b \mid a) = \frac{f_b}{|D_{T,a}|} \tag{3.1}$$

where f_b is the count of occurrences of b in a's training sample,
and $|D_{T,a}|$ is the count of bigrams in a's training sample, T

Additionally, simple Witten-Bell smoothing[5] was applied to each author's bigram distribution in order to deal with unseen bigrams.

3.3.2 Prior probabilities model

The prior probability of an author, also known as the "initial odds,"[28] is the probability that they wrote the wrote the document in question without regard to the contents of the document. In our study, an author's prior probability was based on the number of bigrams in their training sample as a fraction of the number of bigrams in all authors' training samples, representing how prolifically an author writes compared to his or her peers. The prior probability of author a_i, then, is

$$P(a_i) = \frac{|D_{T,a_i}|}{\sum_{a_j \in A} |D_{T,a_j}|}. \tag{3.2}$$

[3] NLTK available at http://www.nltk.org. Accessed 30 June, 2008.

[4] Python available at http://www.python.org. Accessed 30 June, 2008.

[5] Witten Bell Smoothing models the "probability of a previously unseen event by estimating the probability of seeing such a new event at each point as one proceeds through the training corpus." C.f. [24] page 222. In our case, unseen samples were approximated by $T/Z(N + T)$ where T is the number of types, N is the number of samples observed and Z is a scaling factor to ensure mass of the new distribution is 1. Further, T is approximated from the count of all bigram types in the entire corpus to estimate the maximum possible vocabulary size. The exact number chosen for this parameter was of little importance – even drastic experimental manipulation produced no change in results.

3.4 Naïve Bayes Classifier

3.4.1 Bayes' Theorem

Bayes' rule is widely used for deriving the conditional probability of some event, X, given Y, based on the marginal probabilities of X and Y and the probability of Y conditional on X, all of which may be easier to determine. Generally stated,

$$P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)}. \tag{3.3}$$

Applied to determination of authorship for a suspect a when a test feature vector, $\mathbf{F_t}$, is observed,

$$P(a \mid \mathbf{F_t}) = \frac{P(\mathbf{F_t} \mid a)\, P(a)}{P(\mathbf{F_t})}. \tag{3.4}$$

If a set of potential authors, A is known, the most probable among them is the one with the highest probability, or

$$a^* = \operatorname*{argmax}_{a_i \in A} \left[\frac{P(\mathbf{F_t} \mid a_i)\, P(a_i)}{P(\mathbf{F_t})} \right]. \tag{3.5}$$

Because the term $P(\mathbf{F_t})$ does not change between authors, the argmax operator allows us to discard it,

$$a^* = \operatorname*{argmax}_{a_i \in A} \left[P(\mathbf{F_t} \mid a_i)\, P(a_i) \right]. \tag{3.6}$$

Making the "naïve" assumption that each element of the feature vector $\mathbf{F_t}$ is independent of every other element, we can arrive at $P(\mathbf{F_t} \mid a_i)$ by taking the product of each element,

$$a^* = \operatorname*{argmax}_{a_i \in A} \left[P(a_i) \prod_{f_j \in \mathbf{F_t}} P(f_j \mid a_i) \right]. \tag{3.7}$$

Finally, because the product of small probabilities quickly becomes unmanageably small, we instead take the sum of the log-probabilities

$$a^* = \operatorname*{argmax}_{a_i \in A} \left[\log P(a_i) + \sum_{f_j \in \mathbf{F_t}} \log P(f_j \mid a_i) \right]. \tag{3.8}$$

3.4.2 Extension of classifier for ranking

For our evaluation, it is more advantageous to assign each potential author a score[6] by which they may be compared and ranked rather than simply returning the single most probable author.

$$S(a_i \mid \mathbf{F_t}) = \log P(a_i) + \sum_{f_j \in \mathbf{F_t}} \log P(f_j \mid a_i). \tag{3.9}$$

The single most probable author can still be chosen, of course, by

$$a^* = \operatorname*{argmax}_{a_i \in A} S(a_i \mid \mathbf{F_t}). \tag{3.10}$$

3.5 Corrective transformation of results

Observation of all authors' scores from each test sample revealed that a limited number of authors with the highest prior probabilities were overwhelmingly returned as the "correct" author by the classifier. Figure 3.1, a confusion matrix of Author ID's, illustrates this discrepancy. Note that the higher author ID's belong to authors with higher prior probabilities.

Points on the diagonal are authors who were correctly classified, defined as having the highest score on the test document sampled from his or her blog. It is apparent from fig. 3.1 that regardless of the prior probability of the true author for any arbitrary test sample, the Bayesian classifier returned one of the few authors with the highest prior probabilities, that is, authors who wrote quite prolifically.

Examination of the results from a single test in Fig 3.2, where the true author of the document in question was ranked the 49th most probable of 2000 authors, demonstrates that the prior probability of a suspect author is closely correlated with their ranking in this test. It is quite clear from examination of the plot, however, that the true author has a lower prior probability than those who are similarly ranked on their scores. That is, the true author has a much lower prior probability relative to his or her score than do the authors ranked 48th, 47th, 50th, 51st, etc. This trend was observed in many tests when manually examined.

