NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AUTHORSHIP ATTRIBUTION IN THE E-MAIL DOMAIN: A STUDY OF THE EFFECT OF SIZE OF AUTHOR CORPUS AND TOPIC ON ACCURACY OF IDENTIFICATION

by

Kori Levy-Minzie

March 2011

Thesis Advisor:	Craig Martell
Second Reader:	Joel Young

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
2011-03-25	Master's Thesis	2009-03-01—2011-03-25

4. TITLE AND SUBTITLE

Authorship Attribution in the E-mail Domain: A Study of the Effect of Size of Author Corpus and Topic on Accuracy of Identification

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kori Levy-Minzie

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

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

Department of the Navy

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

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. **IRB Protocol Number: n/a**

14. ABSTRACT

We determined that it is possible to achieve authorship attribution in the e-mail domain when training on "personal" e-mails and testing on "work" e-mails and vice versa. These results are unique since they simulate two different e-mail addresses belonging to the same person where the topic of the e-mails from the two different addresses do not intersect. As we only used one classification technique, these results are preliminary and may serve as a baseline for future work in this area. The corpus of data was the entirety of the Enron corpus as well as a subsection of hand-annotated work and personal e-mails. We discovered that there is enough author signal in each class to identify an author in a sea of noise. We included suggestions for future work in the areas of expanding feature selection, increasing corpus size, and including more classification methods. Advancement in this area will contribute to increasing cyber security by identifying the senders of anonymous derogatory e-mails and reducing cyber bullying.

15. SUBJECT TERMS

Machine Learning, Supervised Learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	UU	77	

Standard Form 298 (Rev. 8–98)
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

AUTHORSHIP ATTRIBUTION IN THE E-MAIL DOMAIN: A STUDY OF THE EFFECT OF SIZE OF AUTHOR CORPUS AND TOPIC ON ACCURACY OF IDENTIFICATION

Kori Levy-Minzie
Lieutenant , United States Navy
B.S., Florida Agricultural and Mechanical University, 2004

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
March 2011

Author: Kori Levy-Minzie

Approved by: Craig Martell
Thesis Advisor

Joel Young
Second Reader

Peter J. Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

We determined that it is possible to achieve authorship attribution in the e-mail domain when training on "personal" e-mails and testing on "work" e-mails and vice versa. These results are unique since they simulate two different e-mail addresses belonging to the same person where the topic of the e-mails from the two different addresses do not intersect. As we only used one classification technique, these results are preliminary and may serve as a baseline for future work in this area. The corpus of data was the entirety of the Enron corpus as well as a subsection of hand-annotated work and personal e-mails. We discovered that there is enough author signal in each class to identify an author in a sea of noise. We included suggestions for future work in the areas of expanding feature selection, increasing corpus size, and including more classification methods. Advancement in this area will contribute to increasing cyber security by identifying the senders of anonymous derogatory e-mails and reducing cyber bullying.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

Acknowledgements

There are many people who made this thesis process much less painful than it could have been and I would like to acknowledge them here. First and foremost, I would like to thank my thesis advisor, Dr. Craig Martell, for his guidance and encouragement. Your calming, yet vigorous enthusiasm helped to keep me motivated throughout the process, which was greatly appreciated. I would also like to thank my second reader, Lieutenant Colonel Joel Young, USAF. You managed to scare me into focusing on my writing, which I hope made me sound more academic in my work.

I must also take this opportunity to thank my intern for a summer, Eric Johnson. Without your diligent work annotating the e-mails that was integral to the most interesting part of this thesis, the quality of this work would be much lower. I must also thank our lab assistant, Constantine Perepelitsa, for his vast Unix and regular expression knowledge, which helped get many menial tasks completed in mere seconds.

I would also like to thank my friends in my cohort. Thank you all for keeping me motivated throughout these past two long years. It was helpful knowing that it wasn't just me struggling through some of our classes and the thesis process. I would especially like to thank Randy Honaker for his seemingly daily LaTex knowledge bits. Finding my way through this program would have been much more difficult without you struggling before me and then sharing how to fix it.

I would like to thank my parents Tangerine Levy and Rudolph Minzie as well as my sister Keisha Levy-Minzie for their many phone calls and text messages of encouragement. It was helpful to know that I was making someone proud by finishing my degree. (To clarify, it was my sister sending me text messages since my parents have not yet learned how to use that technological advance.)

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Due to new forms of social media such as e-mail, text messages, blogs, Twitter, and Facebook, the relevance of authorship attribution is clear. These new forms of media not only bring about a new way for people to express themselves and be creative, but also ways to manipulate and harass anonymously. Authorship attribution in older forms of text has been shown to be feasible in earlier work. However, with the advent of new social media and new forms of textual information, there are many new domains that might benefit from authorship attribution research. Particularly, helping to identify authors of anonymous e-mail. If we can create a model of an author from known e-mails and identify that same author using unknown e-mails, we will be taking a large step forward in battling cyber crimes or tracking terrorist social networks.

1.1 Motivation

If it were possible for us to match known authors to e-mails with unknown authors with a high degree of certainty, we could give authorities a new tool to use in the fight against cyber criminals. People who try to commit these acts of cowardice by hiding behind anonymous e-mail address and screen names may think twice about creating those accounts if there were a larger threat of their getting found out.

1.2 Related Work

There has been a lot of work done in areas related to authorship attribution. Much of that work has been done in the e-mail domain using the Enron e-mail corpus. It is the largest naturally occurring publicly available e-mail corpus, and we will use it in our research as well. However, even with the many ways this corpus has been divided and annotated, there has been little work done in the area of delineating *work* e-mails from *personal* e-mails.

1.3 Research Question

This thesis address two research questions. The first question we address is "How feasible is it to do authorship attribution in the e-mail domain using the most simple methods when we have a sufficiently large corpus of data per author?" The second question is "How feasible is it to differentiate between personal e-mails and work-related e-mails given a small data set and

simple classification methods?" The research done in this thesis uses two feature sets and one classifier to answer these questions.

1.4 Results

The results of our research allow us to answer both questions with a "very feasible." We got very high f-scores and accuracies, which leads us to believe we going down the right track. We are far from the end of that track though, since we used such a small data set for one set of our experiments and only used two discriminating features. However, even with such a simple experimental set-up, our results were definitive and telling.

1.5 Future Work

Future work for this domain includes growing the feature set, annotating more data, and tweaking classification techniques. The features used in this research are *unigrams* to create the initial data set, and *e-mail category* in order to simulate to different e-mail addresses belonging to the same author. Using other features sets such as bigrams or e-mail length, for example, for future research may produce better results.

More annotation is necessary to further research in this domain as well. If we try to simulate multiple e-mail addresses for one author using the Enron corpus, we need to create a annotated set of e-mails that is larger than our small test set. We only used a small portion of one author's e-mails.

Our research centered around one classifier, namely naive Bayes. This same work could be done with a number of other classifiers in order to compare results and find the best classification technique for the job.

1.6 Organization of Thesis

The remainder of this thesis is organized in the following way:

- Chapter 2 discusses prior work as it relates to this thesis.
- Chapter 3 describes our experimental design and the data set used for this research.
- Chapter 4 contains the results of our experiments along with analysis of those results.
- Chapter 5 presents our closing remarks and possible areas for future research.

CHAPTER 2:
Prior and Related Work

2.1 Related Work

In this chapter, we review background material motivating and enabling the results presented in this thesis. First, we outline key contributions in the history of authorship attribution. Next, we discuss our key classification technique, naive Bayes, which leads us into a discussion of methods of smoothing.

2.1.1 History of Authorship Attribution

In 1887[1], T.C. Mendenhall published the results of his scientific study of authorship attribution. His approach focused on syntactic characteristics of sample texts from famous literary authors such as Charles Dickens. His idea was an extension of Augustus DeMorgan's hypothesis that comparing mean word length in two texts could indicate whether they were written by the same individual. Mendenhall offered the idea that comparing histograms of word lengths would better illustrate minute differences between authors. Mendenhall called these histograms "characteristic curves of composition" [2]. He hypothisized that authors had uniquely identifable writing styles and these styles would be displayed by comparing each characteristic curve. He compared his process to that of spectral analysis of objects to determine the presence of specific elements. When heated correctly, each element emits a unique light signature that can be used to identify it.

Given the time frame in which Mendenhall's experiments were conducted, his limited results are understandable. In order to generate his characteristic curves, he had to manually count letters used in groups of words from classic authors. This was a time-consuming step. His findings indicate that given a large enough sample for each author, characteristic curves that are sufficiently discriminatory can be produced. Mendenhall also recognized that the benefit of his approach is that it is purely mechanical. Mendenhall claimed his approach could also be applied to other characteristic counts such as word counts per sentence or counts of syllables.[2]

[1] There have been many studies of authorship attribution and an equal number of historical discussions of prior work. This section was influenced by [1]

In 1939, G.U. Yule used histograms in a similar method for authorship attribution. Yule's method focused on sentence length as the discriminating characteristic. He created tables of sentence length distribution of an author in a specific sample of text. The counts of sentence length were grouped by fives, so sentences of length one to five were counted together, six to ten were together and so forth. From these counts, Yule calculated the mean sentence length for the text. He believed that the mean sentence length was enough to identify an author. He computed the mean sentence length for a piece of text, Imatatio Christi, whose authorship had long been disputed. He then computed the mean sentence length for the two people believed to be the authors, Thomas A. Kempis and Jean Charlier de Gerson. He concluded the mean sentence lengths for Imatatio Christi, Kempis, and Gerson were 16.2, 17.9 and 23.4 respectively. Therefore, he classified Kempis as the author, since his mean sentence length matched more closely than Gerson[3].

Conrad Mascol used similar evaluation techniques on the New Testament Epistles by measuring sentences per printed page. Using this method, he concluded that Paul had not written some of the books previously believed to be penned by him by many scholars.[4]

Wilhelm Fucks discriminated between authors using the average number of syllables per word and average distance between equal-syllabled words. He concluded that his method of analysis revealed 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.[5]

The measure of syllable length was utilized by R. Forsyth, D. Holmes, and E. Tse to decide that portions of a version of Cicero's Consolatio, were likely faked. They concluded those portions used language more characteristic of the Renaissance than the Classical time period.[6]

2.2 Stylometric Approaches to Authorship Attribution

Many other textual measures have been proposed that go beyond word and sentence length histograms or syllable counts. In [7], 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.

One of the most prevalent measures in this category is the type-token ratio, that is, the number of unique word types, V , divided by the counted length of the text sample, N. In other words, this is a measure of the scope of the author's vocabulary used in the sample of interest. Unfortunately, the type-token ratio has limited use in authorship studies due to the fact that it 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 is, however, an easily understood starting point for understanding the quantification of an author's style.

Another stylometric measure that is useful is word frequency distribution. George Kingsley Zipf is known for his work regarding word frequency distributions in text. Specifically, Zipf's law gives us a"rough description of the frequency distribution of words in human languages: there are a few very common words, a middling number of medium frequency words, and many low frequency words." [8] Zipf's law states that given a corpus of text, the frequency of any word in the text is inversely proportional to the rank of that word in the frequency table of the same corpus. The simplest way to depict this is by plotting on a log-log graph where the x-axis is the log of the rank order and the y-axis is the the log of the frequency as shown in figure 2.1.

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 more effective when used in conjunction with other measures.

