2009-12

Novel topic impact on authorship attribution

Caver, Johnnie F.

Monterey, California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

NOVEL TOPIC IMPACT ON AUTHORSHIP ATTRIBUTION

by

Johnnie F. Caver

December 2009

Thesis Co-Advisors: Andrew I. Schein
Craig H. Martell

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** December 2009	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Novel Topic Impact on Authorship Attribution		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Johnnie F. Caver			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. 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.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

Several authorship attribution studies have speculated about the existence of a link between topic cues and author style features. This research presents a novel experimental protocol for measuring the impact of topic features on author attribution predictive models. We call our technique "novel topic cross-validation," which consists of holding out a single topic in a test set and iterating over choices of held-out topic to compute an average performance score.

Using the New York Times Annotated corpus, we perform a subset procedure to build a sub-corpus of 18,862 documents, 15 authors, and 23 topics. With this sub-corpus, we perform a novel topic cross-validation. Our experiments differ from previous attempts to model topic/author influence in scope; previous methods were limited to three or fewer topics or authors. Having a larger set of topics and authors should provide researchers with a greater opportunity to explore the variability of style cues represented in sets of authors, as well as the confounding influence of topic. For this reason, we supply document/author/topic identifications so that researchers can build upon our work in a reproducible fashion.

14. SUBJECT TERMS Authorship Detection, Topic Detection, Author-Topic Correlation, Topic-Author Correlation, Maximum Entropy, New York Times Annotated Corpus			**15. NUMBER OF PAGES** 81
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

NOVEL TOPIC IMPACT ON AUTHORSHIP ATTRIBUTION

Johnnie F. Caver
Lieutenant, United States Navy
B.S., Hampton University

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
December 2009

Author: Johnnie F. Caver

Approved by: Andrew I. Schein
Thesis Co-Adviser

Craig H. Martell
Thesis Co-Advisor

Peter J. Denning
Chairman, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Several authorship attribution studies have speculated about the existence of a link between topic cues and author style features. This research presents a novel experimental protocol for measuring the impact of topic features on author attribution predictive models. We call our technique “novel topic cross-validation,” which consists of holding out a single topic in a test set and iterating over choices of held-out topic to compute an average performance score.

Using the New York Times Annotated corpus, we perform a subset procedure to build a sub-corpus of 18,862 documents, 15 authors, and 23 topics. With this sub-corpus, we perform a novel topic cross-validation. Our experiments differ from previous attempts to model topic/author influence in scope; previous methods were limited to three or fewer topics or authors. Having a larger set of topics and authors should provide researchers with a greater opportunity to explore the variability of style cues represented in sets of authors, as well as the confounding influence of topic. For this reason, we supply document/author/topic identifications so that researchers can build upon our work in a reproducible fashion.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I have many people to thank for their guidance and support while completing this thesis process. In particular, I would like to thank:

My thesis co-advisor Professor Craig Martell, for knowing when to lift a heavy hand and when to apply the lightest touch while guiding me through this thesis process. Your passion for teaching and the motivation you showed every day in all that you do was an inspiration and truly made it a joy to learn.

My thesis co-advisor Andrew Schein for all of the support you provided in every aspect of this process. I especially thank you for taking the time to answer all of my questions and carefully explain (and sometimes re-explain) anything that I did not understand. I have learned a great deal from you and I wish could remain here longer in order to have the opportunity to work with you more.

My lab partners and classmates Brian, Dave, Jenny, John, and Constantine for all your help during the course of this curriculum. It has been a privilege serving with each of you and I pray that you all have continued success throughout the remainder of your military careers.

My Mother Betty who retired early from her job, set aside her fear of flying, said goodbye to all of her family and friends, and left the only home she has ever known in South Carolina to come to California and help support my husband, my children, and me during this process. Mom, I simply cannot thank you enough for all of your love and support. I could not have done this without you.

My husband Leonard Sr. who is my greatest supporter and an unlimited source of strength and encouragement, and my children Leonard Jr., Joshua, and Maya for reminding me every day of the things that are most important in life. None of my accomplishments would mean anything without you. You truly represent the best part of every day.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. MOTIVATION

The analytical focus of authorship attribution is on identifying the author of an anonymous text given undisputed knowledge of various communications written by that particular author. Several authorship attribution studies have speculated about the existence of a link between topic cues and author style features. This research presents a novel experimental protocol for measuring the impact of topic features on author attribution predictive models. We call our technique "novel topic cross-validation," which consists of isolating a single topic in a test set, generating training models from the remaining topics, then iterating over choices of held-out topic to compute an average performance score. The underlying idea is to measure the degree of impact topic cues have on the ability of the classifier to predict the author of a particular communication.

Using the New York Times Annotated corpus, we generated two sub-corpuses of data: one consisting of 3,000 documents cross-correlated with 2 authors and 4 topics, and the other consisting of 18,862 documents cross-correlated with 15 authors and 23 topics. From these sub-corpuses, we perform a novel topic cross-validation. Our experiments differ from previous attempts to model topic/author influence in scope, balance, and classification; previous methods were limited to three or fewer topics or authors, using equally balanced data sets and binary pair-wise classifications. Having a larger set of topics and authors and using a multi-classification of un-balanced data should provide researchers with a greater opportunity to explore the variability of style cues represented in sets of authors, as well as the confounding influence of topic. For this reason, we supply document/author/topic identifications so that researchers can build upon our work in a reproducible fashion.

B. ORGANIZATION OF THESIS

This thesis is organized as follows:

- Chapter I discusses the motivation for determining the impact of topic variability on authorship attribution problems.
- Chapter II contains background information on the history of authorship attribution as well as detailed discussions of methods used for modeling of lexical writing-style features with particular focus on word frequencies of the unigram model. In addition, this chapter contains discussions of five previous studies conducted in the realm of author-topic cross correlation.
- Chapter III explains the design and methodology used in the experiments to include information regarding the source of the data, how the data subsets were selected and prepared, and two scenarios used for experimentation. Finally, a description is provided for the classification package and performance measures used in the evaluation process.
- Chapter IV presents results of the experiments for both scenarios and provides detailed analysis for each data set.
- Chapter V contains a summary of the research conducted, along with conclusions and recommendations for future work.

II. BACKGROUND

A. AUTHORSHIP ATTRIBUTION

The authorship attribution problem is a subset of a broader field of linguistic study known as authorship analysis. Authorship analysis is concerned with the identification, exploitation, and characterization of textual features in written communications. The analytical focus in authorship attribution is on identifying the author of an anonymous text given undisputed knowledge of various communications written by that particular author. This authorial "fingerprint" is derived from statistical analysis of various writing-style features within a document to include lexical, character, syntactic, semantic, and application-specific features. A superset of the authorship attribution problem is authorship characterization [1], which is the attempt to infer certain characteristics about an author such as age, gender, language or educational background [2]. Since our goal is to determine the impact of topic cues on authorship attribution, we limit the scope of our research to statistical analysis of lexical writing-style features, particularly the analysis of word frequencies in the unigram model.

B. HISTORY OF AUTHORSHIP ATTRIBUTION

The "Father" of authorship attribution is widely accepted to be 18th century English logician Augustus de Morgan who suggested that the author of a text might be determined by examining the length of words in a document [3]. In 1887, T. C. Mendenhall expanded upon de Morgan's claim by generating a histogram of mean word lengths to discriminate between literary works written by Bacon, Marlowe, and Shakespeare [3] and [4]. Mendenhall's work, though limited by the arduous task of manually counting words in classic works of literary authors, laid the foundation for textual features and computational techniques used later in authorship analysis research. This subsequent research included

the work of G.U. Yule, who in [5] made conclusions regarding disputed works based on a frequency distribution over average sentence lengths. In [6], Conrad Mascol measured frequency distributions over the average number of sentences on a printed page. The work of Wilhelm Fucks in [7] attributed authorship based on the frequency distribution over word syllables.