[6] This "score" does not represent an absolute "probability" that the given author wrote a text sample. Instead it is a strictly comparative measure within a single test. In particular, no evidence was found to suggest that this score represents a level of confidence in the classification or other such metric that could be compared between test samples.

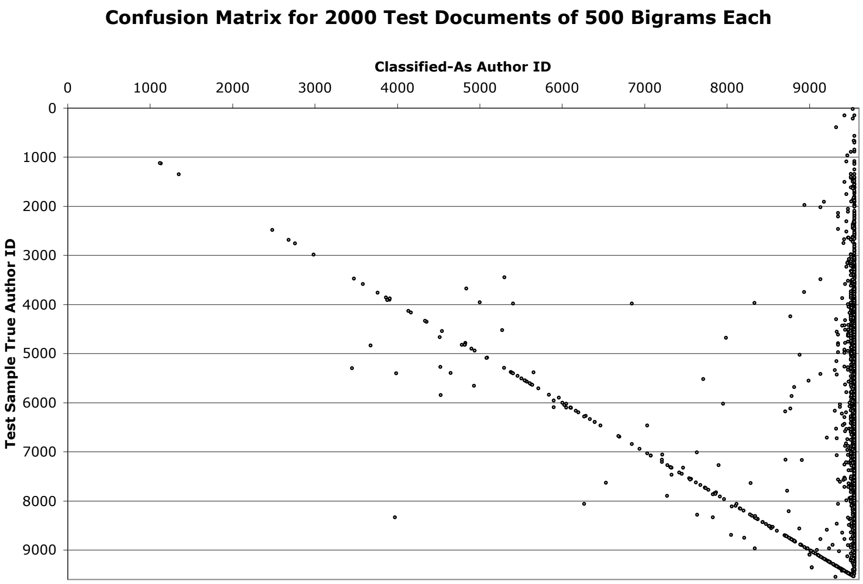


Figure 3.1: Confusion Matrix for All Test Documents of 500 Bigrams Each

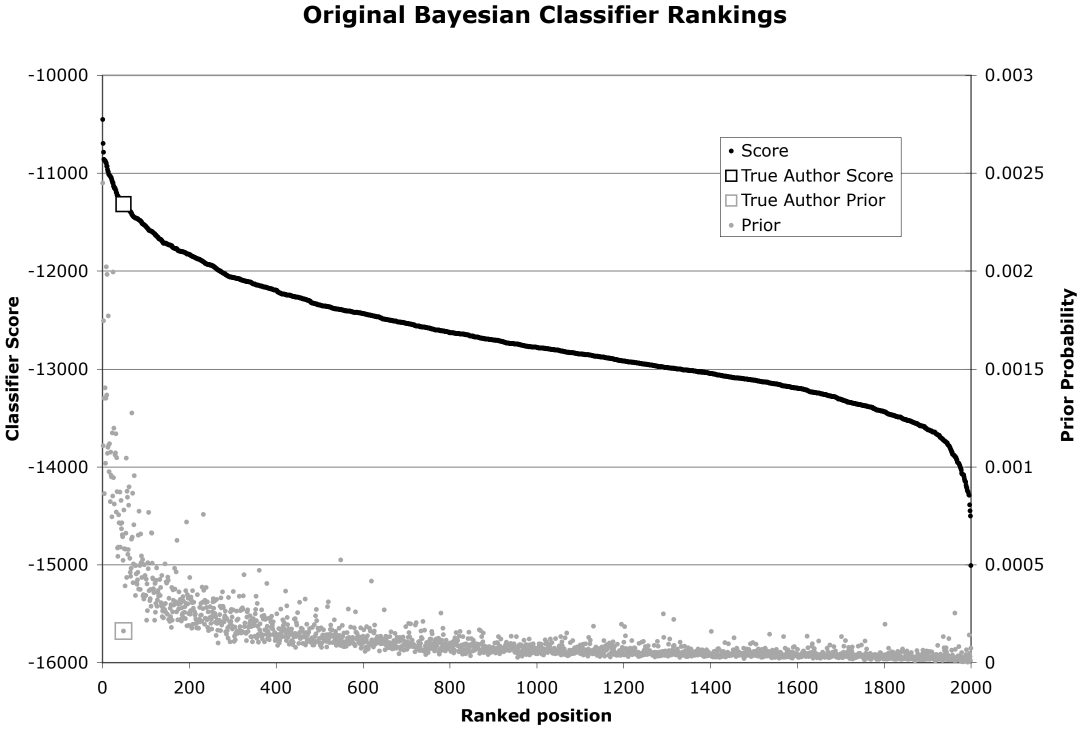


Figure 3.2: Full results of a single test, before transformation, ordered by score

Removing the prior probability term in the Bayesian classifier had little effect in correcting this influence. Instead, we normalized the scores and negative-log-priors such that the maximum (best) score became zero and the lowest score became -1, and the most prolific author had a log-prior of 1 and the least prolific had a log-prior of zero.