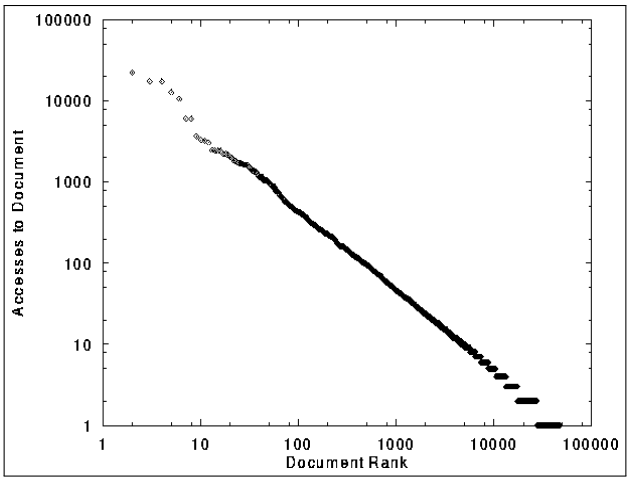

Figure 2.1: Zipf's Law Graphical Representation

Borrowing from thermodynamics, we utilize the concept of entropy, H, defined as the following:

$$H(P) = -\sum_{x=1}^{N} P(x) lg P(x)$$

where x ranges over words in the vocabulary, lg is log_2 and P is the probability distribution over the vocabulary. In this context, entropy is a measure of predictability of the next word; the lower the entropy, the more predictable. Note that the highest entropy is when all words are equally likely.

2.3 Lexical Approaches to Authorship Attribution

The above approaches to a more stylometric way to do authorship attribution try to generalize a text sample based on the statistics of its construction. However, a different approach would be to examine the distribution of the actual words, and their comparative usage between texts. In most cases, these lexical techniques do not approach the level of semantic analysis, where the words have some *meaning* to the classifier. Instead the strings themselves are simply counted.

In their landmark 1963 and 1964 studies on the Federalist papers, [10], Mosteller and Wallace examine the Federalist papers with statistical analysis of word frequencies. According to [10], 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 Consiti-

tution. Of the 77 essays 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, 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.

In [9], Ellegard used the Junius Letters as his corpus to exercise his extremely labor intensive method to building word frequency distributions for determining authorship. He built a "distinctiveness" measure that is similar in concept to *tf-idf* (described below), where words that appeared frequently (and similarly infrequently) in the known works of the authors in question, but did not appear as frequently in the works of opposing authors, were highly weighted. Next, Ellegard manually counted each of these weighted "plus" and "minus" words in each of the Junius Letters for each of the authors. Then he obtained a similarity score for each author over each document. Ellegard concluded that Sir Phillip Francis, the suspected author of the letters, was the author. However, even what seems to be a sound approach to authorship identification has its' flaws. Ellegard included content words in his lists of "plus" and "minus" words. In terms of tf-idf, it has become commonplace to consider a word with a high score to be distinctive of the primary topic of a document. Since this is what Ellegard's method is doing, it has the potential to match two distinct authors who are writing about similar topics rather than one author writing about a different topic.

Term frequency - inverse document frequency (tf-idf) is method used for data mining in text that ranks the importance of a word in a document by assigning a weight. The weight is calculated using the following set of equations:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.1}$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{2.2}$$

$$(tf - idf)_{i.j} = tf_{i,j} \times idf_i \tag{2.3}$$

where in 2.1, the numerator is the number of occurrences of the term and the denominator is the

size of the document and in 2.2, the numerator is the number of documents in the corpus and the denominator is the number of documents the term appears in.

2.4 Machine Learning Techniques

This section provides an overview of relevant techniques, naive bayes and smoothing. These techniques are those used for our experimental design and analysis.

2.4.1 Naive Bayes

Naive Bayes is a probabilistic binary classifier. The "naive" part of the classifier comes from the strong independence assumption made about the features. Without this independence assumption, a naive Bayes classifier simply becomes a Bayesian classifier. The basis of a Bayesian classifier is the Bayes equation:

$$P(C|\hat{F}) = \frac{P(\hat{F}|C)P(C)}{P(\hat{F})} \tag{2.4}$$

where C is a class and $\hat{F}$ is a feature vector.

Since this is a probabilistic classifier, our goal is to find the argument that when plugged in, maximize 2.4. That is:

$$\arg\max_{C} \frac{P(\hat{F}|C)P(C)}{P(\hat{F})} \tag{2.5}$$

It is important to note that as we progress through the classes, since the feature vector does not change, the $1/P(\hat{F})$ term will always remain constant. We can then name that constant α. Rewriting 2.5 using the new term α yields:

$$\arg\max_{C} \alpha P(\hat{F}|C)P(C) \tag{2.6}$$

Note that the argument that maximizes a function also maximizes a scalar times that same

function. Because of this, we can pull the scalar α out of the argmax and rewrite 2.6 as:

$$\arg\max_{C} P(\hat{F}|C)P(C) \tag{2.7}$$

At this point, we have a Bayesian classifier. As was discussed earlier, what makes this a naive Bayes classifier is the strong independence assumption that the features in the feature vector are conditionally independent given the class, which lets us put 2.7 in the form:

$$\arg\max_{C} P(C)\prod_{i=1}^{n} P(f_i|C) \tag{2.8}$$

2.4.2 Smoothing

Naive Bayes classifiers work well when all features in the test data have been seen before in the training data. However, their shortfalls become obvious when a term in the test data has not been seen in the training data for that particular author. If that happens, the probability of that term given that author is zero. That zero causes the probability for the entire document to be zero, which is unrealistic. The more realistic case is that we did not capture the author's entire vocabulary in the training data. One common approach to solving this problem is smoothing. Smoothing is the process of taking a small portion of the probability mass that we have seen and distributing it to the zero count or low count terms.

The simplest way to smooth is to add some number to every count in the data. This method is called Laplace, or Add-K smoothing where K can be any positive number. A common K is one, where one is added to every count so any term that wasn't seen is treated as having been seen once. All terms that had a count other than zero, now have a count one greater than before. To get probability, we normalize by the original token counts plus the size of the vocabulary, since one was added for each vocabulary item.

$$P_\ell(w_i) = \frac{C(w_i) + 1}{N + V}$$

The downside to Laplace smoothing is that it takes too much probability mass from the high count terms and gives too much probability mass to the zero and low count terms, causing it to not perform as well as other smoothing techniques that are more stingy about probability mass redistribution[11].

Witten-Bell smoothing outperforms Laplace smoothing while remaining relatively simple to implement[12]. Witten-Bell smoothing estimates the probability of an unseen word based on the frequency that we have seen new words in the past[12]. The unigram formula, from[13], is:

$$P_{WB}(w_i) = \frac{C(w_i)}{N + T} \tag{2.9}$$

if

$$C(w_i) > 0$$

$$P_{WB}(w_i) = \frac{T}{N + T} * \frac{1}{Z} \tag{2.10}$$

if

$$C(w_i) = 0$$

$C(w_i)$ = count of word w_i (number of tokens).
T = number of distinct words (types).
N = total number of words (tokens) seen.
Z = estimated number of unseen words.

Without the Z term, the second formula is the total probability mass assigned to unseen words. The Z term is used to determine how much probability each occurrence of a new word is assigned. All the above counts refer to what has been seen in the training data for this author.

The formula for bigrams, from[14], is:

$$P_{WB}(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{N(w_{i-1}) + T(w_{i-1})} \tag{2.11}$$

if

$$C(w_{i-1}, w_i) > 0$$

$$P_{WB}(w_i|w_{i-1}) = \frac{C(w_{i-1})}{N(w_{i-1}) + T(w_{i-1})} * \frac{1}{Z(w_{i-1})} \tag{2.12}$$

if

$$C(w_{i-1}, w_i) = 0$$

$C(w_{i-1}, w_i)$ = count of bigrams consisting of word w_{i-1} followed by word w_i
$T(w_{i-1})$ = Number of distinct words (types) seen to the right of word w_{i-1}.
$N(w_{i-1})$ = Total number of words (tokens) seen to the right of word w_{i-1}.
$Z(w_{i-1})$ = Estimated zero counts; the number of bigrams starting with w_{i-1} that do not occur in the training set. If V is the number of words (unigram types) in the vocabulary, then $Z(w_{i-1}) = V - T(w_{i-1})$.

The bigram formula can easily be extended to arbitrary length n-grams, by replacing (w_{i-1}) with $(w_{i-n+1} \dots w_{i-1})$.

The disadvantage of the bigram and n-gram versions of Witten-Bell smoothing is that if the preceding words do not occur in the training data, the smoothed probability is zero[14]. In other words, if we have never seen the preceding terms, the number of words (tokens) and distinct words (types) seen follow those terms in zero $(N = T = 0)$. This problem can be solved with back-off, such as in the formulas described in[12] and[8], however this adds to the complexity of the implementation.

2.5 Metrics

In this thesis, we will use metrics to measure how well our classifier is completing the task at hand. The metrics we will use are based on combining the counts of true positives (tp), true

negatives (tn), false positives (fp) and false negatives (fn). The metrics of interest are accuracy, precision, recall, and f-score.

2.5.1 Accuracy

Accuracy measures the proportion of items correctly classified. It is defined as:

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{2.13}$$

Using a confusion matrix, accuracy is the sum of the values on the diagonal of the matrix divided by the sum of all values in the matrix.

2.5.2 Precision

Precision measures the proportion of things labelled X that actually are X. In other words, if we are looking for e-mails written by author A, precision measures how many of those labelled as being written by author A actually were. Precision is defined as:

$$precision = \frac{tp}{tp + fp} \tag{2.14}$$

2.5.3 Recall

Recall determines how well a classifier is at correctly classifying all possible true items. In other words, recall measures what proportion of the actual emails written by A did we label as "A". Recall is defined as:

$$precision = \frac{tp}{tp + fn} \tag{2.15}$$

2.5.4 F-score

F-score is the harmonic mean of precision and recall. The harmonic mean is used, since it punishes an increase in one dimension at the expense of another. F-score is defined as:

$$f - score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{2.16}$$

2.6 Conclusion

In this chapter, we reviewed material relevant to the history of authorship attribution. We discussed the classification model used in this thesis. Finally we reviewed smoothing techniques that were used with our classifier. Next we will describe our experimental set-up.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:
Experimental Design

3.1 Introduction

In this chapter, we discuss all phases of our experimental design. First, we will present a description of the data. Next, we will describe the process used for converting the raw data into a usable format for machine learning purposes. Then, we discuss the features selected for the experiments. Finally, we present the details of experimental setup for each experiment.

3.2 Description of Data

The data used for these experiments comes from the Enron E-mail Corpus. The corpus is plain text e-mails from 150 executives in the company at the time of the investigation. There are many versions of the corpus since many people have used this same data for other natural language experiments. The iteration of the corpus that we are using is tagged by author. Every e-mail in the corpus is in the folder of the person who wrote it.

3.3 Converting Raw Data

In order to start running experiments, the data needed to be cleaned up to get it into a usable format. First we removed the header information at the beginning of each e-mail. This was accomplished using the NLTK package in python. We used a function that looked for the end of the header of the e-mails, essentially isolating the body of the e-mail with out including any text from the header and tokenizing it. Then, we converted each e-mail body to a word count vector in the NPSML format.

3.4 NPSML

NPSML stands for Naval Postgraduate School Machine Learning. NPSML is a line oriented file format where there is one record per line in the file. Each entry on a line is delimited by a single space. The first three columns in the line contain the identifying information about the record. Column one is a unique key identifying the origin of the line. Column two is a weight, which, for our experiments, is always set to 1.0. Column three is the category label. The category label should be set to NA for an unsupervised task. The remaining columns of the record are a series

of feature-value pairs, where the feature is the name of the feature and the value is a floating point count of that feature.

3.5 Data Segmentation

After the data was in NPSML format, we built data to run a preliminary experiment, using just five of the authors for identification. We choose to use the five authors in the corpus that came first alphabetically. All of the e-mails written by these authors had to be grouped to create the full data set. Our python script concatenated each author's individual NPSML file into one big file. For the later experiments, we used all data from all authors so we again concatenated NPSML files into one large file that included all data in the corpus.

3.6 Cross Validation

Next we split the data into test and training sets for the classifier. We shuffled the large NPSML files randomly. We used 5-fold cross validation to get the test and training sets. Each test set was 20% of the total number of e-mails in the corpus. Each training set was the other 80% in that iteration.