The most notable study in authorship attribution, however, was published in the mid-1960s by Mosteller and Wallace. As described in [8], this research, known as "The Federalist Papers," was conducted on a series of political essays written by John Jay, Alexander Hamilton, and James Madison. These essays were anonymously published in local newspapers to "persuade the citizens of the State of New York to ratify the Constitution" [8]. There was wide consensus among literary scholars on the authorship of all but twelve of these essays, which Mosteller and Wallace were able to attribute to Madison using a Bayesian statistical classification method on the frequency of a small set of context-free words [9].

Identification of these textual features, such as word length, sentence length, word-syllables, and function words used in the above-mentioned studies, later became known as writing-style or "stylometric" features used in the field of authorship analysis.

C. LEXICAL WRITING-STYLE FEATURES

Let us consider a text to be a collection of words and punctuation logically ordered to form sentences, which are in turn logically ordered to form paragraphs. Then the decomposition of the text would result in lexical features, such as words, sentences, paragraphs and punctuation, whose length, ordering, diversity, and frequency of use could be exploited to identify certain characteristics of an author's writing style. These lexical characters, known as word lengths, sentence lengths, vocabulary richness, and word n-grams, have proven successful in discriminating authors in a variety of studies using various computational techniques.

1. Word Frequencies

As the phrase implies, *word frequency* is the measure of the number of times words occur in the text of a document. The foundation for use of this distribution over words is rooted in Zipf's Law, which states that the most frequently used word in a text will appear approximately twice as often as the second most frequent, which occurs twice as often as the third, etc. Thus, the acceptance of the premise that each author has a unique style or "fingerprint" of writing has led to the use of this frequency distribution over words to determine the author of a written communication [10]. This "frequentist" approach has been used with numerous writing style features over many domains to include use of word n-grams, sentences, punctuation, characters, and character n-grams in literary works, news articles, blogs, chat, and on-line forums.

2. Unigram Model

The unigram model, also known as the "bag-of-words" model, is generated using the individual words of a document without regard to context or word order. Word frequencies in this model are calculated based on the total number of times a word appears in a document. Types are defined in [10] as distinct words that appear in a document, whereas tokens are defined as the individual occurrences of the word types. For example, the previous sentence has 25 tokens but only 20 types, since the words "types," "are," "defined," "as," and "they" appear multiple times.

Punctuation and capitalization must be carefully considered in a unigram model. If we consider a document to be a vector of words, the presence of punctuation and capitalization may increase the dimensionality of the vector space; however, removal of such may introduce ambiguity. For example, the third sentence of the previous paragraph contains the word "types" at the beginning and end of the sentence. The word is capitalized at the beginning of the sentence and a period is appended to the word at the end of the sentence, thus creating two different vector space dimensions. Moreover, a third dimension

would be introduced if the word also appeared in the middle of the sentence, instead of simply computing three occurrences of the same word. In addition, let us consider the abbreviation “U.S.” and the word “us.” Both words would increase the frequency count of type “us” if capitalization and punctuation were removed, which would not convey the intended meaning. Given the focus of our research, we believe the dimensionality reduction gained from removing punctuation would be more closely aligned to the true frequency distribution over words in each document and that the ambiguity introduced with the removal of capitalization would have minimal impact on the outcome of our results.

D. ATTRIBUTION METHODS

In [9], Stamatatos discusses a profile-based and an instance-based approach to authorship attribution. In the profile-based approach, a model is developed for each author, using a combination of various documents written by that particular author [9]. A simple procedure for developing this model would be to concatenate all the writings by a particular author into one document, then generate a model for that author based on the result, thus disregarding any style differences associated with the author’s individual writings [9]. In the instance-based approach, a model of each individual writing is generated, and then all models for each author are combined into one, thus accounting for any individual differences in documents written by a particular author [9]. We use the instance-based approach in our research in order to more accurately constrain the model generated for each author.

E. PRIOR WORK

Our research revealed five studies investigating the effect of topic on authorship attribution.

1. Investigating Topic Influence in Authorship Attribution

The first study, conducted by Mikros and Argiri in [11], tested topic-neutrality of stylometric features used in authorship attribution by performing a two-way ANOVA test to determine the interaction between authors and topics [11]. They tested the impact of topic on authorship attribution using the following stylometric features:

- Vocabulary "richness"
- Sentence length
- Function words
- Average word length
- Character frequency

The corpus they used consisted of 200 modern Greek electronic newswire articles written by two authors about two topics [11]. The data set was completely balanced, with each author writing 100 articles, half of which were written about one of two topics. Corpus statistics are identified in Table 1.

	Topics					
	Culture		***Politics***		***Total***	
Authors	Texts	Words	Texts	Words	Texts	Words
Boukalas	50	41,107	50	21,561	100	62,668
Maronitis	50	30,645	50	28,850	100	59,495
Total	***100***	***71,752***	***100***	***50,411***	***200***	***122,163***

Table 1. Distribution of words and texts across topic and author categories [From 11].

They reported a 96% overall accuracy for author classification and a 79.5% overall accuracy for topic classification across all features tested, as identified in Table 2.

Overall Author Classification accuracy = 96%	***Predicted author***	
Author	Boukalas (%)	Maronitis (%)
Boukalas	97	3
Maronitis	5	95
Overall Topic Classification accuracy = 79.5%	***Predicted topic***	
Topic	Culture (%)	Politics (%)
Culture	76	24
Politics	17	83

Table 2. Cross-validated classification results in author and topic categorization using only topic-neutral features [From 11].

From the results of the two-way ANOVA test, they concluded that there is a significant correlation between the stylometric features and topic text, and that use of such features in authorship attribution over multi-topic corpora should be done with caution.

2. Measuring Differentiability: Unmasking Pseudonymous Authors

The second study, conducted by Koppell, Schler, and Bonchek-Dokow in [12], explored the "depth of difference" between topic variability in authorship attribution using an "unmasking" technique [12]. The intuition behind this technique is to gauge how fast the cross-validation accuracy degrades during the process of iteratively removing the most distinguishable features between two classes. They used a corpus of 1,139 Hebrew-Aramaic legal query response letters written by three distinct authors about three distinct topics as identified in Table 3.

	Ritual	Business	Family
Author 1 (Yosef)	328	55	143
Author 2 (Feinstein)	157	46	120
Author 3 (Halberstam)	138	70	82

Table 3. Number of responsa written by each author on each topic in legal responsa corpus [From 12]

They used a binary classification to evaluate each author over all documents written about the same topic and to evaluate all topics over all documents written by the same author. Using their "unmasking" technique, they demonstrated consistently high accuracy scores for different author pairs on a single topic even as features were removed; however, there was a decline in accuracy as features were removed from same-author pairs on different topics [12]. Based on this result, they assessed that there were fewer features associated with the topic compared to those associated with the author and therefore these topic features were quickly eliminated during the removal process; thus, making it easier to distinguish one author from another [12]. They concluded that it is more difficult to distinguish writings by the same author on different topics than writings by different authors on the same topic [12].

3. Analyzing E-mail Text Authorship for Forensic Purposes

The third study, conducted by Corney in [2], showed that the topic did not adversely affect the identification of the author in e-mail messages. In order to support this claim, Corney used a corpus of 155 e-mail messages from three distinct authors about three distinct topics. He then developed a model for each of the three authors, using one of the three topics. Next, he used a support vector machine to test for authorship on e-mails from the remaining two topics. He reported a success rate of approximately 85% when training on one topic and testing on the others, which was consistent with the rate of success for authorship attribution across all topics [2]. Classification results are identified in Table 4.

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

Table 4. Classification results for the *food* and *travel* topics from the *discussion* data set using the *movies* topic classifier models [From 2]

We attribute Corney's results to the length and structure of e-mail communications. Often, the most discriminatory words associated with topic are in the subject of an e-mail and, therefore, if only the body of the e-mail text is evaluated, the impact of content-specific words could easily be negligible.