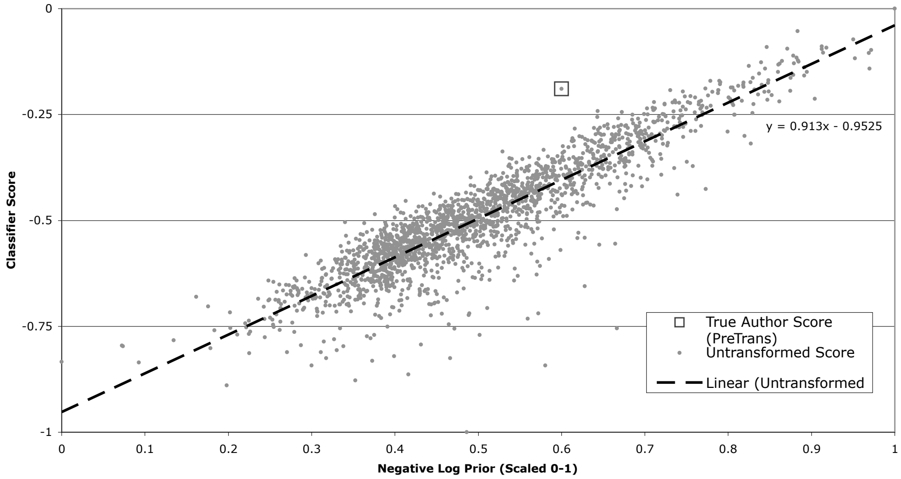


Figure 3.3: Normalized results of a single test, before transformation

Plotting the output from a single test in Figure 3.3 allows a least squares linear regression to be performed and a scaling factor $\hat{\beta}$, unique to the current test, to be obtained.

$$\hat{\beta} = \frac{\sum_k \left(\log P(a_k) - \overline{\log P(a)}\right)\left(S(a_k) - \overline{S(a)}\right)}{\left(\log P(a_k) - \overline{\log P(a)}\right)^2} \tag{3.11}$$

Using the slope of the regression line, $\hat{\beta}$, as a corrective factor allows a modified score to be calculated for each data point in the test results, shown in figure 3.4. The results can then be re-sorted on this new score and the most probable author determined by the maximum S'.

$$\begin{aligned} S'(a_i \mid \mathbf{F_t}) &= S(a_i \mid \mathbf{F_t}) - \hat{\beta} \log P(a_i) && (3.12) \\ &= (1 - \hat{\beta}) \log P(a_i) + \sum_{f_j \in \mathbf{F_t}} \log P(f_j \mid a_i) && (3.13) \end{aligned}$$

S' for each is a corrected score where we are essentially discounting more or less of the influence of that author's prior probability based on the scaling factor $\hat{\beta}$ as determined by the slope of the regression line through all authors' scores.

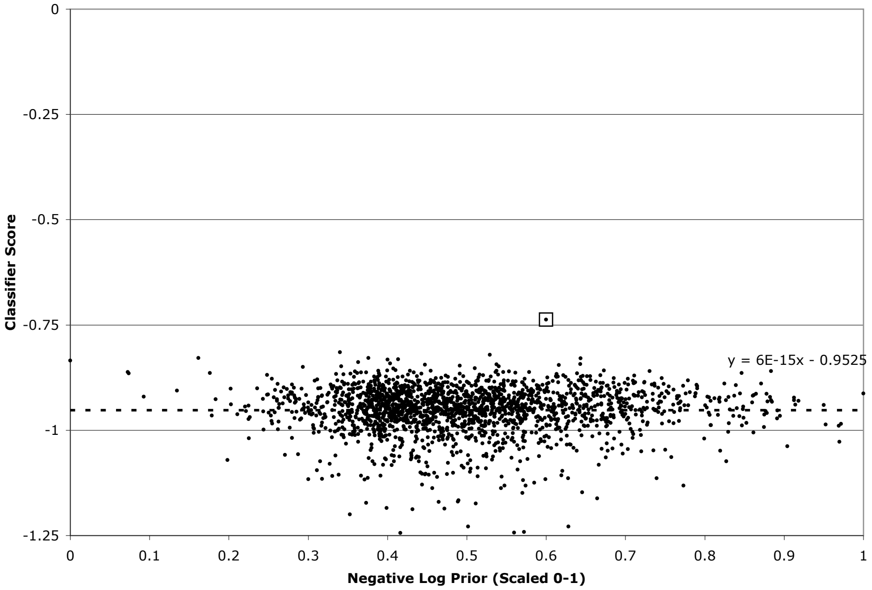


Figure 3.4: Normalized results of a single test, after corrective transformation. Compare to Figure 3.3