3.7 Classification

The classification for all experiments was done using the naive Bayes package created in the Naval Postgraduate School Natural Language Processing Lab. These tool take an NPSML training file as input and generate a naive Bayes model as output. The learning part of the process can use either Laplace add one-smoothing, Witten-Bell, or Good-Turing smoothing. For our experiments, we only used Laplace add-one smoothing. The classification part of the process takes the model generated by the learning portion and a NPSML test file as input. The output was two column text where the first column was the truth of what the data should be classified as and the second was what the classifier predicted.

3.8 Creating Work vs. Personal Data Set

An important aspect of our work is determining how critical topic is to authorship attribution. More specifically, if topics related to the work environment and topics related to one's personal life can equally determine authorship. To begin to evaluate this, we needed a corpus of text that is classified as being related to work or to one's personal life. The Enron e-mail corpus was not labelled in this way.

Using the most prolific author in the corpus, we annotated each e-mail for that author as being work related or personal. This author was known as Mann-k in the corpus. Mann-k had 4,443 e-mails in the corpus. We annotated approximately 33% of that total where 1411 of those were labeled as work and 292 were labelled as personal. We labelled 12 as ambiguous. These e-mails contained words that could be considered related to either the work or personal group. 44 e-mails we labelled as other. These e-mails were ones that seemed to have no relevance either way or were a simple one or two word reply to a previously received message in the thread. This gives us a total of 1759 annotated e-mails.

Using those labelled work and personal data sets, we ran two sets of experiments. To create the data set, we used the entire labelled work file and personal file, concatenated, shuffled, and split as described in the earlier sections. The data set of the next experiment was created using random under-sampling.**??** Random under-sampling is a method used to force the number data points in each set to be equal. In our case, we had more data points in our work set. The number of times we randomly sampled from the work set was equal to the total number of data points in the personal set. Then we ran the same experiment described above with this new data set.

3.9 Experiments

- First Five Authors (FFA): Five authors in the data set, objective was to identify each of the five authors
- First Five Authors Plus Other (FFAPO): Same five authors from FFA plus the e-mails from the other 145 authors in the corpus classified as ”other” to simulate noise. Objective was to identify each of the five authors.
- Most Prolific Five Authors Plus Other (PFAPO): Same setup as FFAPO except the five authors were now the most prolific writers in the corpus. Objective was to identify each of the five authors.
- Work versus Personal (WvP): Took the work and personal e-mails from the most prolific writer in the corpus as the data set. Objective was to correctly identify work and personal e-mails.
- WvP Randomized Under-Sampling (WvPRUS): Same setup as WvP except manually forced the number of work e-mails in the data set to match the number of personal e-mails by randomizing the work e-mails and then picking the first 583 of them.

- Train on Work-Other, Test on Personal-Other and Vice Versa (TWTPVV): Tried to identify the most prolific author by training on a mix of the work e-mails and random e-mails from all other authors in the corpus and testing on a mix of personal e-mails and random e-mails from all other authors in the corpus. Essentially, this is an attempt to identify an author in a sea of noise by training on work or personal and testing on the other.

3.10 Conclusion

In this chapter, we outlined a description of the raw data and the process used to convert the raw data into workable data that could be used in our experiments. We also discussed the set-up for our experiments using our naive Bayes classifier. Next we present the results of our experiments and an analysis of these results.

CHAPTER 4:
Results

4.1 Introduction

In this section, we present the results of our experiments. First we will discuss the identification experiment over five authors. Then we will describe the results of growing the data set by doing the identification experiment over all 150 authors in the corpus. Next, we present the results of using six classes for identification, the five from the first experiment and an "other" class. Next, we show the results of using six classes but with the five author classes being the most prolific in the corpus. Then, we present the results of two experiments for identification over two classes, work e-mail and personal e-mails, where the first experiment consists of disproportionally-sized data sets, and the second consists of evenly-sized data sets. Finally, we present the results of experiments that try to identify the most prolific author from a sea of noise by training on work e-mails and testing on personal e-mails and vice versa.

4.2 Finding Five Author from Five

The premise behind all of our experiments was to use the most basic tools and do the least amount of refining to the data as possible to see how easy it is to identify an author in a noisy environment where the noise was other authors. The first experiment used five authors. We would only use text written by these five authors and our task was to identify each author correctly. The five authors that we chose were the first that came alphabetically in the corpus by last name. Table 4.1 shows the total number of e-mails written by the author in the corpus as well as the proportion of the corpus this represents, Maximum Likelihood Expectation (MLE), for each author in this experiment.

Author	E-mails	MLE
Allen-p	904	0.141
Arnold-j	1535	0.206
Bass-e	1614	0.343
Beck-s	1580	0.277
Brawner-s	202	0.032

Table 4.1: 5 Author E-mail Counts and MLE

We took all of the e-mails written only by those five authors and created the working data set for this experiment. We then created five even splits of the data and ran each set through the naive Bayes classifier. Table 4.2 shows that the MLE for each author remained the same for each split.

Author	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Allen-p	0.141	0.141	0.141	0.141	0.141
Arnold-j	0.206	0.206	0.206	0.206	0.206
Bass-e	0.343	0.343	0.343	0.343	0.343
Beck-s	0.277	0.277	0.277	0.277	0.277
Brawner-s	0.032	0.032	0.032	0.032	0.032

Table 4.2: 5 Author MLEs for five test splits

The results of feeding these five folds into the classifier are shown in Table 4.3.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.660	0.720	0.768	0.718	0.717

Table 4.3: 5 Author Accuracy

We see from the results that fold three gave us the best accuracy of the five folds. This is a totally random occurrence though, since the five splits were generated randomly. The take away from these results is that there is some authorship signal in the text that naive Bayes is able to learn since the accuracies that resulted are better than the baseline accuracy that would've resulted for the author with the highest MLE.

4.3 FFAPO

In this section, we will discuss the results from the first of two experiments involving six classes. We took the same five authors from the previous experiments and added all of the other e-mails in the corpus. Any e-mail that was not written by one of the original five authors was classified as "other" for this experiment, essentially creating noise from 145 unknown authors for the classifier. Table 4.4 shows the total e-mails for each class. Obviously, since "other" is our placeholder for the 145 other authors in the corpus, the count for this class is significantly larger than then other five classes.

Tables 4.5 through 4.9 show the mean results of the experiment for each individual author. The other category had the best numbers but that is expected since the number of 'other' e-mails was

Author	E-mails
Allen-p	904
Arnold-j	1535
Bass-e	1614
Beck-s	1580
Brawner-s	202
Other	90477

Table 4.4: E-mail Totals per Class

a significant portion of the data set. For this reason, we will ignore the results for the 'other' category and simply focus more closely on the results for each author individually. The three metrics of interest are precision, recall, and f-score. Beck had the highest mean precision in the experiment with .637 while the lowest precision was Brawner. Bass had the highest mean recall with .366 while Brawner had the lowest with .030. Similarly for f-score, Bass had the highest with .405 while Brawner had the lowest again with .037. At this point, an interesting phenomenon worth noting is the correlation between the three measures and the number of e-mails written by an author. The one counter to this is when looking at precision. As noted earlier, Beck had the highest precision but Bass actually wrote more e-mails than Beck.

	Mean	Standard Dev.
Precision	0.216	0.060
Recall	0.033	0.007
F-score	0.057	0.010

Table 4.5: Results for Allen

	Mean	Standard Dev.
Precision	0.462	0.026
Recall	0.208	0.019
F-score	0.286	0.018

Table 4.6: Results for Arnold

This correlation between number of e-mails written and our measures for identification was the impetus for the next experiment, PFAPO. As discussed in Chapter 3, in the PFAPO experiment, we chose our five authors to be the most prolific in the entire corpus for a couple of reasons. First, we hoped to see the same correlation between the number of e-mails written and ease of

identification. Second, we hoped to also see an actual increase in accuracy for each author due to the increased number of e-mails written.

	Mean	Standard Dev.
Precision	0.454	0.037
Recall	0.366	0.007
F-score	0.405	0.017

Table 4.7: Results for Bass

	Mean	Standard Dev.
Precision	0.638	0.036
Recall	0.249	0.012
F-score	0.358	0.011

Table 4.8: Results for Beck

	Mean	Standard Dev.
Precision	0.052	0.029
Recall	0.030	0.011
F-score	0.037	0.017

Table 4.9: Results for Brawner

4.4 PFAPO

In this section, we discuss the results of the second of the six class experiments, PFAPO. The tables 4.11 through 4.15 show these results. The first thing to note is the level of increase in quantity of e-mails written per author. (See Table 4.10)

Author	E-mails	MLE
Mann	4400	0.227
Dasovich	3930	0.203
Jones	3810	0.197
Shackleton	3774	0.195
Kaminski	3463	0.179

Table 4.10: 5 Proflific Author E-mail Counts and MLE

In FFAPO, the author with the least e-mails had 202 e-mails while the most had 1614 e-mails. In PFAPO, we see the average number of e-mails written was significantly greater. The author

with the least had 3,463 e-mails while the most had 4,440. The highest precision was Mann with a mean value of .828 while the lowest was Jones with .730. Recall was highest from Kaminski with .895 and lowest from Jones with .563. Finally, the highest f-score was .845 from Kaminski and the lowest was from Jones with .636.

This data showed us that there was definitely some relationship between accuracy and e-mail quantity. The metrics for each author were all significantly higher than the metrics in the FFAPO experiment. In comparing the f-scores in the two experiments, we saw that the lowest in FFAPO was .037 from Brawner while the lowest in PFAPO was .636 from Jones. Exploring this connection further, we saw that there was a significant disparity between the number of e-mails written by those same two authors. Brawner wrote 202 e-mails while Jones wrote 3,810. Jones wrote 18.861 times more than Brawner and Jones' f-score is 17.189 times higher. When we compare the highest f-scores in the two experiments, the highest in FFAPO was .405 from Bass while the highest in PFAPO was .845 from Kaminski. Bass wrote 1614 e-mails while Kaminski wrote 3463, which was 2.146 times more. The f-score for Kaminski was 2.086 times higher than Bass'.

We can also say that this data supported our assumption that a significant increase in e-mail quantity would result in a significant increase in accuracy. Earlier, we showed that accuracy increased eighteen fold between the two experiments' authors who had the least number of e-mails and accuracy increased two times between the authors who had the highest number of e-mails. In fact, each of the metrics for all authors in the PFAPO experiment were higher than their respective values for all of the authors in the FFAPO experiment. Figure 4.1 shows the mean precision for the authors in the two experiments. They are listed in the order they were classified so the first pair of bars is the first author tested in FFAPO against the first author tested in PFAPO. This graph is meant to show that precision went up across the board from FFAPO to PFAPO. Similarly, recall and f-score also increased, which is shown in Figures 4.2 and 4.3.

We can see that this data supports some of our hypotheses from above, but it also contradicts one. Earlier, we stated that writing more e-mails correlates to better results. For the most part, the data from the FFAPO experiment supported this hypothesis. The data from the PFAPO experiment however, does not. If we look at the most prolific author, Mann, we would assume

	Mean	Standard Dev.
Precision	0.828	0.014
Recall	0.782	0.009
F-score	0.804	0.009

Table 4.11: Results for Mann

	Mean	Standard Dev.
Precision	0.777	0.007
Recall	0.718	0.022
F-score	0.746	0.011

Table 4.12: Results for Daskovich

	Mean	Standard Dev.
Precision	0.730	0.023
Recall	0.563	0.012
F-score	0.636	0.014

Table 4.13: Results for Jones

	Mean	Standard Dev.
Precision	0.747	0.015
Recall	0.813	0.012
F-score	0.779	0.007

Table 4.14: Results for Shackleton

	Mean	Standard Dev.
Precision	0.801	0.021
Recall	0.895	0.011
F-score	0.845	0.010

Table 4.15: Results for Kaminski

from our hypothesis that she would have the highest scores associated with her. However, she only has the highest precision in this experiment. The highest recall and f-score both come from Kaminski, who happens to be the author who wrote the least number of e-mails in this experiment. This development leads us to believe that the phenomenon that occurred in the FFAPO experiment was just a coincidence of the input data.