4. Author Identification on the Large Scale

In contrast to results obtained by Corney in [2], the fourth study, by Madigan et al. in [13], tested the effect of topic on authorship attribution in Usenet postings by two distinct authors over three distinct topics, as outlined in Table 5:

Author	GUNDOG-L	BGRASS-L bluegrass music	IN-BIRD-L birds of Indiana	TOTAL DOCUMENTS
drxxx@aol.com	10	24		34
bbxxx@inetdirect.net	6		19	25

Table 5. Two Authors from Listserv for cross-topic experiment: number of postings per group [After 13]

Just as with Corney in [2], they constructed a model of each author on one of the three topics and tested for authorship on postings written about the remaining two topics. Their results demonstrated poor performance by the unigram model; however, their bi-gram parts-of-speech model proved to be one of the best [13].

5. Outside the Cave of Shadows: Using Systematic Annotation to Enhance Authorship Attribution

Finally, the fifth study, conducted by Baayen et al. in [14], used principal components analysis (PCA) and linear discriminant analyses (LDA) to evaluate the effectiveness of grouping text by author, using stylometric features. Their data set consisted of 72 documents written by eight students. Each student wrote a total of 24 documents in three different genres about three different topics [14].

After conducting a linear discrimination for authorship using a pairwise leave-one-out cross-validation, PCA and LDA were unable to effectively distinguish the text of different authors, thus suggesting too much similarity in the training background of different authors [14]. However, when they compensated for the imbalance in the topic coverage between the text of two authors by leaving out the corresponding text of the same topic by the non-target author, they were able to achieve approximately a 10% increase in classification accuracy. This led to the conclusion that strict control of topic in cross-validation resulted in a significant increase in classification accuracy [14].

6. How Our Research is Different

Mikros and Argiri's research, conducted in [11], focused on statistical analysis of test results in order to determine the existence of a cross-correlation between author style and topic text, whereas our study is a simulation of what happens when a model encounters a novel topic.

Research conducted by Koppell, Schler, and Bonchek-Dokow in [12] explored the discriminatory nature of content-specific words on a small sample, whereas our study focuses on a testing protocol for performing this type of research, and provides a much larger data set for researchers to develop their methods.

Research conducted by Corney in [2] used a training model of only one topic and tested for authorship with e-mails pertaining to the remaining two topics written by each of the authors, whereas our research holds out a single topic for testing then generates a model for each author over the remaining 22 topics.

The study conducted by Madigan et al. in [13] uses a data set that has only one topic in common between authors, as identified in Table 5 where certain cells are empty. In addition, the cross-validation technique used by Madigan et al. differs from the novel topic cross-validation technique introduced in our study,

in that the training model for each author was based on a single topic and the test for authorship was conducted using a different topic for each author.

The imbalance compensation conducted by Baayen et al. in [14] consisted of removing documents from the test set that were written about topics for which the model had not been trained. This differs from our technique which intentionally holds out all documents pertaining to one topic for testing in order to asses the classifiers ability to predict the author of a communication for which a training model has not been developed.

F. MAXIMUM ENTROPY MODELS

Many authorship attribution NLP tasks can be formulated as statistical classification problems where the task is to make a probabilistic estimate of a class based on certain linguistic characteristics [15]. Hence, various classifiers with statistically based algorithms have proven to be effective in making certain predictions. These classifiers include Bayesian classifiers, support vector machines, and neural networks. Several studies demonstrated the superior results from maximum entropy classifiers in a variety of natural language processing tasks to include partial parsing [16], sentence boundary detection [17] and [18], prepositional phrase attachment [19], part-of-speech tagging [20], text segmentation [21], word morphology [22], language modeling [23], text classification [24], conversation thread extraction [25], and information extraction [26]; thus, we chose to use a maximum entropy classifier in our research to explore the effect of topic variability on authorship attribution.

1. Entropy and Maximum Entropy Defined

In information theory, entropy represents a measure of the amount of information in a system [27]; the lower the entropy, the greater the amount of information that can be obtained from the system. The uncertainty with regard to this information measurement is associated with a random variable, x, and is a

function of the variable's probabilities [10]. Hence, the formula for entropy is expressed as follows:

$$H\left(p\right) = H(X) = -\sum_{x} p(x) \log_2 p(x),$$

where $p(x)$ represents the probability at the value x [10]. The formula for the joint entropy of two variables, x and y, is given by the formula [10]:

$$H(X,Y) = -\sum_{x \in \mathrm{X}, y \in \mathrm{Y}} p(x,y) \log_2 p(x,y)$$

The formula for the conditional entropy of Y given some unknown value for X is given by the formula [10]:

$$H(\mathrm{Y}|\mathrm{X}) = -\sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} p(x,y) \log_2 p(y \mid x).$$

Entropy is maximized when all values of x are equally likely (e.g., uniform distribution). This property, known as maximum entropy, represents the greatest degree of uncertainty in the information. It is given by the probability distribution whose entropy is greater than or equal to all other members of a specified class, C, of distributions satisfying any constraints on the system [28]. Hence, the maximum entropy distribution, p^*, is the probability distribution with the highest entropy [10] and is given by the formula [28]:

$$\left(p\right)^* = \arg\max_{p(x) \in C} H\left(p\right).$$

2. The Principle of Maximum Entropy

According to the principle of maximum entropy, if there is no information regarding the distribution of a particular class, then the true distribution, is the one that maximizes the amount of uncertainty in the system subject to any given constraints [15].

The underlying theme of this principle was conveyed roughly 200 years ago in Laplace's "Principle of Insufficient Reason," which states the best strategy

to use when differentiating between the probabilities of two events when no information is given is to consider them both equally likely [28]. Given that such a choice for equal likelihood can seem just as arbitrary as any other choice in the absence of supporting information, E. T. Janes offers a justification as to why the maximum entropy is the one that should be chosen over all other distributions satisfying any constraints on the system [28]:

> The fact that a certain probability distribution maximizes entropy subject to certain constraints representing our incomplete information, is the fundamental property which justifies use of that distribution for inference; it agrees with everything that is known, but carefully avoids assuming anything that is not known…

3. Maximum Entropy Principle Example

The following joint probability distribution example was derived from the maximum entropy discussion in [15] and the randomness and probability discussion in [29].

Using a simple joint probability distribution to demonstrate the maximum entropy principle, consider a system that correlates high blood pressure and high cholesterol in adult males of a specific ethnicity between the ages of 35 and 50. A review of medical data provided by a nation-wide medical association revealed that 40% of the adult males in this category had high blood pressure and high cholesterol ($h_b h_c$) and 30% of adult males in this category had high blood pressure and normal cholesterol ($h_b n_c$). The joint probability of this system is depicted in Table 6:

$p(h_b, h_c)$	h_b	n_b	
h_c	.4		
n_c	.3		
total	.7		1.0

Table 6. Joint probability distribution representing only system constraints [After 15]

Hence, the above system has the following constraints:

- $p(h_b h_c) + p(h_b n_c) = 0.70$
- $p(h_b h_c) + p(h_b n_c) + p(n_b h_c) + p(n_b n_c) = 1$, where $p(n_b h_c)$ represents the probability of normal blood pressure and high cholesterol and $p(n_b n_c)$ represents the probability of normal blood pressure and normal cholesterol.

Based on the joint probability distribution outlined in Table 6 we can see that there are infinitely many distributions that will satisfy the constraints on this system. One such distribution is identified in Table 7.

$p(h_b, h_c)$	h_b	n_b	
h_c	.4	.1	.5
n_c	.3	.2	.5
total	.7	.3	1.0

Table 7. Joint probability distribution satisfying system constraints but not representing maximal entropy [After 15]

If we wish to maximize the entropy over all probability distributions that satisfy the constraints on the system, we must identify the distribution that produces the least amount of randomness or variability. That is, we must find the probability distribution, which results in an equal distribution over all unknown information about the system. This system is represented in Table 8.