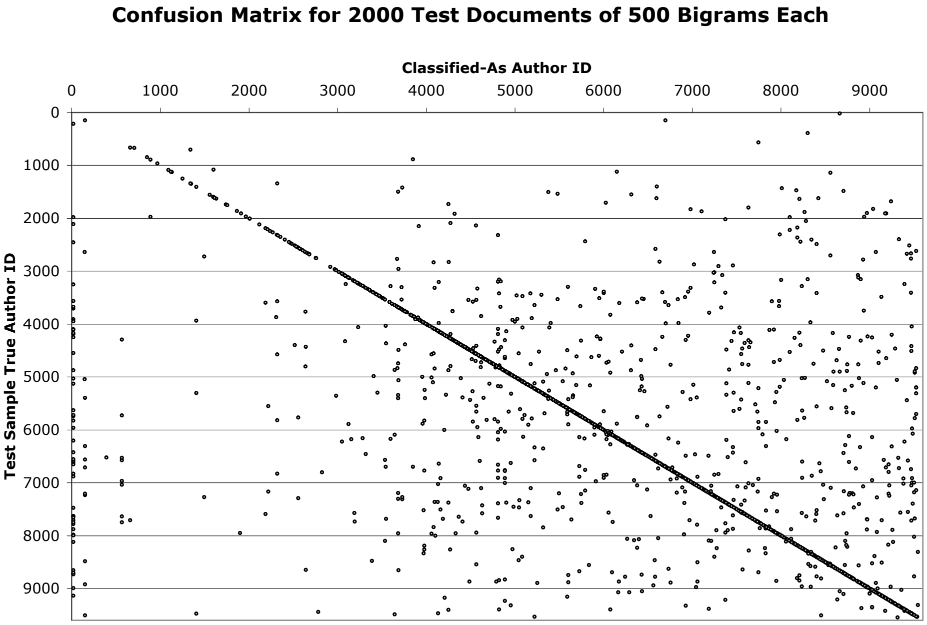
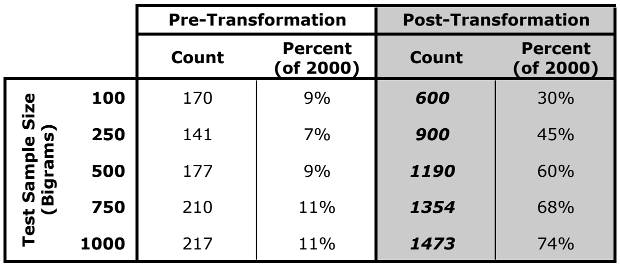


Figure 3.5: Confusion Matrix for All Test Documents of 500 Bigrams Each after corrective transformation. Compare to Figure 3.1

In figure 3.4, a plot of S' for the same example test as figure 3.3, the true author is assigned the highest S' and is, visually, clearly distinguishable. Examination of many test documents suggests that this pattern occurs with great regularity.

Comparing the confusion matrix from before corrective transformation, figure 3.1, to the confusion matrix of all tests after corrective transformation has been performed, figure 3.5, it is clear that the transformation not only dramatically improves classifier accuracy, with many more points on the diagonal, but it also removes the strong bias toward classifying all samples as one of the few authors with the highest priors.

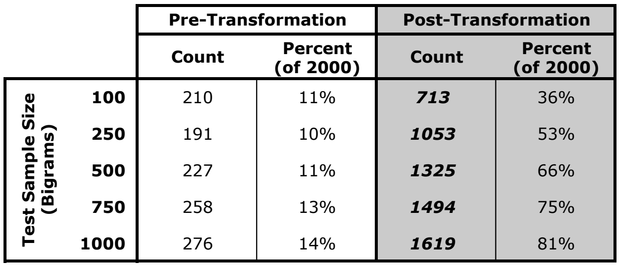


		Pre-Transformation		Post-Transformation	
		Count	Percent (of 2000)	Count	Percent (of 2000)
Test Sample Size (Bigrams)	**100**	170	9%	***600***	30%
	250	141	7%	***900***	45%
	500	177	9%	***1190***	60%
	750	210	11%	***1354***	68%
	1000	217	11%	***1473***	74%

Table 3.1: Count of authors classified exactly correctly among 2000 suspects, for 2000 test documents

3.6 N-Percent-Correct Threshold

In many cases it is desirable for an author with a score in some top threshold of all scores to be regarded as a "match" rather than the author with the single maximum score. This threshold could be used to reduce the search space of many thousands of potential authors to a few likely candidates with a high degree of certainty, making the job of a human investigator or more sophisticated classification much more manageable. Regarding the problem as a task of authorship *discovery* rather than authorship *verification*, the utility of this metric is apparent.



		Pre-Transformation		Post-Transformation	
		Count	Percent (of 2000)	Count	Percent (of 2000)
Test Sample Size (Bigrams)	**100**	210	11%	***713***	36%
	250	191	10%	***1053***	53%
	500	227	11%	***1325***	66%
	750	258	13%	***1494***	75%
	1000	276	14%	***1619***	81%

Table 3.2: Relaxing an "exactly correct" classification to tests where the true author was ranked *first or second*