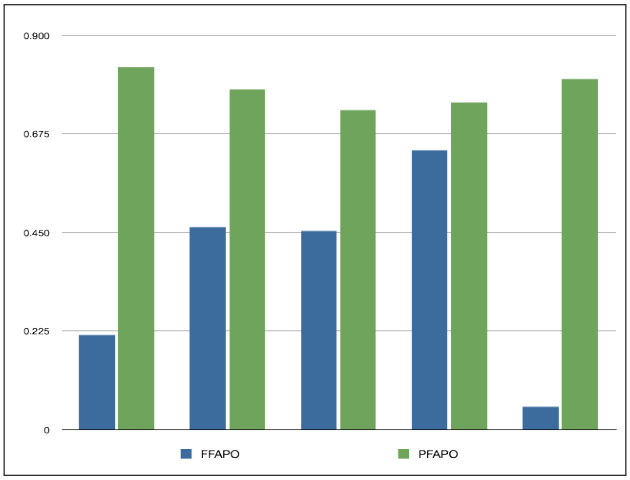

Figure 4.1: Mean: Precision

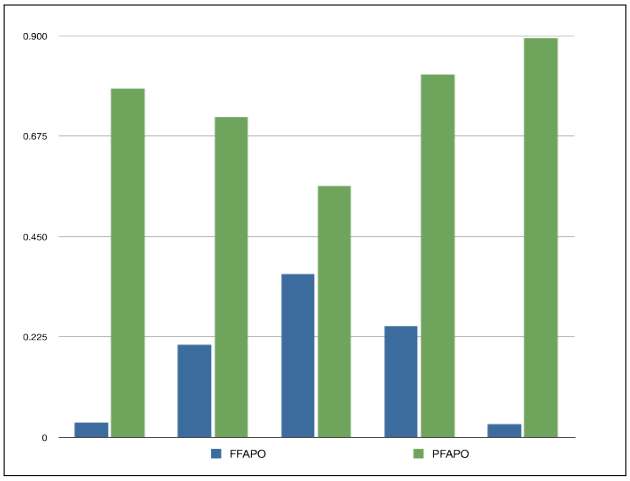

Figure 4.2: Mean: Recall

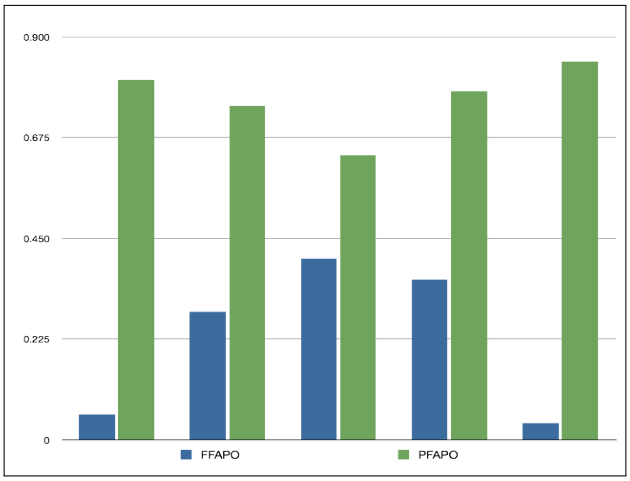

Figure 4.3: Mean: F-score

4.5 WvP

In this section, we will analyse the results of the WvP experiment. For this experiment, we took the hand-annotated e-mails described in Chapter 3 and ran them through the naive Bayes classifier. The impetus for this experiment was to see if there was enough signal to differentiate between work and personal e-mails.

Table 4.17 shows us the mean precision for work e-mails was .992 while the mean precision for personal e-mails from Table 4.18 was .683. The reason for the large gap between the two may have to do with the size of the data set and the prior. The number of work e-mails in the corpus (1411) was so much larger than the number of personal e-mails (292), that the prior could be dominating the classifier, causing it to choose work more often than it should.

The mean recall value for work from Table 4.17 is .911 while Table 4.18 shows the value for personal is .962. The first thing that jumps out is the difference between the precision and recall values for personal e-mails. Recall is defined as the fraction of e-mails that are actually personal that are correctly labelled as personal. These two numbers show us that the classifier had trouble precisely calling an e-mail personal, but it found most of the personal e-mails in the data set. In other words, the classifier is using the personal label too "loosely". The low precision tells us that the total number of e-mails that the classifier calls personal is actually more than the total number of personal e-mails in the data set, causing the precision to drop to .683.

Table 4.17 shows the mean f-score for work e-mails and Table 4.18 shows us the mean for the personal e-mails. We see that the f-score for personal e-mails is lower than work e-mails. This is expected, since f-score is a function of both precision and recall.

4.6 WvP: Randomized Under-sampling

For our next experiment, we use a technique known as Randomized Under-Sampling to essentially remove the prior and even out the data set.

Table 4.18 shows that the precision for personal e-mails increased over the previous experiment to .904 while the precision for work dropped to .968. This may be a direct result of adjusting the prior since there seems to be even more signal associated with the personal e-mails.

The values for recall did not change very much for the personal e-mails We see from Table 4.18 that recall went up slightly by .009. The recall for work e-mails went down to .895. A possible explanation for this revolves around the reduction of the prior. Since the two data sets were the same size for this experiment, there was less domination of the prior than in the previous experiment where there were more work e-mails in the data set. There seems to be more confusion between the work and personal e-mails. More of the work e-mails are being incorrectly called personal by the classifier, which caused the drop in recall.

Since f-score is a function of the same values that drive both precision and recall, the changes that occur in both f-scores are expected. The f-score for work dropped to .930. We expected to see this since both precision and recall for work both decreased as well. The f-score for personal increased to .903, which was also expected. As noted above, the recall for personal had a very small increase. Precision, however, had a large increase, which explains the resulting increase in f-score.

4.7 TWTPVV

For this pair of experiments, we use the hand annotated data set of work and personal e-mails described in Chapter 3 to identify the most prolific author from a sea of noise. For our training set, we used a mix of the work e-mails written by Mann and an equal number of random e-mails written by other authors in the corpus. Our test set was a mix of personal e-mails written by Mann and an equal number of random e-mails written by other authors in the corpus. The objective of this experiment was to simulate having only one class of e-mails from an author, work or personal, and using that class to create a model that would still be able to pick out the author of interest from a large set of other authors.

Table 4.16 shows us the mean precision for training on work and personal. Training on work resulted in a precision of .896 while training on personal resulted in a precision of .826. These are both very interesting numbers in that they show that there is enough signal in each of the two classes to pick out the other in a sea of noise relatively well. These values are not that much lower than the values from the previous experiments, which means the noise that was introduced had an effect on the signal, but one that was effectively managed by the classifier.

The mean recall for work and personal from Table 4.16 are .820 and .589 respectively. The large difference between these two values may be due to the amount of signal in the e-mail categories. The personal e-mails may have less of the author's signal and therefore result in lower chance

of identifying the the work e-mails. This seems to contradict the results from the WvPRUS experiment, but the high recall there was from looking for only one of the two classes at a time. The disparity between the two values in TWTPVV show that there is a more identifying signal in the work e-mails than the personal e-mails.

Table 4.16 also includes values for f-score. The f-score for work e-mails was .856 while the f-score for personal was .688. As discussed in Chapter 2, since f-score is a function of both precision and accuracy, these values are expected.

	Precision	Recall	F-score
Work/Noise	0.896	0.820	0.856
Personal/Noise	0.826	0.589	0.688

Table 4.16: Work/Noise, Personal/Noise Mean Values

4.8 Conclusion

In this chapter, we analysed the results of our experiments. With the data that we collected and the knowledge that we have learned from it, we are now able to make conclusions and recommendations for future work to continue making strides in this domain.