$p(h_b, h_c)$	h_b	n_b	
h_c	.4	.15	.55
n_c	.3	.15	.45
	.7	.3	1.00

Table 8. Joint probability distribution satisfying constraints on the system with maximal entropy [After 15]

4. Maximum Entropy Modeling in Natural Language Processing

In natural language processing, a maximum entropy model is a flexible structure that allows unrestricted use of contextual information from various sources and combines them for classification purposes [10] and [15].

As described by R. Rosenfeld in [30], this approach to natural language processing constructs a single, combined model by extracting general observations from a collection of samples, known as training data. The knowledge gained from each sample imposes a set of constraints on the model. These constraints are normally expressed as marginal distributions requiring only that the combined estimate equal a certain probability mass on average in order to avoid any inconsistencies. The model chooses the function with the most uniform distribution (i.e., the highest entropy) from among the set of all probability functions that satisfy all of the constraints. Thus, all constraints are taken into consideration and no assumptions are made outside what is known from the data.

G. EVALUATION MEASURES

Accuracy, balanced accuracy, precision, recall, and F-score are metrics commonly used to evaluate statistical natural language processing models. Each metric uses the count of positive and negative instances to predict the true classification of the data. Computations for linguistic classification of documents are defined using the following components:

- *TruePositives* - The total number of documents from the test set which belonged to the target class and were correctly labeled as such by the classifier.
- *TrueNegatives* - The total number of documents from the test set which did not belong to the target class and were correctly labeled as such by the classifier.

- $FalsePositives$ - The total number of documents from the test set that did not belong to the target class but were erroneously labeled as belonging to the target class by the classifier.
- $FalseNegatives$ - The total number of documents from the test set that belonged to the target class, but were erroneously labeled as not belonging to the target class by the classifier.

1. Accuracy

Accuracy is the percentage of documents classified correctly by the system. The formula for accuracy is as follows [10]:

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + FalsePositives + TrueNegatives + FalseNegatives}$$

2. Balanced Accuracy

Balanced accuracy is calculated by applying a weighted average of the percentage of documents written by each author to the accuracy computation. Appendix A contains the code used to compute the balanced accuracy scores.

3. Precision

Precision measures the proportion of documents from the test set that were actually written by the target author and were correctly classified as such by the system. The formula for calculating precision is as follows [10]:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

4. Recall

Recall measures the proportion of documents from the test set that were classified as the target by the system and were actually written by the target author. The formula for calculating recall is as follows [10]:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

5. F-score

The F-score is the harmonic mean of precision and recall and is used to ensure that a favorable result for one metric is not achieved at the expense of the other. The formula for calculating the F-score is as follows [10]:

$$f-score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

III. EXPERIMENTAL DESIGN AND METHODOLOGY

A. SOURCE OF DATA

The New York Times (NYT) Annotated Corpus is a collection of 1,855,658 XML documents representing nearly all articles published in the NYT between January 1987 and June 2007. Each XML document contained one New York Times article along with meta-data identifying information pertaining to the document to include the document's title, author, and topic. Although 99.95% of the documents contained tags for the topic, only 48.18% of the documents contained tags for the author [31]. Therefore, in order to cross-correlate authors with topics and vice-versa for each document, a relational database was created and populated using a subset of 871,050 of the NYT Annotated Corpus documents that were tagged with both author and topic as well as title.

B. DATA SELECTION

Documents written by a single author about a single topic were selected from a relational database in order to generate the following two subsets of data used to conduct these experiments: a binary data set and a multi-category data set. The binary data set was balanced across two authors and unbalanced across four topics. The multi-category data set was unbalanced across 15 authors and unbalanced across 23 topics.

1. Relational Database

A MySQL relational database was created and populated with the following five tables of information extracted from the subset of 871,050 XML documents of the NYT Annotated Corpus: document, author, topic, writtenBy, and writtenAbout. Full details of the database structure are given in Appendix B.

2. Single Author and Single Topic

We separated a total of 646,742 documents in the database because they were written by more than one author or were identified with more than one topic. The remaining 224,308 documents were used to generate two subsets of documents of sufficient size with a minimum number of samples from each author and each topic.

3. Binary Data Subset

In the binary data subset, a total of 3,000 documents were selected from the 224,308 that were written by a single author about a single topic. This subset consisted of documents written by two distinct authors who wrote an equal number of documents. These documents were about four distinct topics that appeared in at least 500 of the 3,000 documents. Table 9 is a list of authors along with the corresponding total number of documents in the subset written by each author. Table 10 is a list of topics along with the corresponding total number of documents in the subset written about each topic. The average vocabulary size over all documents was 282.57 with a minimum vocabulary size of 2 and a maximum of 1,304.

MYSQL DATABASE AUTHOR ID	AUTHOR	AUTHOR TOTAL DOCS	MYSQL DATABASE TOPIC ID	TOPIC	TOPIC TOTAL DOCS
A100024	Dunning, Jennifer	1500	T50031	Music	1
			T50048	Motion Pictures	6
			T50050	Dancing	1,467
			T50128	Theater	26
A100078 A105328	Holden, Stephen	1500	T50031	Music	500
			T50048	Motion Pictures	494
			T50050	Dancing	6
			T50128	Theater	500

Table 9. Author-Topic Data Correlation

MYSQL DATABASE TOPIC ID	TOPIC	TOPIC TOTAL DOCS	MYSQL DATABASE AUTHOR ID	AUTHOR	AUTHOR TOTAL DOCS
T50031	Music	501	A100024	Dunning, Jennifer	1
			A100078	Holden, Stephen	500
T50048	Motion Pictures	500	A100024	Dunning, Jennifer	6
			A100078/A105328	Holden, Stephen	494
T50050	Dancing	1473	A100024	Dunning, Jennifer	1,467
			A100078	Holden, Stephen	6
T50128	Theater	526	A100024	Dunning, Jennifer	26
			A100078/A105328	Holden, Stephen	500

Table 10. Topic-Author Data Correlation

4. Multi-category Data Subset

In the multi-category data set, a total of 18,862 documents were selected from the 224,308 documents that were written by a single author about a single topic. This subset consisted of documents written by a total of 15 distinct authors and about 23 distinct topics. Table 11 lists the topic categories along with their corresponding topic identifications. The minimum number of documents written by a particular author was 730 and the maximum number was 2,912. The minimum number of documents written about a particular topic was 35 and the maximum number was 2,907. The average vocabulary size over all documents was 306.12 with a minimum vocabulary size of 25 and a maximum of 2,889. Appendix C contains a matrix of the author-topic total document counts.

T50014	Books and Literature	T50187	Appointments and Executive Changes
T50031	Music	T51556	Deaths (Obituaries)
T50013	Baseball	T50172	Advertising and Marketing
T50128	Theater	T50383	Golf
T50012	Football	T50368	Boxing
T50048	Motion Pictures	T50273	Horse Racing
T50015	Art	T50222	Photography
T50097	Basketball	T50338	Soccer
T50050	Dancing	T50049	Suspensions, Dismissals and Resignations
T50006	Television	T50214	Cooking and Cookbooks
T50115	Hockey, Ice	T50077	Food
T50136	Restaurants		

Table 11. Multi-category data set topic categories and identifications

C. DATA PREPARATION AND FEATURE SELECTION

A query of the MySQL database was conducted in order to generate a directory of files containing only the text portion of the XML documents. The text of each document was stored as a separate text file named for the author, topic, and document identifications (i.e., A100024_T50006_0046467.txt). It is important to note that the regular expression that extracts the text used to populate the text field of the document table did not always capture the lead paragraph of the document since the tagging of the NYT Annotated corpus in the XML distribution was not always consistent.[1] Documents used in these experiments were removed from the subset in cases where this discrepancy resulted in an empty text file.