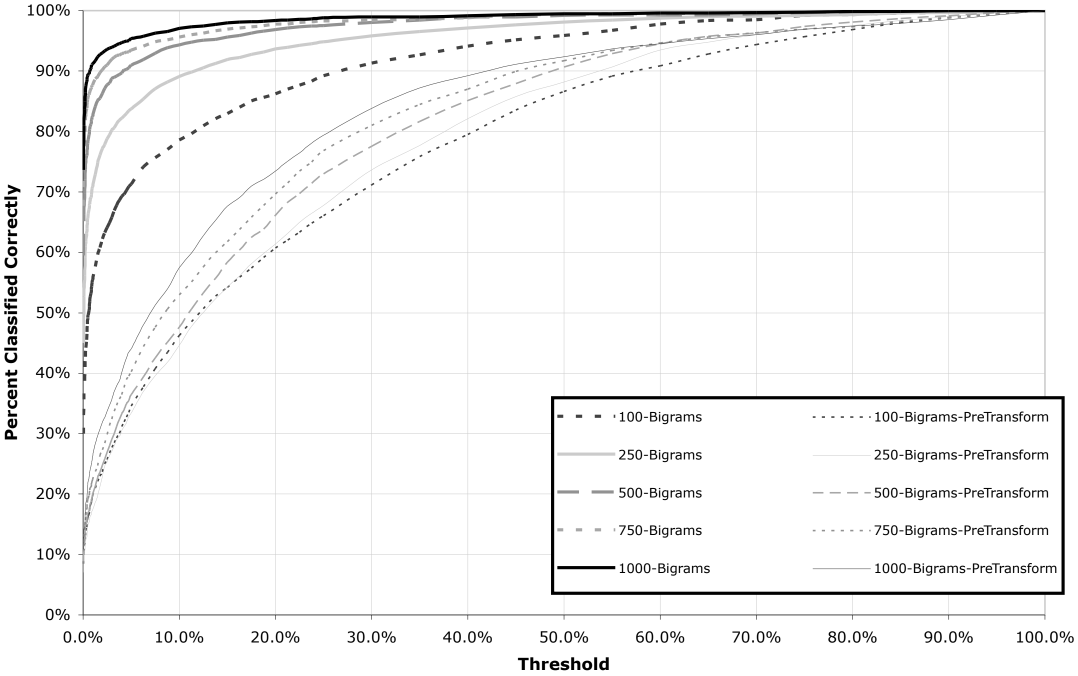


Figure 3.6: Classifier accuracy at progressively relaxing n-percent-correct threshold

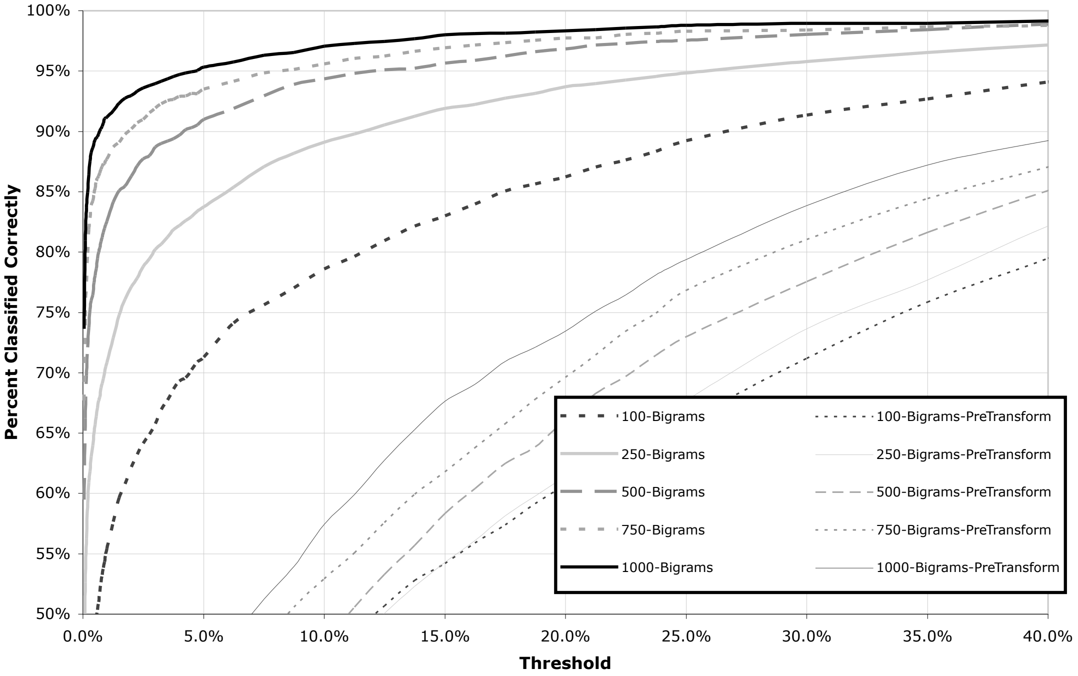
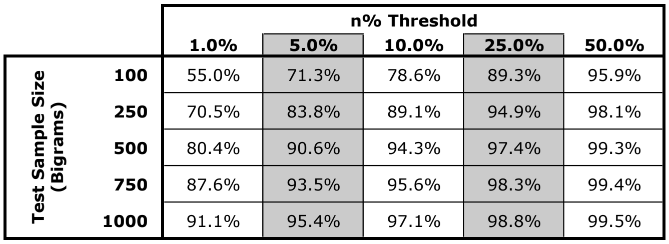


Figure 3.7: Detailed view of classifier accuracy at progressively relaxing n-percent-correct threshold

Table 3.2 indicates the number of correct classifications made among 2000 test samples if we define a correct classification as the true author being within the top $n = 0.1\%$ when sorted on S', that is, ranked first or second. Supposing we are willing to accept a new set of suspects that is a fraction the size of the original set[7], accuracy can be improved significantly.