	Precision	Recall	F-score
Normal	0.992	0.911	0.950
RUS	0.968	0.895	0.930

Table 4.17: Work Mean Values

	Precision	Recall	F-score
Normal	0.693	0.962	0.805
RUS	0.903	0.969	0.935

Table 4.18: Personal Mean Values

	+	-
+	2572	561
-	22	250

Accuracy 0.829

Table 4.19: Confusion Matrix for Work

	+	-
+	522	18
-	61	565

Accuracy 0.932

Table 4.20: Confusion Matrix for Work RUS

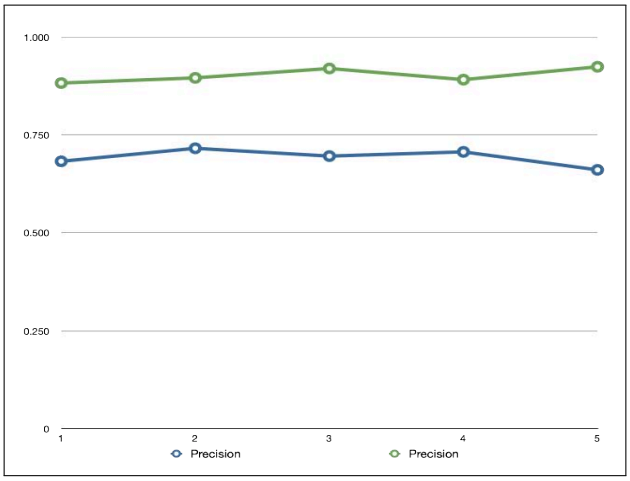

Figure 4.4: Personal: Precision Normal vs. RUS

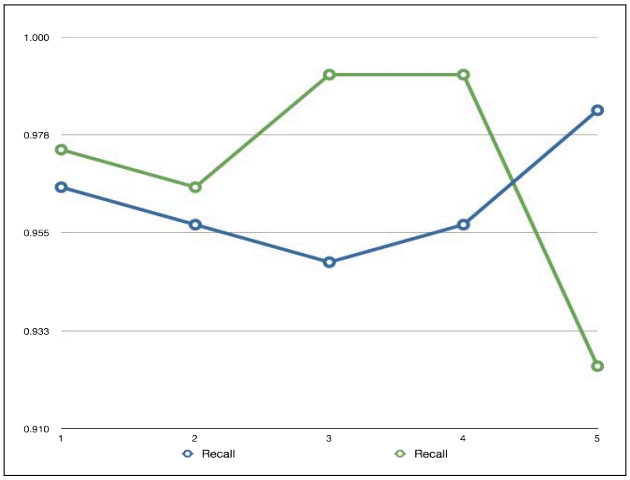

Figure 4.5: Personal: Recall Normal vs. RUS

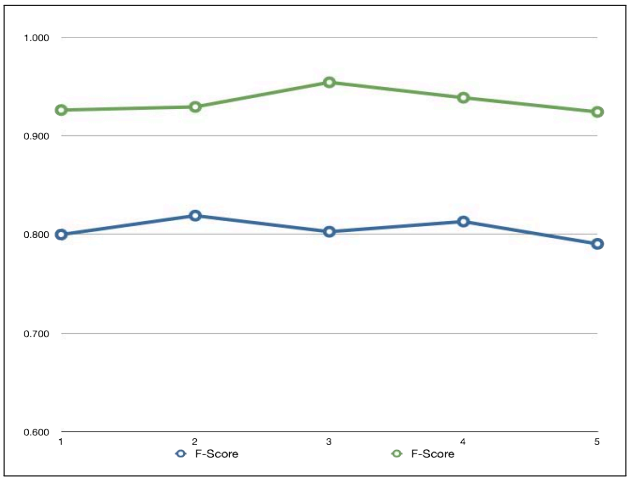

Figure 4.6: Personal: F-score Normal vs. RUS

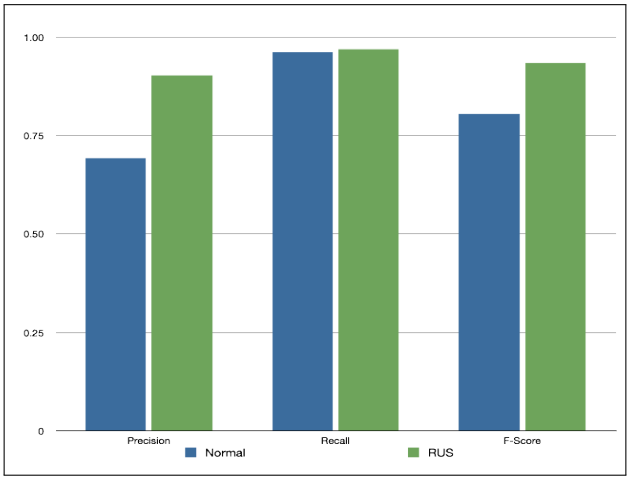

Figure 4.7: Personal: Means Normal vs. RUS

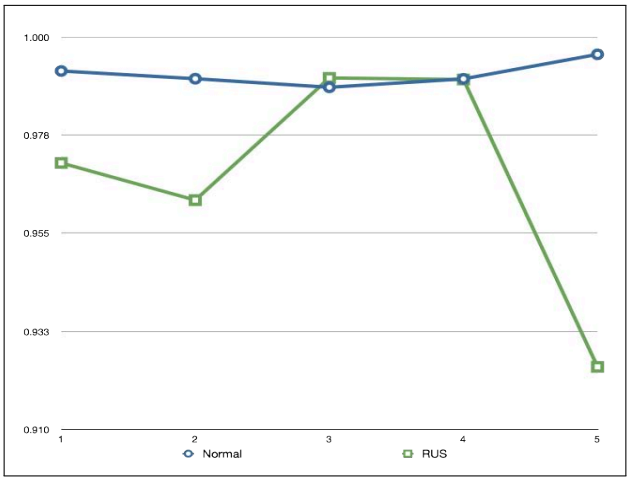

Figure 4.8: Work: Precision Normal vs. RUS

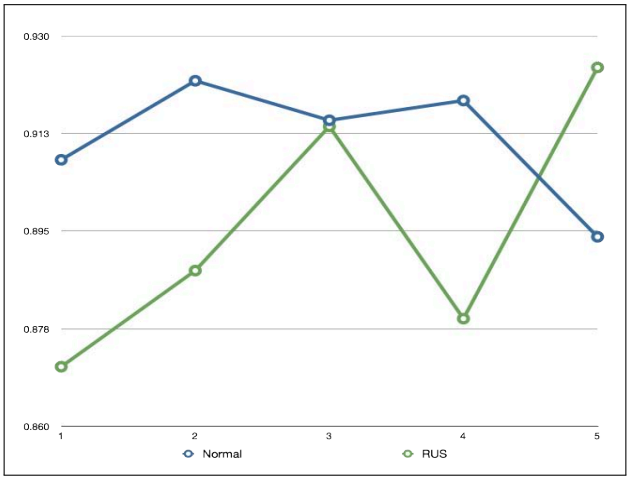

Figure 4.9: Work: Recall Normal vs. RUS

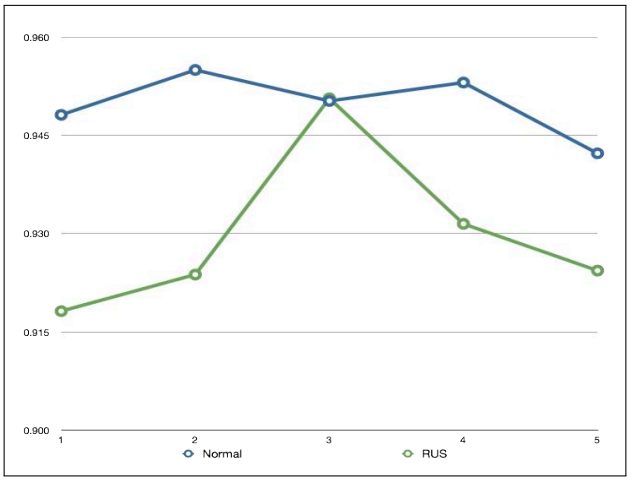

Figure 4.10: Work: F-score Normal vs. RUS

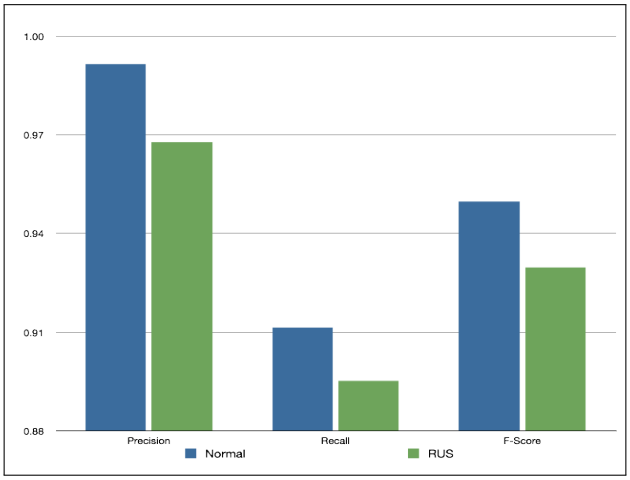

Figure 4.11: Work: Means Normal vs. RUS

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
Conclusion and Future Work

5.1 Summary

In this thesis, we attempted to answer two questions. The first question we addressed was "How feasible is it to do authorship attribution in the e-mail domain using the most simple methods when we have a sufficiently large corpus of data per author?" The second was "How feasible is it to differentiate between personal e-mails and work-related e-mails given a small data set and simple classification methods?" The answer to both of these questions is "very feasible."

We started in Chapter 2 where we presented the important concepts and prior work related to our experiments. We described naive Bayes as our primary classification method for all of our experiments. Finally, we discussed smoothing techniques that we used in order to combat the zero and low count problem.

After laying out the prior and related work and concepts, we moved to Chapter 3 where we present our experimental design. We started with a description of the data that we used and the process we used to go from raw data to data ready for experimentation. Then we discussed the method of classification. Next, we described the process used to create the "work" and "personal" data sets that necessary for the second set of experiments. This was an important step, since there was no data set using the Enron corpus that was split in this way. Finally, we gave a brief description of each of the experiments that we ran.

In Chapter 4, we analysed the results of our experiments. We started with our First Five Authors experiment, which showed us that there was a strong enough authorship signal in these e-mails to use in order to do better than MLE authorship attribution. First Five Authors was followed up with the results from the First Five Authors Plus Other experiment where we showed that adding a new class of "noise" represented by e-mails from 145 other unseen authors, make it significantly more difficult to identify the authors we were looking for. Next we presented the results of the Prolific Five Authors Plus Other experiment, where the data revealed that having a larger amount of representative data for the target authors significantly helped fight the noise we introduced and increased the classifier's accuracy.

For our second set of experiments, we began with the results for our Work Versus Personal experiment. These results told us that we were able to delineate between a work related e-mail and a personal e-mail sufficiently well. Since there were questions about the prior dominating those results, we ran a follow-on experiment called Work Versus Personal Randomized Under-Sampling. These results confirmed that we were able to identify a work e-mail or a personal e-mail, and, as a result of reducing the prior, our accuracy at doing so increased.

Our final set of experiments were an attempt to use one class of e-mails to identify one author of interest in a different class of e-mails in a sea of noise and vice versa. This set of experiments was meant to simulate only having access to, for example, work e-mails from an author to build a model of and then using that model to identify the same author via their personal e-mails out of a sea of other e-mails from unknown authors. Our results show by having only one class of e-mail from which to create a model, it is feasible to have some success at identifying an author from the class not used to create that model.

Our results are very promising in that we got some high f-scores and accuracies. This does not mean that research in this area is complete, since this was a very directed study where only certain techniques were used. We also created a hand-annotated data set that was relatively small due to the time cost of hand-annotation. The next section will describe future work that would be needed to make further progress in this domain.

5.2 Future Work

5.2.1 Expanding Feature Selection

Our research focused on only one feature of email, unigrams. It would be interesting to see what results would occur when we use different features such as bigrams, orthogonal sparse bigrams, e-mail length, etc.

5.2.2 Increasing Corpus Size

For our second set of experiments, we focused on trying to differentiate between work e-mails and personal e-mails. The problem that we faced was that there was no corpus of e-mail that was already split up in this manner. This means we had to create our own data set by annotating before experimentation. Future research endeavours should include expanding the annotation efforts on the the Enron e-mail corpus to make a work and personal data set for more authors in the corpus, eventually getting to the point where each author has a work and personal set. Further experimentation could explore how well we do when trying to identify multiple authors

with two writing styles. That is, how well do we do at correctly differentiating author one's work e-mail and author one's personal e-mail from author two's?

5.2.3 Another Authorship Attribution Method

When we created our work and personal e-mail data sets, only used one author and annotated about 33% of her e-mail set. Another method for authorship attribution to try could be using the other 67% as test data to a classifier to see how well it annotates the rest of the data.

5.3 Closing Remarks

Authorship attribution is a problem that has been attacked from many different angles. Methods that we know work continue to be implemented on new data sets such as e-mails, SMS, and twitter feeds. With an increased focus in this age of technology on things such as cyber bullying and chat room predators, there are many scenarios where authorship attribution in the e-mail domain could be very useful. The first is in relation to cyber bullying. Children are under an increased threat of cyber bullying via e-mail, SMS, and social media outlets such as Facebook and Twitter. Cyber bullies can create an anonymous bully account and proceed to terrorize they targets with little to no fear of every being discovered. Perfecting methods for authorship attribution in the e-mail domain is the just the beginning to finding a way to fight this activity. Also, a person with multiple e-mail accounts could be involved in any number of activities such as preying on young children to plotting terrorist activities. If authorities had access to one of the suspects e-mail accounts, such as a work account, they could use similar authorship attribution methods to build an author model of suspects and narrow the amount of data that needs to be searched. Given these scenarios, it is clear that authorship attribution in the e-mail domain should continue to be an area of focus.

THIS PAGE INTENTIONALLY LEFT BLANK

REFERENCES

[1] Grant Gehrke. Authorship discovery in blogs using bayesian classification with corrective scaling. Master's thesis, Naval Postgraduate School, 2008.

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

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

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

[5] Wilhelm Fucks. On mathmatical analysis of style. *Biometrika*, 39(1/2):122–129, 1952.

[6] RS Forsyth, DS Holmes, and EK Tse. Cicero, sigonio, and burrows: Investigating the authenticity of the consolatio. *Lit Linguist Computing*, 14(3):375–400, 1999.

[7] D. I. Holmes. The analysis of literary style - a review. *Journal of the Royal Statistical Society*, 148(4):328–341, 1985.

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

[9] Alvar Ellegard. A statistical method for determining authorship: The junius letters 1769-1772. *Gothenburg Studies in English*, 13, 1962.

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

[11] D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. London: Prentice Hall, 2009.

[12] S.F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Harvard University Tech. Rep.*, 1998.

[13] I.H. Witten and T.C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.