The records in the database differentiated documents written by authors whose names appeared in all capital letters and those written in upper and lower case letters. Therefore, for these experiments, documents written by author Stephen Holden identified by author ID A100078 and A105328 were combined and documents written by author Stuart Elliott identified by author ID A104872 and A111915 were combined.

Punctuation was removed from the text of the documents by replacing all non-alphanumeric characters with the empty string. In addition, all letters were converted to lower case to more accurately reflect the dimensionality of the vector space.

Finally, to facilitate use of unigram word features, data was processed into word grams by tokenizing words on whitespace.

[1] Some XML documents in the NYT Annotated Corpus contained an XML tag for a lead paragraph then repeated the lead paragraph twice in the XML tag for the full text whereas other documents did not.

Some XML documents in the NYT Annotated Corpus contained a lead paragraph and a full text that repeated the lead paragraph twice while other documents did not.

D. MAXIMUM ENTROPY GA MODEL (MEGAM) OPTIMIZATION PACKAGE

We used the Maximum Entropy GA Model (MEGAM) Optimization Package, written by Hal Daume[2]. This package optimizes logistic regression classifiers through efficient implementation of the conjugate gradient method for binary problems and the limited memory BFGS (Broyden-Fletcher-Goldarb-Shanno) method for multiclass problems. The training algorithms used for these two methods are presented in Appendix D and can be found on Daume's website, along with an unpublished paper describing the algorithms employed.

As described by Daume on his website, this software can be used to solve three types of problems:

- binary classification (classes are 0 and 1)
- binomial regression (classes are real values between 0 and 1)
- multiclass classification (classes are 0, 1, 2, etc.)

The software takes a set of training vectors as input and uses an iterative optimization process to produce a set of weights. These weights are then used in conjunction with a set of test vectors to generate probabilities that are used to predict the class. Figure 1 graphically depicts the MEGAM classification process.

[2] Available for download from the University of Utah School of Computing website: http://www.cs.utah.edu/~hal/megam/index.html.

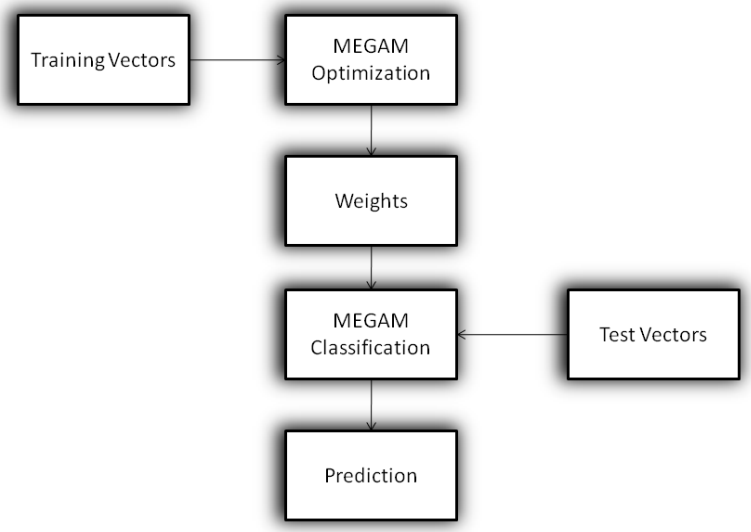

Figure 1. MEGAM Optimization Package Classification Process [After 10]

E. SCENARIO 1: STANDARD 10-FOLD CROSS-VALIDATION

In the first of two evaluation scenarios, we conducted a randomized 10-fold cross-validation for both the binary and multi-category data sets. The MEGAM classifier was trained on 90% of the documents, then tested on the remaining 10% for each fold using a binary classification for the data set with two authors and using a multiclass classification for the data set with 15 authors. The 10% of test documents in each fold consisted of 10% of the documents written by each author with the last fold also including any remaining documents not tested in folds one through nine.

F. SCENARIO 2: NOVEL TOPIC CROSS-VALIDATION

In the second scenario, we conducted a leave-one-topic-out n-fold cross-validation where n represented the total number of topics in the data set. In each fold of the experiments, the MEGAM classifier was tested on all documents pertaining to one topic and trained on all other documents pertaining to the

remaining n-1 topics. There were a total of four topics in the binary data set, and a total of 23 topics in the multi-category data set.

G. PERFORMANCE MEASURES

Accuracy, balanced accuracy, precision, recall, and F-score were the performance measures used to evaluate the results of the experiments. All metrics were computed for each fold of the binary data set; however, only accuracy and balanced accuracy were computed for the multi-category data set. Table 12 depicts the confusion matrix used to compute the precision, recall, and F-score for the two authors in the binary data set.

	GROUND TRUTH	
	A100024	**A100078/A105328**
A100024	TruePositive	FalsePositive
A100078/A105328	FalseNegative	TrueNegative

Table 12. Binary Data Set Classification Confusion Matrix

The balanced accuracy was necessary in the novel topic cross-validation of both data sets in order to provide an indication of the degree to which errors are made on less frequent topic categories. The balanced accuracy scores computed in the standard 10-fold cross-validation of the binary data set were virtually identical to the standard accuracy scores because the binary data set was balanced across authors in the data selection process.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS AND ANALYSIS

A. RESULTS

We introduced two scenarios used for experimentation. Results obtained from the first scenario were used to establish the baseline for comparing against our novel topic cross-validation technique (the second evaluation scenario). The first scenario used a standard 10-fold cross-validation to compute performance measures which were then compared to results obtained from the second scenario. The second evaluation scenario used an n-fold cross-validation technique where n, representing the total number of topics in the data set, was iterated over leaving all documents pertaining to one topic out for testing while training over all documents pertaining to the remaining n-1 topics.

The degradation in performance presented from our experiments was computed from the difference between results reported in the two scenarios. This difference represented the effect certain content-specific words associated with a particular topic had on the classifiers ability to detect the author having modeled writings from the author over the remaining n-1 topics. The greater the degradation in performance, the more discriminatory words associated with topic text negatively impacted the author prediction models.

1. Scenario 1: Standard 10-Fold Cross-Validation

a. Binary Data Set

In the binary data set, the test sets consisted of a total of 300 documents, 150 of which were written by each of the two authors. The training sets consisted of a total of 2,700 documents, 1,350 of which were written by each of the two authors. The average vocabulary for the test sets was 20,410 with a minimum vocabulary of 19,752 and a maximum vocabulary of 20,885. The average vocabulary size for the training sets was 63,641 with a minimum

vocabulary of 63,267 and a maximum vocabulary of 63,701. The results of the 10-fold cross-validation for the binary data set are described in Table 13. The accuracy results for the 10 folds represented the total number of observations for the data set. The precision, recall, F-score, accuracy, and balanced accuracy were calculated for each fold, and then the results were averaged to establish the scores provided as a snapshot in Table 14.