The curves in figures 3.6 and 3.7 represent classifier accuracy as the n-percent threshold is progressively relaxed and as more or less testing data is available. For example, suppose we wish to identify the author of a sample of 500 bigrams from among 2000 possible authors. If we are willing to accept a new set of 100 possible authors, a reduction of the search space to just 5.0% of it's original size, the probability that the true author is among the new subset is 91.0%. If 1000 bigrams were available the probability of a "correct" classification rises to 95.4%. Even with 100 bigram test sample sizes, after transformation we can reduce the search space by half with greater than 95% assurance that the true author is in the new subset of potential authors.



		n% Threshold				
		1.0%	**5.0%**	**10.0%**	**25.0%**	**50.0%**
Test Sample Size (Bigrams)	**100**	55.0%	71.3%	78.6%	89.3%	95.9%
	250	70.5%	83.8%	89.1%	94.9%	98.1%
	500	80.4%	90.6%	94.3%	97.4%	99.3%
	750	87.6%	93.5%	95.6%	98.3%	99.4%
	1000	91.1%	95.4%	97.1%	98.8%	99.5%

Table 3.3: Classifier accuracy for progressively relaxing n-percent threshold, where classification within the top n-percent of all scores is considered "correct"

Figures 3.6 and 3.7 also include plots of classifier accuracy before transformation. For test sample sizes of 250 bigrams or larger the search space could be reduced by half with 90% or greater accuracy, but attempting to reduce the size further resulted in severely degraded accuracy. Only with the corrective scaling were we able to both limit the search space to a reasonably small size and to do so with a high degree of accuracy.

[7] We also discount the possibility that the true author may not be represented in the original search space at all.

CHAPTER 4: Discussion of Results

4.1 Effect of test sample size

One area in which this thesis differs from similar studies in authorship attribution is our examination of the influence of limited test sample sizes on classification accuracy. It is easily conceivable that practical authorship attribution problems would require methods that are accurate even on very short samples of text, such as might be found in a very short blog post, a comment on a blog, a short email, or a sentence or two appearing on a wiki. Past research has explored the possibility of authorship attribution where the very nature of the text is short, such as in poetry [32], but we are aware of none that addresses the possible degradation of accuracy in cases where only a few sentences are available.

The spacing of the curves in figure 3.6 reveals insight into the classifier's performance on smaller test sample sizes. As we increased the available test sample size from 100 to 250 to 500, and so on, the accuracy improved logarithmically, with diminishing returns from increased data.

This supports the obvious intuition that the more test data we can gather, the better our classifier's performance will be. The time penalty is not so significant that we would ever want to limit the size artificially for actual problems of determining authorship. It does also suggest, though, that this type of classifier is a good choice for situations where the available test data is very limited. A test sample size of 500 bigrams was often used as the baseline for comparison in this study, and represents a level where the search space may be reduced most dramatically with a high level of confidence, for example reducing the search space to 5.0% of it's original size with an accuracy over 90%, or classifying over half of the test samples exactly correct. 500 bigrams is, however, more text than may be available in many problems. For reference, this paragraph is less than 200 bigrams in length and is typical of the size of a single blog post in our corpus.

4.2 Inclusion of content words

We must point out that the inclusion of content words as features for the classifier may be a critical flaw in the application of this approach to real-world problems of authorship discovery. We suspect the results would not have been so positive had we restricted our study to function words or otherwise abstracted away the effect of topic and context. In particular, it is impractical or impossible to construct a large scale scenario where the classifier is trained on samples of one subject and tested on samples of text on a significantly different topic, but we suspect that attribution would be given to the authors of training samples which align more closely to the test on topic than on authorship.

For example, in an intelligence situation, suppose we wish to discover the author of an anonymous text sample that discusses detailed plans to use homemade chemical weapons. The true author of the message maintains a blog where he discusses his daily life but does not address his clandestine activities and therefore uses few of the same context specific words in his blog as were used in the test sample. Perhaps other indicative terms or idiosyncratic spelling would improve the true author's rank slightly, but his score would be quickly swamped by other bloggers such as legitimate chemical engineers who use the same context specific words in their training samples.

Of course, situations also exist in which matching the topic and context are advantageous, such as in a plagarism investigation. Suppose a sample of text from a paper is believed to have come from a blog source but does not give credit and does not reuse exact text strings from the original source, making it difficult to find the original source. Bayesian classification using all words as features would be more likely to reveal the source, causing it to emerge from the many blogs on other topics and hopefully returning it in a small subset of possible blogs which match most closely.

4.3 Corrective Scaling

Unfortunately the full explanation of why performing regression on the results of a classifier has such a dramatic effect is not known. In particular, we have not determined whether we can arrive at the same scaling factor, $\hat{\beta}$, through some other means or otherwise replicate its effect. For example, a more sophisticated back-off scheme for determining feature probabilities would reduce the occurrences of unseen bigrams, which we suspect would improve results without post-classification transformation and would be likely to make post-classification transformation less beneficial or entirely unnecessary. Other possibilities are that latent semantic analysis or some other transformation of the space before classification could have a similar effect.