[14] B. Dorr and C. Monz. Cmsc 723: Introduction to computational linguistics, lecture 5. http://www.unimacs.umd.edu/~christof/courses/cmsc723-fall04/lecture-notes/Lecture5-smoothing-6up.pdf.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A:
Tables

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.167	0.171	0.200	0.231	0.313	0.216	0.060
Recall	0.028	0.033	0.044	0.033	0.027	0.033	0.007
F-Score	0.048	0.056	0.073	0.058	0.050	0.057	0.010

Table A.1: Allen Complete Results

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.454	0.492	0.467	0.475	0.424	0.462	0.026
Recall	0.192	0.212	0.231	0.186	0.218	0.208	0.019
F-Score	0.270	0.296	0.309	0.267	0.288	0.286	0.018

Table A.2: Arnold Complete Results

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.478	0.408	0.496	0.462	0.424	0.454	0.037
Recall	0.379	0.363	0.363	0.363	0.362	0.366	0.007
F-Score	0.423	0.384	0.419	0.407	0.391	0.405	0.017

Table A.3: Bass Complete Results

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.638	0.655	0.579	0.642	0.675	0.638	0.036
Recall	0.234	0.247	0.266	0.244	0.256	0.249	0.012
F-Score	0.343	0.359	0.364	0.353	0.372	0.358	0.011

Table A.4: Beck Complete Results

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.029	0.053	0.031	0.100	0.050	0.052	0.029
Recall	0.025	0.020	0.025	0.050	0.024	0.030	0.011
F-Score	0.027	0.034	0.028	0.067	0.032	0.037	0.017

Table A.5: Brawner Complete Results

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.952	0.952	0.953	0.952	0.952	0.952	0.0004
Recall	0.984	0.982	0.982	0.985	0.983	0.983	0.001
F-Score	0.968	0.967	0.967	0.968	0.967	0.968	0.0005

Table A.6: Other Complete Results

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.831	0.807	0.845	0.833	0.823	0.828	0.014
Recall	0.792	0.780	0.777	0.790	0.770	0.782	0.009
F-Score	0.811	0.793	0.809	0.811	0.796	0.804	0.009

Table A.7: Six class results for Mann

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.789	0.775	0.772	0.776	0.772	0.777	0.007
Recall	0.700	0.753	0.723	0.710	0.704	0.718	0.022
F-Score	0.742	0.764	0.747	0.742	0.736	0.746	0.011

Table A.8: Six class results for Dasovich

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.705	0.712	0.755	0.725	0.751	0.730	0.023
Recall	0.561	0.547	0.578	0.572	0.559	0.563	0.012
F-Score	0.625	0.619	0.655	0.639	0.641	0.636	0.014

Table A.9: Six class results for Jones

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.765	0.758	0.751	0.731	0.731	0.747	0.015
Recall	0.792	0.821	0.816	0.817	0.821	0.813	0.012
F-Score	0.778	0.788	0.782	0.772	0.774	0.779	0.007

Table A.10: Six class results for Shackleton

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.826	0.780	0.820	0.786	0.791	0.801	0.021
Recall	0.888	0.914	0.894	0.887	0.891	0.895	0.011
F-Score	0.856	0.842	0.855	0.834	0.838	0.845	0.010

Table A.11: Six class results for Kaminski

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.927	0.933	0.930	0.931	0.929	0.930	0.002
Recall	0.941	0.933	0.941	0.936	0.937	0.938	0.003
F-Score	0.934	0.933	0.936	0.933	0.933	0.934	0.001

Table A.12: Prolific Six class results for Other

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.683	0.716	0.696	0.707	0.661	0.693	0.022
Recall	0.966	0.957	0.948	0.957	0.983	0.962	0.013
F-Score	0.800	0.819	0.803	0.813	0.791	0.805	0.011

Table A.13: Results for Personal

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.992	0.990	0.989	0.990	0.996	0.992	0.003
Recall	0.908	0.922	0.915	0.918	0.894	0.911	0.011
F-Score	0.948	0.955	0.950	0.953	0.942	0.950	0.005

Table A.14: Results for Work

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.883	0.896	0.920	0.891	0.924	0.903	0.018
Recall	0.974	0.966	0.991	0.991	0.924	0.969	0.028
F-Score	0.926	0.929	0.954	0.939	0.924	0.935	0.012

Table A.15: Results for Personal-RUS

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Mean	Standard Dev.
Precision	0.971	0.963	0.991	0.990	0.924	0.968	0.027
Recall	0.871	0.888	0.914	0.879	0.924	0.895	0.023
F-Score	0.918	0.924	0.951	0.932	0.924	0.930	0.013