Fold	Train Vocab	Test Vocab	TP	TN	FN	FP	Correct	Incorrect	Precision	Recall	F-score	Accuracy	Balanced Accuracy
1	63,355	20,407	143	157	0	0	300	0	1.0000	1.0000	1.0000	1.0000	1.0000
2	63,298	20,626	147	153	0	0	300	0	1.0000	1.0000	1.0000	1.0000	1.0000
3	63,267	20,885	144	155	1	0	299	1	1.0000	0.9931	0.9965	0.9967	0.9966
4	63,459	20,420	147	149	4	0	296	4	1.0000	0.9735	0.9865	0.9867	0.9868
5	63,584	20,471	167	132	1	0	299	1	1.0000	0.9940	0.9970	0.9967	0.9970
6	63,348	20,844	147	152	1	0	299	1	1.0000	0.9932	0.9966	0.9967	0.9966
7	63,573	19,752	139	159	1	1	298	2	0.9928	0.9928	0.9928	0.9933	0.9933
8	63,520	20,211	145	154	1	0	299	1	1.0000	0.9931	0.9965	0.9967	0.9966
9	63,701	20,062	153	145	2	0	298	2	1.0000	0.9870	0.9935	0.9933	0.9935
10	63,608	20,417	156	143	1	0	299	1	1.0000	0.9936	0.9968	0.9967	0.9968
Average	63,471	20,410							0.9993	0.9920	0.9956	0.9957	0.9957
StdDev									0.0023	0.0075	0.0039	0.0039	0.0038

Table 13. Binary data set standard 10-fold cross-validation results

	Average	Std.Deviation
Accuracy	0.9957	0.0039
Bal.Accuracy	0.9957	0.0038
Precision	0.9993	0.0023
Recall	0.9920	0.0075
F-score	0.9956	0.0039

Table 14. Binary data set 10-fold cross-validation performance measure results

b. Multi-category Data Set

In the multi-category data set, the test set for folds one through nine consisted of a total of 1,880 documents and the test set for fold ten consisted of 1,942 documents. Furthermore, the training set for folds one through nine consisted of a total of 16,982 documents and the training set for fold ten consisted of a total of 16,920 documents. The average vocabulary for the test set was 56,198 with a minimum vocabulary size of 55,221 and a maximum vocabulary size of 56,926. The average training vocabulary size was 159,599

with a minimum vocabulary size of 159,269 and a maximum vocabulary size of 159,891. The results of the 10-fold cross-validation for the multi-category data set are described in Table 15. The accuracy results for the 10 folds represented the total number of observations for the data set. The accuracy and balanced accuracy were calculated for each fold. The results were then averaged to establish the accuracy and balanced accuracy scores provided as a snapshot in Table 16.

Fold	Train Docs	Test Docs	Train Vocab	Test Vocab	Correct	Incorrect	Accuracy	Balanced Accuracy
1	16,982	1,880	159,681	56,514	1,346	534	0.7160	0.6887
2	16,982	1,880	159,601	55,792	1,000	880	0.5319	0.4139
3	16,982	1,880	159,891	55,666	1,119	761	0.5952	0.5972
4	16,982	1,880	159,504	56,508	1,317	563	0.7005	0.6918
5	16,982	1,880	159,869	55,221	1,051	829	0.5590	0.4663
6	16,982	1,880	159,741	55,824	1,145	735	0.6090	0.5142
7	16,982	1,880	159,607	56,101	989	891	0.5261	0.3908
8	16,982	1,880	159,473	56,606	1,086	794	0.5777	0.6010
9	16,982	1,880	159,269	56,826	957	923	0.5090	0.3741
10	16,920	1,942	159,353	56,926	1,000	942	0.5149	0.3855
Average			159,599	56,198			0.5839	0.5123
StdDev							0.0737	0.1248

Table 15. Multi-category data set standard 10-fold cross-validation results

	Average	Std.Deviation
Accuracy	0.5839	0.0737
Bal.Accuracy	0.5123	0.1248

Table 16. Multi-category data set accuracy and balanced accuracy performance measure results

2. Scenario 2: Novel Topic Cross-Validation

a. Binary Data Set

In the binary data set, a novel topic cross-validation was used to compute the results of the author prediction task. That is, of the four topics covered in the data set, all documents pertaining to one topic were held out for testing and a training model for each author was developed using all documents pertaining to the remaining three topics for each fold of cross-validation. For this

data set, there was only a miniscule difference between the accuracy and balanced accuracy results of the standard 10-fold cross-validation as would be expected given the fact that each author wrote an equal number of documents and a randomized but balanced cross-selection of documents was chosen for each fold of the cross-validation. There was, however, a slight increase in the balanced accuracy of the novel topic cross-validation as compared to the standard 10-fold cross-validation, which we attributed to the one topic category that was three times the size of any of the other topics.

With regard to the precision, recall, and F-score performance measures, there was a 39.5% degradation in precision and a 16.4% degradation in recall, resulting in a 40.6% degradation in F-score from the standard 10-fold cross-validation as computed from results provided in Tables 14 and 17. There was also a 12.3% decline in accuracy and a decline of 8.9% in balanced accuracy against the standard 10-fold cross-validation as computed from results provided in Table 18. The full performance measures for each topic category are detailed in Appendix E.

	Average	Std.Deviation
Accuracy	0.8722	0.2129
Bal.Accuracy	0.9060	0.0916
Precision	0.6031	0.3909
Recall	0.8269	0.1959
F-score	0.5888	0.3101

Table 17. Binary data set novel topic cross-validation precision, recall, and F-score performance measure results

	Accuracy		**Balanced Accuracy**	
	Mean	**Std.Dev.**	**Mean**	**Std.Dev.**
Standard 10-fold cross-validation	0.9957	0.0039	0.9957	0.0038
Novel topic cross-validation	0.8722	0.2129	0.9060	0.0916

Table 18. Binary data set accuracy and balanced accuracy performance measure results

b. Multi-category Data Set

In the multi-category data set, a novel topic cross-validation was used to compute the results of the author prediction task. That is, of the 23 topics covered in the data set, all documents pertaining to one topic were held out for testing and a training model for each author was developed using all documents pertaining to the remaining 22 topics for each fold of cross-validation. There was a 30.3% difference between accuracy and balanced accuracy results in the multi-category data set as would be expected given the wide variation in the number of documents written about each topic. There was also a 27.6% decline in accuracy and a decline of 41.7% in balanced accuracy against the standard 10-fold cross-validation as computed from results provided in Table 19. Appendix F provides detailed results for each topic category.

	Accuracy		Balanced Accuracy	
	Mean	Std.Dev.	Mean	Std.Dev.
Standard 10-fold cross-validation	0.5839	0.0737	0.5123	0.1248
Novel topic cross-valication	0.3079	0.3431	0.0952	0.0688

Table 19. Multi-category data set accuracy and balanced accuracy performance results

B. ANALYSIS

In our analysis, we considered the highest and lowest accuracy results, top and bottom three accuracy results, the total test and train set documents, the total test and train set vocabulary counts, and the top and bottom 50 word tokens.

1. Binary Data Set

Figures 2 and 3 graphically depict the 4-fold cross validation accuracy and balanced accuracy results for the binary data set, respectively. As a reminder, in our binary data set, we designate author A100024 as the target and author

A100078/A105328 as the non-target for classification purposes in order to compute precision and recall.