The shortcoming of this method is that it requires accurate estimates of prior probabilities for all authors, which may not always be known. In our study, the prior probabilities were determined by the fraction of the entire training corpus that was attributed to each author in question. If this does not accurately represent the true prior probabilities of all authors, scaling the flawed priors by the factor $\hat{\beta}$ would not be likely to produce meaningful results.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
Summary and Future Work

5.1 Summary of Experiment in Bayesian Classification

In this experiment we constructed bigram word frequency models for 2000 authors at a time and used them to then classify an unseen test sample from each author. The classifier used was based on Bayes' rule, but used a scoring scheme to return authors in ranked order from most probable to least. Further, least squares regression on the scores from each classification test allowed us to compute a scaling factor, $\hat{\beta}$, which was used to discount each potential author's score. Reordering the results by the modified score produced dramatic improvements demonstrated in table 3.1.

We also introduce the concept of an 'n-percent-correct threshold.' When the list of suspect authors is extremely large, it is not only extraordinarily difficult to reliably classify a document in question to the single true author, but it may not always be necessary. Many cases exist where returning a subset of the original search space that is some n% the size of the original can be a very useful result. In figures 3.6 and 3.7 we demonstrate that as the threshold is relaxed, say from 1% to 5%, the cumulative percentage of test documents classified "correctly" within the threshold increases significantly. For large sample sizes, we are able to achieve 95% accuracy by defining a "correct" result as being classified within the top 5%, a reduction from 2000 possible authors to just 40 in our experiments.

As expected, when the test samples were allowed to be larger, classifier accuracy improved. However, the classifier performed well (90% accuracy when reducing the search space to 1/4 its original size, for example) even on samples as small as 100 bigrams. Additionally, there were diminishing returns with test sample sizes above 500 bigrams, suggesting that large samples are not required to use this technique effectively.

5.2 Practical Application of Techniques

Despite the possibility that inclusion of content words introduces a significant flaw into this technique, we believe it has a high level of utility for practical problems. Though 2000 suspects is significantly larger than any other known studies of authorship attribution, discovering an author among the approximately *113 million* or more active bloggers[1] is a thoroughly daunting task. Even if this technique cannot provide investigators with the single true author on the first pass, it certainly can give them a starting point for further investigation.

The scalability of this technique would need to be further examined and optimized before it could be used in real world situations. Building bigram frequency models for 10,000 authors took several hours of processing time, and classification of a document among 2000 suspects required 1-2 minutes per document in question. If the search space was enlarged beyond what could fit into memory, disk access delay caused classification times to degrade to several minutes per document in question. These are acceptable times in research, but building and storing models is not likely to be practical in a deployed system.

5.3 Future Directions

The potential for further study in this area is exciting and limitless. Unfortunately the influence of content words on this particular technique has dissuaded us from pursuing it further as it exists, but could spawn additional studies to determine the best method to abstract away the influence of topic and context.

5.3.1 Abstracting away topic and context influence

We have already begun investigating the possibility of tagging blog data with part-of-speech tags and building frequency models for POS n-grams. This abstraction would not only remove the actual content from language (e.g., abstracting the use of both ‘dog’ and ‘computer’ to the same token, ‘noun’) but would reduce the feature vector sizes by orders of magnitude and decrease the sparsity of an individual’s model.

As a consequence of compacting the feature vector sizes, we have been able to examine the use of Markov chains to model a particular individual’s language use with encouraging results. Comparing probabilities of suspect authors using first-order chains has not performed as well as simple chi-squared comparison of frequency distributions for POS n-grams

Early results also suggest that these techniques are better suited for smaller scale problems such as determining authorship among sets of suspects no larger than 100-200.

[1] Source: Blogs currently tracked by Technorati.com as of June 2008.

5.3.2 Further examination of authorship discovery

We believe additional techniques for authorship discovery should be explored in order to effectively and efficiently address the many real-world situations in which they could be useful. This could include application of a combination of

- Forensic techniques such as IP address geolocation from data logs
- Text statistic measures such as sentence or word lengths, in particular used as heuristics to quickly discount and eliminate potential authors from the search space
- Determining an author's age, location, education level or other metadata from the text of the blog itself if it is not provided by the author.
- Methods to quantify style such as parsing sentences and determining an author's preferential sentence structure or generative grammar rules
- Automatically discovering interconnectedness of suspect authors to the same entities as a document in question
- Discovering language use patterns which are likely to result from an author intentionally obfuscating their identity such as using words with similar meanings and connotations in two documents without repeating the actual grapheme itself.

Additionally, the determination of whether training and testing data must be from the same source could have significant implications for practical authorship discovery. For example, we suspect that an email could be classified accurately using training data from blogs, or that an addition to a wiki could be attributed by examining potential authors' blogs. Testing of this hypothesis would require a very specific corpus. Validation of the technique, though, could be of great use, particularly in criminal investigations where use of such a classification as evidence would require that it be accepted by the scientific community and that such a study be published in the scientific literature.