Table A.16: Results for Work-RUS

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B:
Graphs

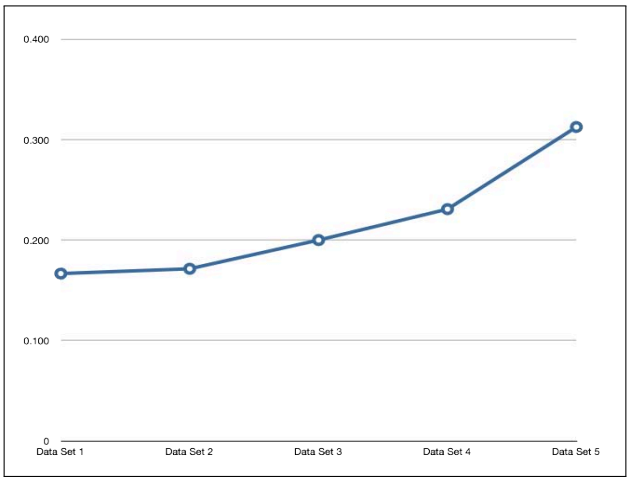

Figure B.1: Allen: Precision

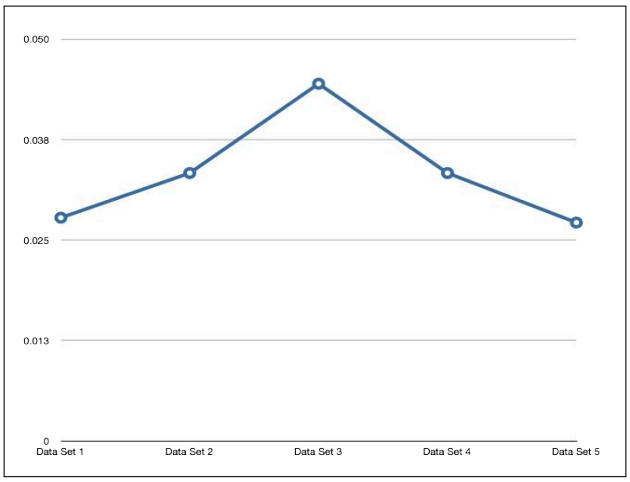

Figure B.2: Allen: Recall

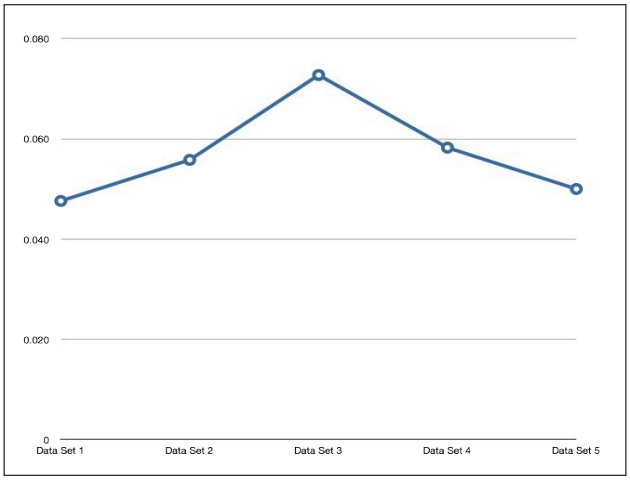

Figure B.3: Allen: F-score

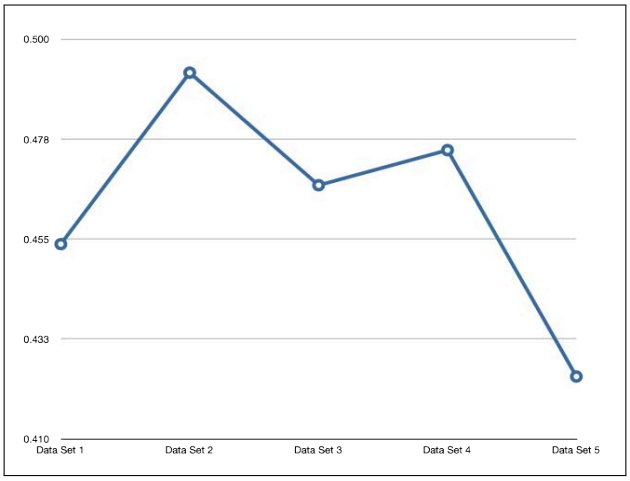

Figure B.4: Arnold: Precision

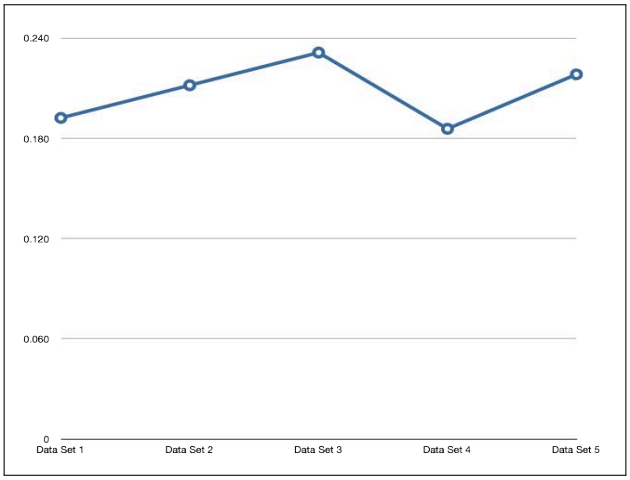

Figure B.5: Arnold: Recall

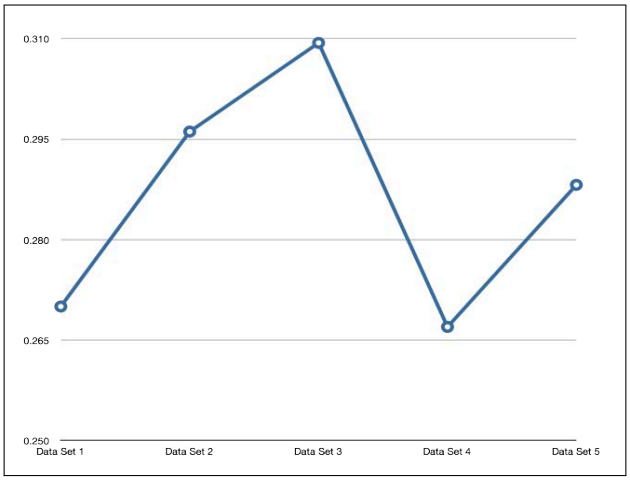

Figure B.6: Arnold: F-score

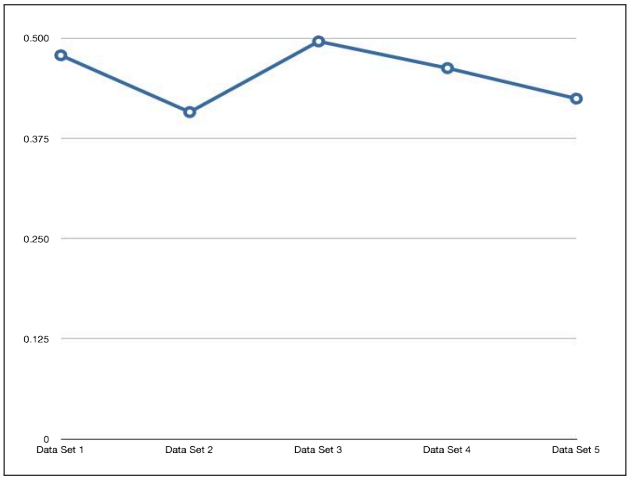

Figure B.7: Bass: Precision

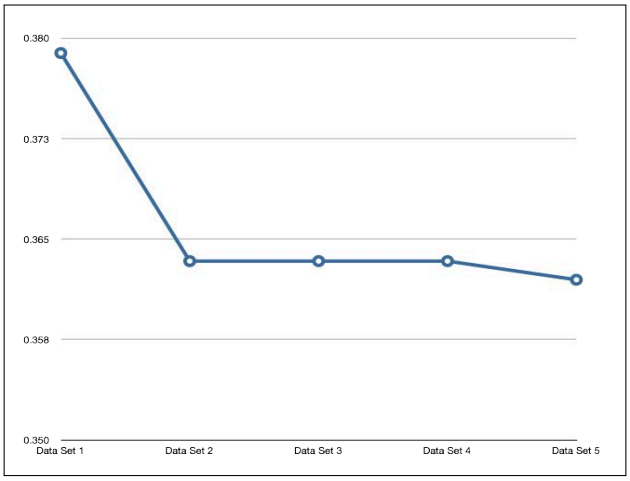

Figure B.8: Bass: Recall

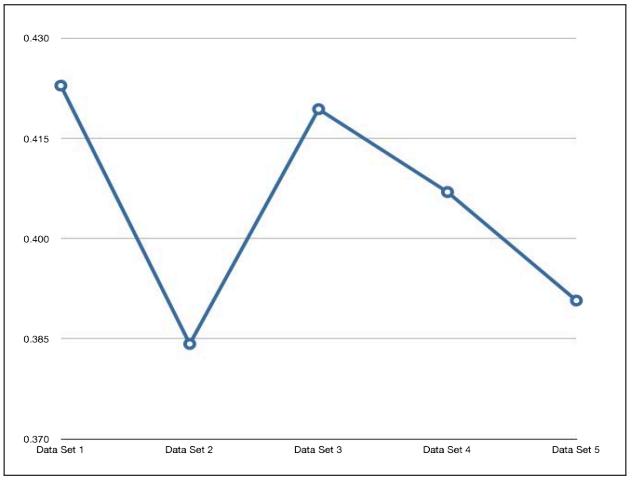

Figure B.9: Bass: F-score

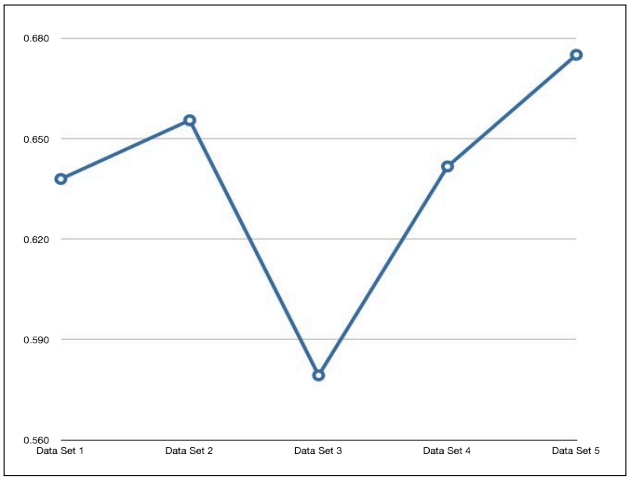

Figure B.10: Beck: Precision

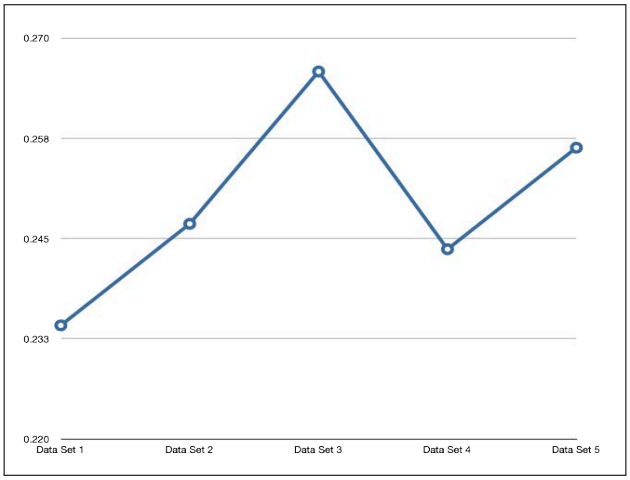

Figure B.11: Beck: Recall

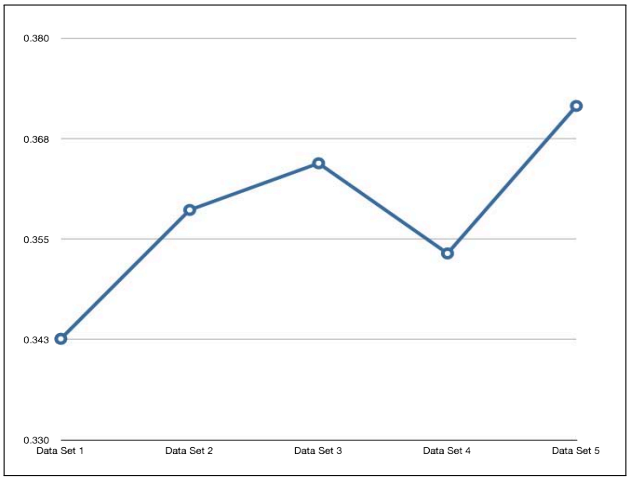

Figure B.12: Beck: F-score

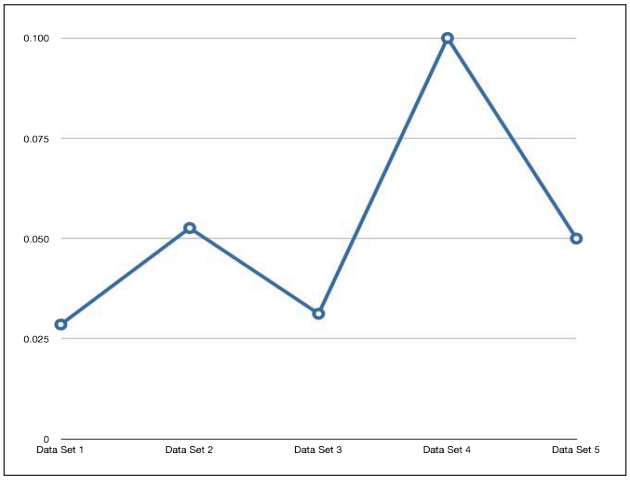

Figure B.13: Brawner: Precision

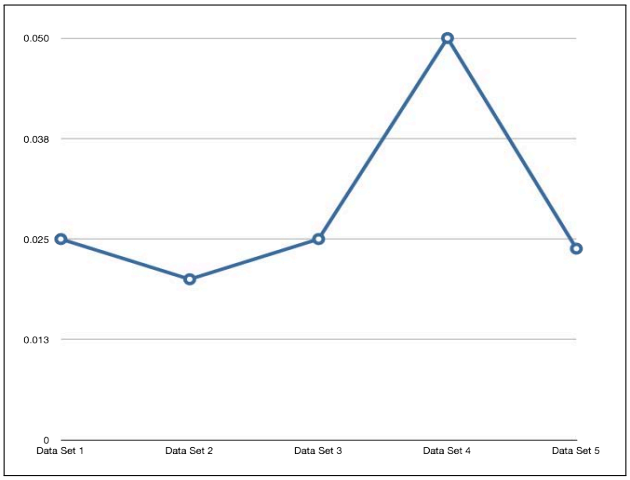

Figure B.14: Brawner: Recall

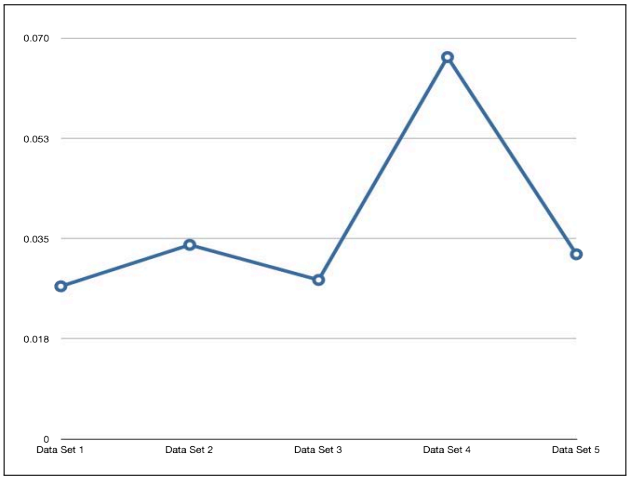

Figure B.15: Brawner: F-score

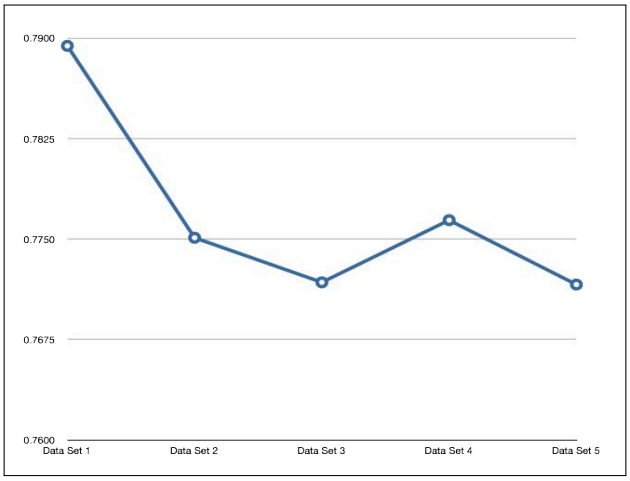