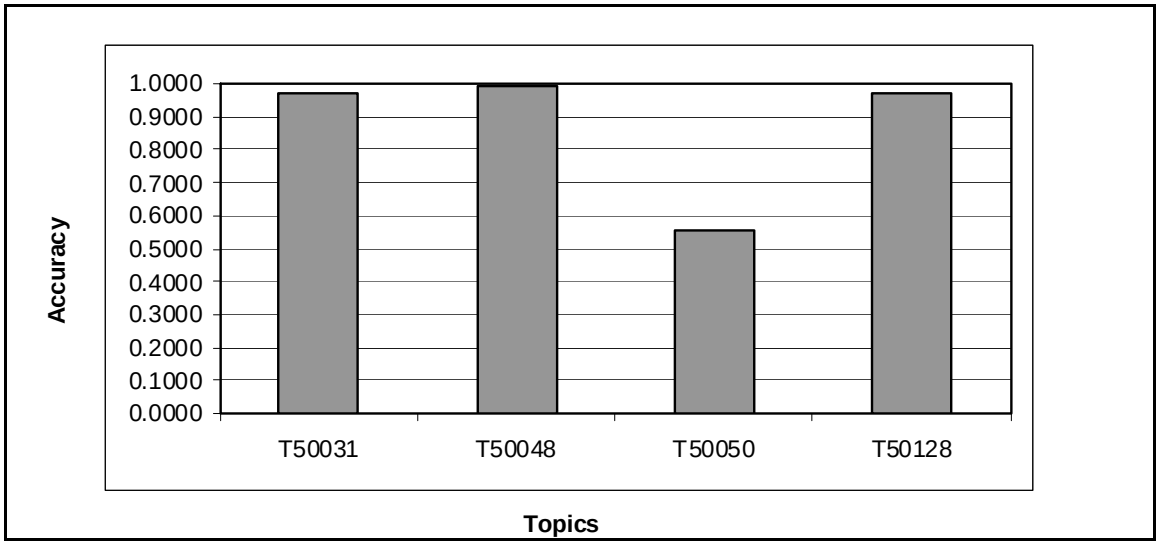

Figure 2. Binary data set novel topic cross-validation **accuracy** results

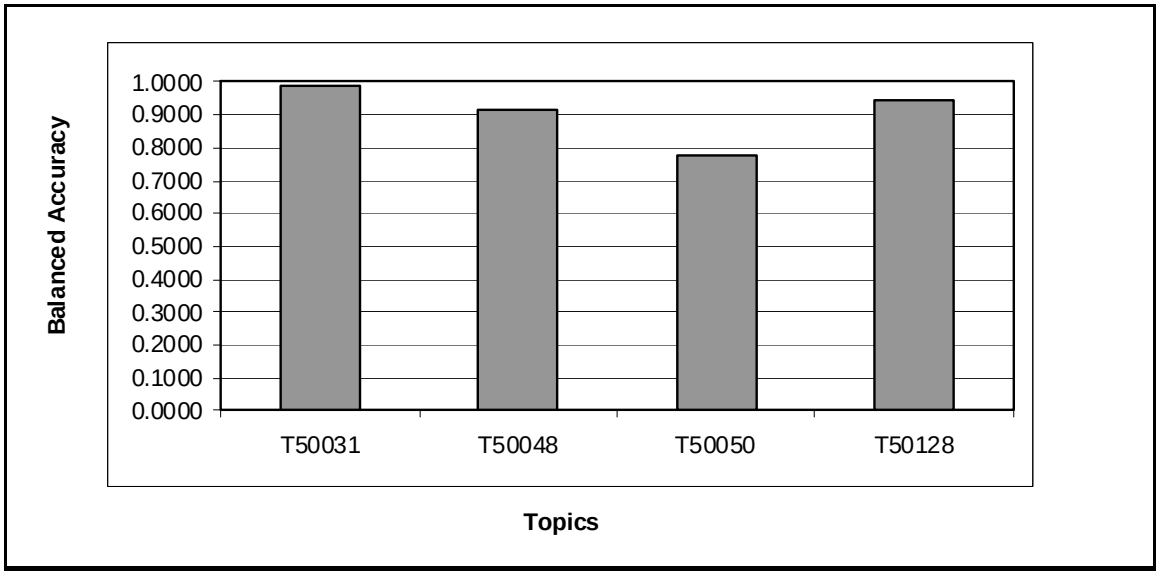

Figure 3. Binary data set novel topic cross-validation **balanced accuracy** results

a. Highest Precision and Lowest Recall Scores

In the binary data set, the precision for each topic category ranged from 6.6% to 100.0% and recall for each topic category ranged from 55.1% to 100%. In order to determine what may have accounted for the highest and lowest precision and recall scores, first consideration was given to the number of documents in the test and train sets. For this data set, there were a total of 1,473 test documents in the topic category T50050 (Dancing) which resulted in a perfect precision score compared to the roughly 500 documents in the other three topic categories. However, more importantly, of these 1,473 documents, 1,467 were written by the target author, and only 6 were written by the non-target author. It stands to reason, that calling everything the target author would yield the highest precision results; however, it is important to note that this topic category yielded the lowest recall, accuracy, and balanced accuracy results at 55.1%, 55.3%, and 77.5%, respectively.

b. Lowest Precision and Highest Recall Scores

The lowest precision of 6.6% was obtained in the topic category T50031 (Music) in the binary data set. Of the 501 test documents in this category, one was written by the target author, and the other 500 were written by the non-target author. Once again, given the overwhelming imbalance of test documents for each author, it stands to reason that calling everything by the target author would yield the worst precision results and the best recall. It is important to note that this topic category yielded the lowest F-score at 12.5% and the highest recall and balanced accuracy results at 100% and 98.6%, respectively.

c. F-scores

Analysis of F-scores in the binary data set for each topic category of our novel topic cross-validation, revealed an interesting phenomenon, which was difficult to account for. That is, the lowest F-score of 12.5% was obtained

from the topic category T50031 (Music) which resulted in the second highest accuracy and highest balanced accuracy scores of 97.2% and 98.6%, respectively. The second lowest F-score of 71% was obtained from the topic category T50050 (Dancing) which resulted in the lowest accuracy and balanced accuracy results of 55.3% and 77.5%, respectively. These results were consistent with the non-biased representation of the F-score in relation to precision and recall for the most prolific target in an overwhelmingly unbalanced data set; however, there was no discernable pattern useful for analysis of results.

d. Accuracy and Balanced Accuracy Scores

We attribute the difference between accuracy and balanced accuracy to disparity in the ratio of train to test documents in the topic category folds. There was a 5 to 1 ratio of train to test documents in the most prolific topic category where as the remaining three categories had approximately a 1 to 1 ratio of train to test documents. For example, topic category T50050 (Dancing) had approximately five training documents for every one test document as compared to the topic category T50048 (Motion Pictures) which had approximately one training document for every 1 test document.

The balanced accuracy of the novel topic cross-validation in the binary data set was 3.3% higher than the average accuracy. We attribute this increase in accuracy to the categorical imbalance of topics used in each fold of cross-validation. One topic was approximately 3 times more prevalent in the data set than each of the other three topics. That is, topic category T50050 (Dancing) comprised 49.1% of the total documents in the data set whereas topic categories T50031 (Music), T50048 (Motion Pictures), and T50128 (Theater) comprised approximately 16.9% each of the remaining documents.

e. Overall F-score, Accuracy, and Balanced Accuracy Results

Finally, in the binary data set, the overall F-score, accuracy, and balanced accuracy results of the novel topic cross-validation as compared to standard 10-fold cross-validation demonstrated a significant degradation in the classifiers ability to make an accurate prediction.

2. Multi-category Data Set

Figures 4 and 5 graphically depict the novel topic cross-validation accuracy and balanced accuracy results for the multi-category data set, respectively.

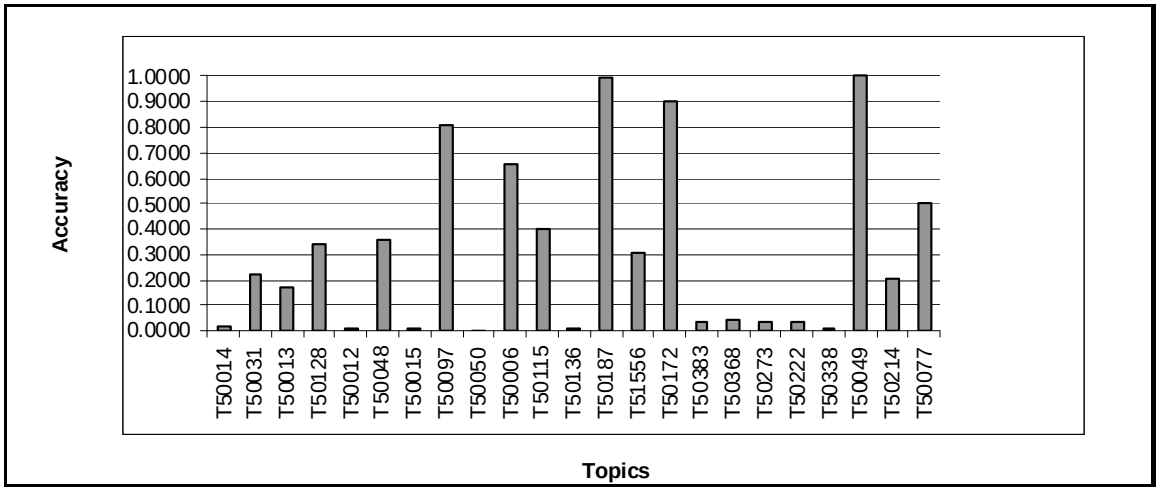

Figure 4. Multi-category data set novel topic cross-validation **accuracy** results