5.3.3 Study of stylochronometry in blogs

Further areas of interest which arise from authorship attribution studies in blogs include the possibility of studying stylochronometry, quantifying the progression of an author's writing style through the lifespan of his or her blog. This could be used to determine whether an author is likely to develop as a writer, perhaps improving, through the act of maintaining a blog and whether his or her progression is comparable to that of a writer from another genre. It may be possible, as well, to determine when an unknown sample was written in relation to dated blog posts. This type of study has been performed on classic literature, but blogs provide a convenient corpus for testing these techniques because it is easy to draw a test sample with an absolutely known date and time.

5.3.4 Discovering new blogs of interest

Finally, study of the language of blogs could be of great public interest. Authorship discovery techniques on a large scale may also be useful for discovering blogs of interest. For example, blogs which are similarly ranked by some metric could potentially share topical coverage, stylistic preferences, or both. A reader who “likes” one blog could use these techniques to find others which they may be interested in reading in a much more robust way than keyword searching or other currently available techniques.

Bibliography

[1] Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. An experiment in authorship attribution. *6es Journées internationales d'Analyse statistique des Données Textuelles*, 2002.

[2] C. E. Chaski. Who's at the keyboard: Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1), 2005.

[3] Malcolm Walter Corney. Analysing e-mail text authorship for forensic purposes, March 2003.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2001.

[5] Alvar Ellegård. A statistical method for determining authorship: the Junius Letters 1769-1772. Gothenburg Studies in English No. 13, Acta Universitatis Gothenburgensis, 1962.

[6] Eric N. Forsyth. Improving automated lexical and discourse analysis of online chat dialog. Master's thesis, Naval Postgraduate School, September 2007.

[7] RS Forsyth, DI Holmes, and EK Tse. Cicero, Sigonio, and Burrows: investigating the authenticity of the Consolatio. *Lit Linguist Computing*, 14(3):375–400, 1999.

[8] Wilhelm Fucks. On mathematical analysis of style. *Biometrika*, 39(1/2):122–129, 1952. ISSN 00063444.

[9] Michael Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 611. Association for Computational Linguistics, Morristown, NJ, USA, 2004.

[10] Felix Golcher. A new text statistical measure and its application to stylometry. In *Proceedings of Corpus Linguistics*. University of Birmingham, 2007.

[11] Jack William Grieve. Quantitative authorship attribution: A history and an evaluation of techniques. Master's thesis, Simon Fraser University, Summer 2005.

[12] Susan C Herring. Weblogs as a bridging genre. *Information Technology and People*, 18: 142–171(30), 2005.

[13] D. I. Holmes. The analysis of literary style–a review. *Journal of the Royal Statistical Society. Series A (General)*, 148(4):328–341, 1985. ISSN 00359238.

[14] D. I. Holmes. A stylometric analysis of mormon scripture and related texts. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 155(1):91–120, 1992. ISSN 09641998.

[15] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32(4):485–525, 2006. ISSN 0891-2017.

[16] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution, 2003.

[17] M. Koppel, J. Schler, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*. American Association for Artificial Intelligence, 2006.

[18] Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM, New York, NY, USA, 2004. ISBN 1-58113-828-5.

[19] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660. ACM, New York, NY, USA, 2006. ISBN 1-59593-369-7.

[20] Jane Lin. Automatic author profiling of online chat logs. Master's thesis, Naval Postgraduate School, March 2007.

[21] Edward Loper and Steven Bird. Nltk: The natural language toolkit, May 2002.

[22] T. Mala and T. V. Geetha. Visualizing author attribution using blobby objects. In *CGIV '07: Proceedings of the Computer Graphics, Imaging and Visualisation*, pages 460–464. IEEE Computer Society, Washington, DC, USA, 2007. ISBN 0-7695-2928-3.

[23] M. Malyutov. Authorship attribution of texts: A review. *General Theory of Information Transfer and Combinatorics*, pages 362–380, 2006.

[24] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 6th edition, 2003.

[25] Conrad Mascol. Curves of pauline and pseudo-pauline style i. *Unitarian Review*, 1888.

[26] T. C. Mendenhall. The characteristic curves of composition. *Science*, 9(214):237–246, 1887. ISSN 00368075.

[27] Gilad Mishne. *Applied Text Analytics for Blogs*. PhD thesis, Universiteit van Amsterdam, 2007.

[28] Frederick Mosteller and David L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963. ISSN 01621459.

[29] Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1st edition, 1964.

[30] Scott Nowson. *The Language of Weblogs: A study of genre and individual differences*. PhD thesis, University of Edinburgh, 2006.

[31] Alex Riba and Josep Ginebra. Diversity of vocabulary and homogeneity of style in *Tirant lo Blanc*. *7es Journées internationales d'Analyse statistique des Données Textuelles*, 2004.

[32] Conrad Sanderson and Simon Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 482–491. Association for Computational Linguistics, Sydney, Australia, July 2006.

[33] Diane J. Schiano, Bonnie A. Nardi, Michelle Gumbrecht, and Luke Swartz. Blogging by the rest of us. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 1143–1146. ACM, New York, NY, USA, 2004. ISBN 1-58113-703-6.

[34] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[35] G. Udny Yule. On sentence- length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3-4):363–390, 1939.

[36] H. Zhang. The optimality of naive bayes. In *Proceedings of the 17th International FLAIRS conference*, 2004.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 6:
Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudly Knox Library
   Naval Postgraduate School
   Monterey, California