Figure B.16: Daskovich: Precision

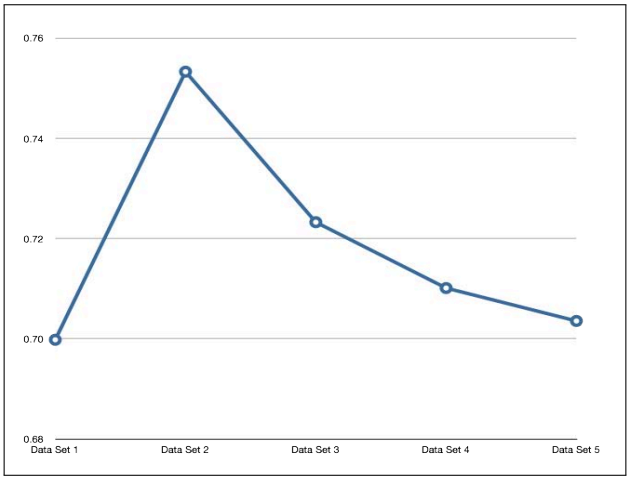

Figure B.17: Daskovich: Recall

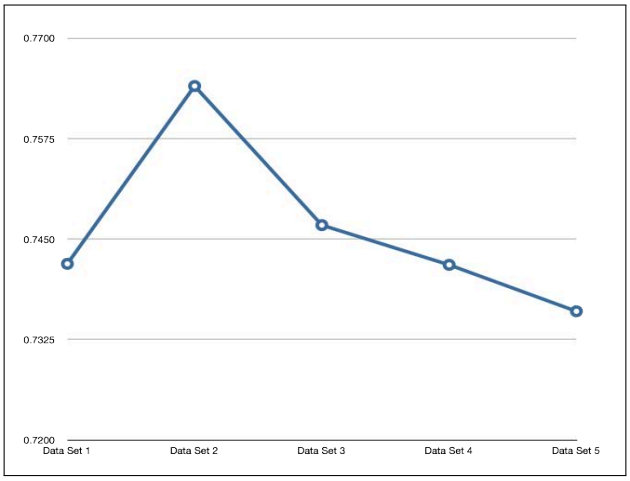

Figure B.18: Daskovich: F-score

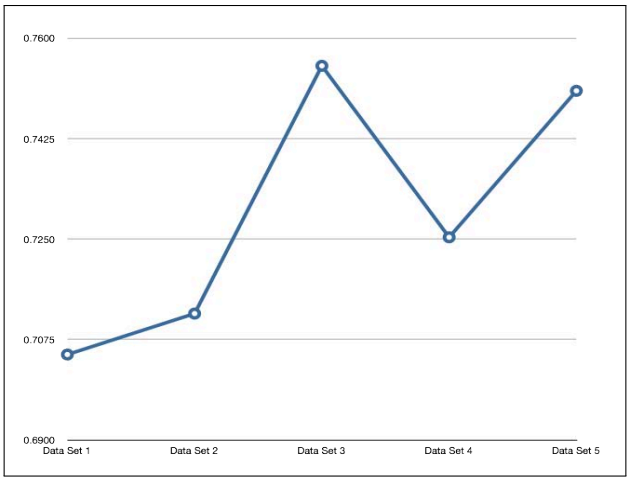

Figure B.19: Jones: Precision

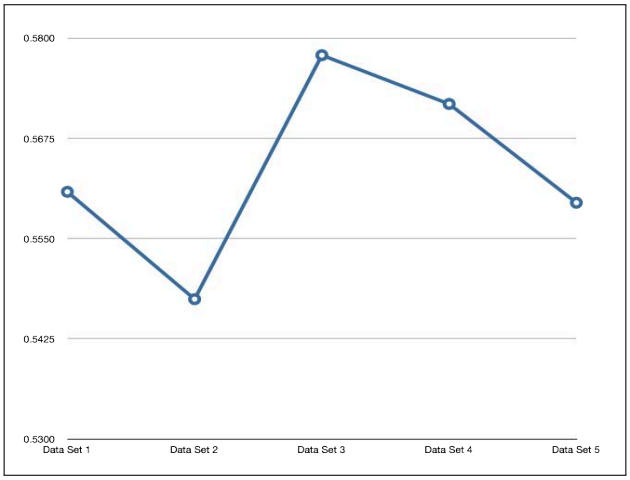

Figure B.20: Jones: Recall

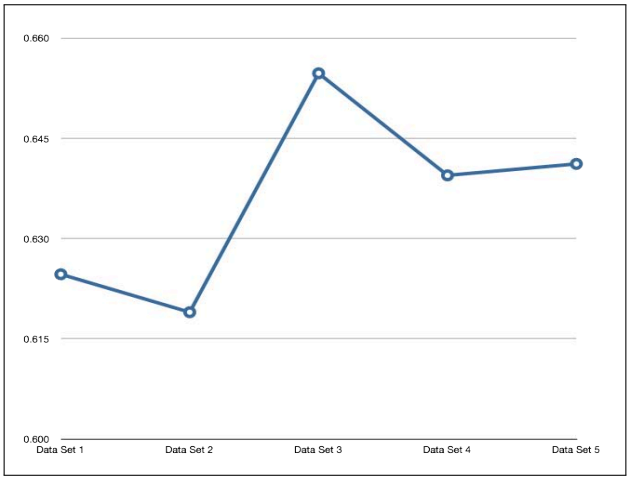

Figure B.21: Jones: F-score

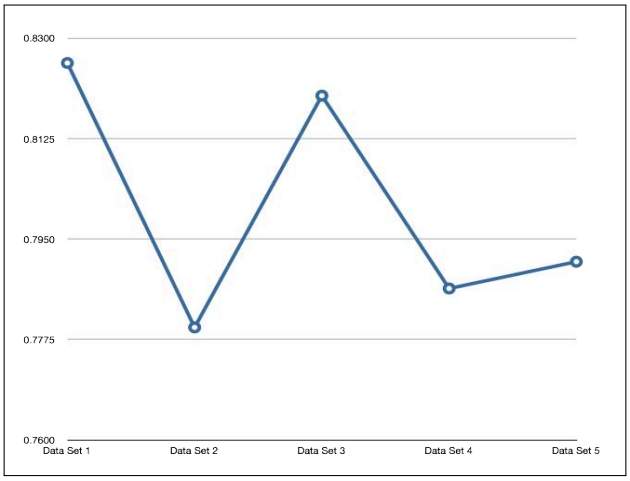

Figure B.22: Kaminski: Precision

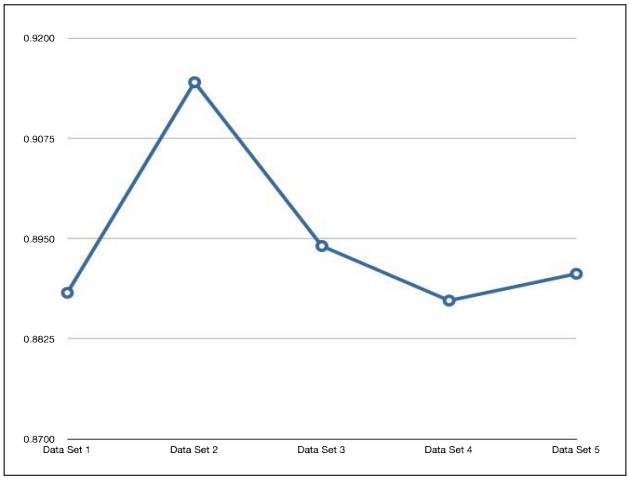

Figure B.23: Kaminski: Recall

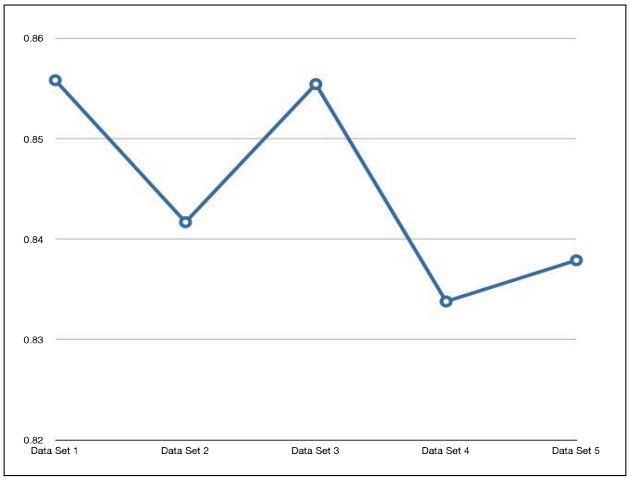

Figure B.24: Kaminski: F-score

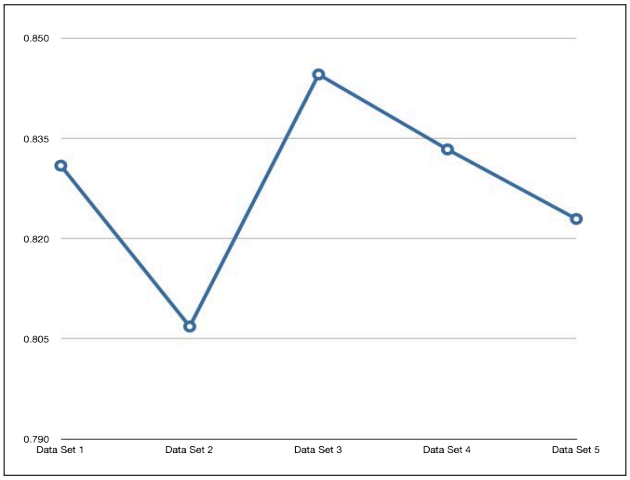

Figure B.25: Mann: Precision

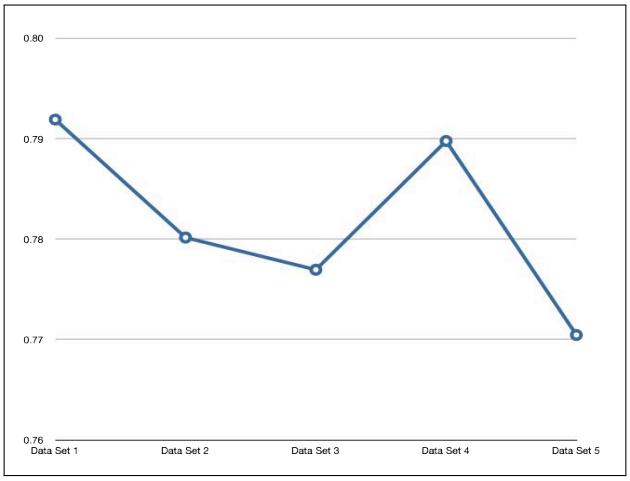

Figure B.26: Mann: Recall

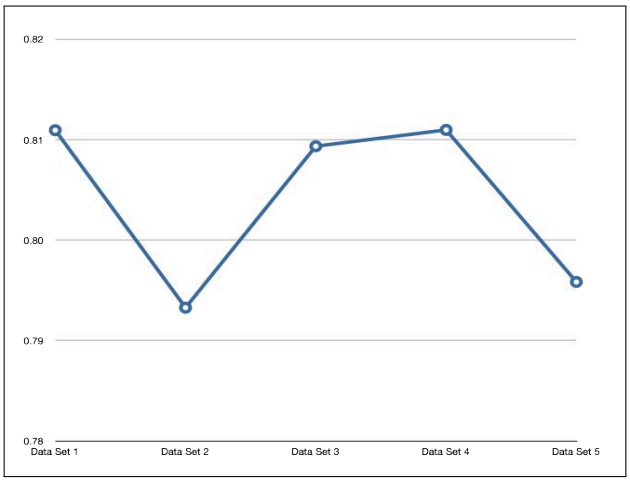

Figure B.27: Mann: F-score

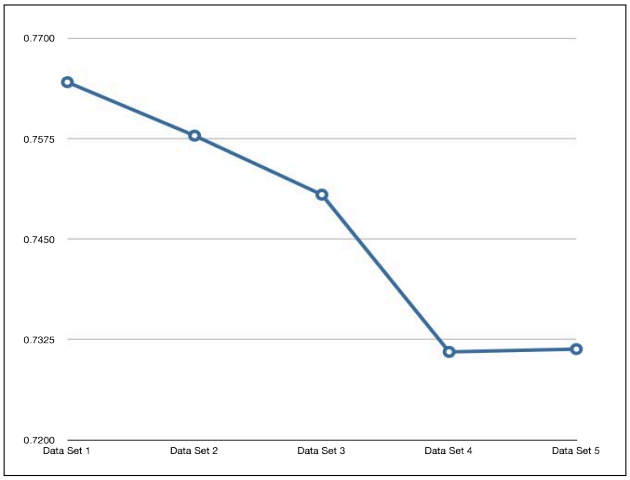

Figure B.28: Shack: Precision

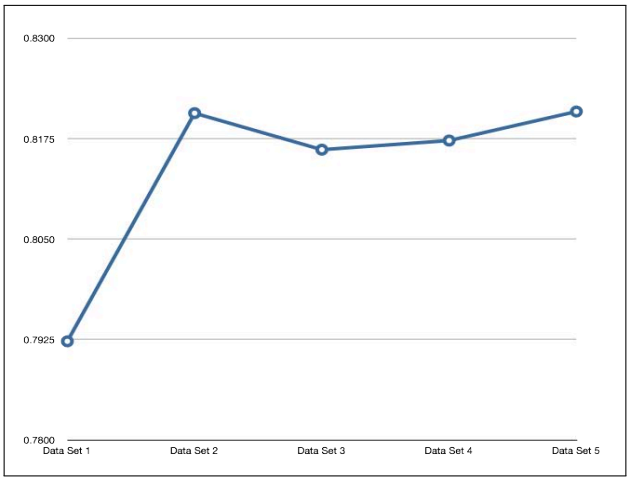

Figure B.29: Shack: Recall

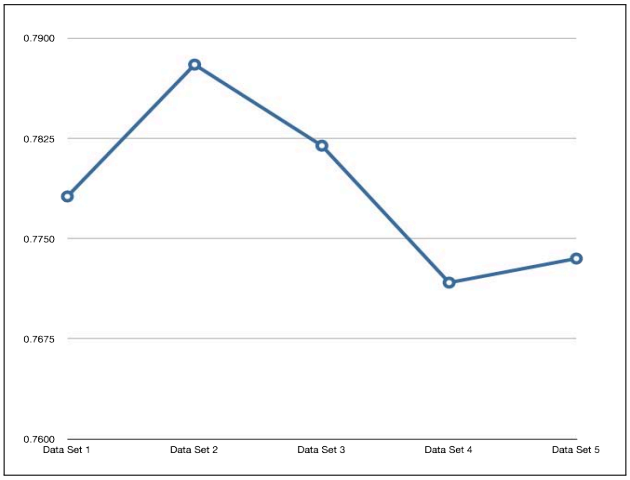

Figure B.30: Shack: F-score

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Navy Representative
   Naval Postgraduate School
   Monterey, California