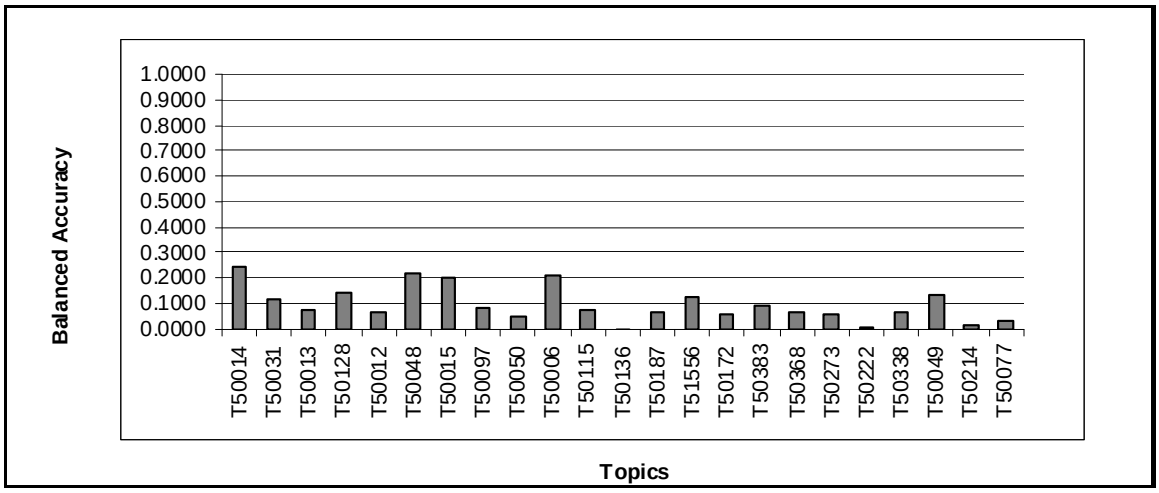

Figure 5. Multi-category data set novel topic cross-validation **balanced accuracy** results

a. Highest Accuracy Score

In the multi-category data set, the highest accuracy obtained was 100% in the topic category T50049 (Suspensions, Dismissals, and Resignations). In order to determine what may have accounted for the perfect accuracy score, we first considered the number of documents in the test and train sets for this category. The test fold for topic category T50049 (Suspensions, Dismissals, and Resignations) had a total of 64 test documents and 18,798 training documents. We then compared accuracy results against a comparable test/train split fold. The test fold for topic category T50077 (Food) had the same number of documents in the test and train sets, however, the resulting accuracy for this topic fold was 50%. Thus, we ruled out the number of documents in the test/train split as accounting for this perfect accuracy score.

Next, we considered the size of the vocabulary in the test/train split. The topic category T50077 (Food) had the closest vocabulary size with only 290 fewer vocabulary words in the training set and 1,643 more vocabulary words in the test set, yet this category only yielded a 50% accuracy. Thus, more words were tested.

The topic category T50049 (Suspensions, Dismissals and Resignations) had the smallest test vocabulary count at 2,860 even though this category was not the category with the minimum number of test documents. There were two other categories with fewer test documents (i.e., T50006 (Television) with 35 test documents and T51556 (Deaths (Obituaries) with 55 test documents), yet they obtained accuracy percentages in the 60's and 30's, respectively. The next highest vocabulary word count from our perfect accuracy category was 4,503 in the topic category T50077 (Food) with 50% accuracy, yet the third lowest vocabulary word count was in the topic category T50187 (Appointments and Executive Changes) with a vocabulary word count of 5,791; however, this topic category resulted in the second highest accuracy score of 99.6%. Thus, there was no discernable correlation between vocabulary size in the test/train split and accuracy score.

Finally, we considered the topic area of the top three and bottom three accuracy category folds. We reviewed the top 50 tokens and the bottom 50 tokens. The top 50 word tokens were mostly content-free words also referred to as stop-words (i.e., "the", "of", "and", "is", "over", "by", etc.). The bottom 50 words, however, tended to be words that only appeared in the vocabulary once and were not necessarily indicative of the topic at hand. By observing the word vector of the test set, we noticed a large number of content-free words in the word vectors for the top three performing accuracy categories. In contrast, however, we observed many more content-specific words in the word vectors for the bottom three performing accuracy categories. That is, we noticed many words that would only be used in the context of the topic at hand.

b. Lowest Accuracy Score

In the novel topic, cross-validation of the multi-category data set the lowest accuracy obtained was 0.2% in the topic category T50050 (Dancing). This testing fold consisted of 1,543 documents with a training set of 17,319 documents. The topic category had only 19 fewer documents for testing and

training as topic category T50128 (Theater) which ranked 9 of 23 in accuracy. The second lowest ranking accuracy was from the topic category T50012 (Football) which consisted of 1,044 test documents and 17,818 training documents and a test set vocabulary of 21,668 and a training vocabulary of 164,601. Thus, we concluded that there is no correlation between the number of vocabulary words tested, or modeled for training. There does, however, appear to be a correlation between the generality of the subject matter and the level of accuracy. That is, the more general the topic, the higher the accuracy and the more specific the topic, the lower the accuracy result. Table 21 depicts the top and bottom three topic category rankings for accuracy. From this observation, we could reasonably assess that the more general the topic category, the better the classifier was at discriminating between authors given documents written about topics for which the classifier had never seen before.

Top 3 Accuracy Categories		Bottom Three Accuracy Categories	
Topic	**Accuracy**	**Topic**	**Accuracy**
Suspensions, Dismissals, and Resignations	1.0000	Dancing	0.0026
Appointments and Executive Changes	0.9966	Football	0.0057
Advertising and Marketing	0.9059	Soccer	0.0064

Table 20. Novel topic cross-validation top and bottom three accuracy result topic categories

c. Highest and Lowest Balanced Accuracy Scores

In the multi-category data set, the highest balanced accuracy obtained was 24.2% in the topic category T50014 (Cooking and Cookbooks) and the lowest balanced accuracy obtained was 0.08% in the topic category T50136 (Restaurants). Our belief was that the balanced accuracy would highlight the degree to which errors were being made on the less frequent topic categories; however, no discernable pattern was detected from the balanced accuracy results identified in Appendix F in order to confirm this belief.

d. Overall Accuracy and Balanced Accuracy Scores

The overall results showed a 27.6% decline in accuracy and a 41.7% decline in balanced accuracy as computed from results provided in Table 22. The full performance measures for each topic category of the binary data set are detailed in Appendix F.

	Average	Std.Deviation
Accuracy	0.3079	0.3431
Bal.Accuracy	0.0952	0.0688

Table 21. Novel topic cross-validation multi-category data set accuracy and balanced accuracy performance measure results

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY AND RECOMMENDATIONS

A. SUMMARY AND CONCLUSIONS

This research investigated the impact of a novel topic on the ability of a maximum entropy classifier to discriminate between authors in binary and multi-classification authorship attribution problems. In order to study novel topic impact on author identification, we used two subsets of data from the New York Times Annotated corpus. The first data set of 3,000 documents was balanced across two authors and unbalanced across four topics. The second data set of 18,862 documents was unbalanced across 15 authors and unbalanced across 23 topics. A baseline was established for both data sets using a standard 10-fold cross-validation where test documents for each fold consisted of 10 percent of the documents written by each author. The final fold also included any documents not tested in the first nine folds. The remaining 90 percent of documents were used for training in each fold. Performance measures were then averaged, and compared against a novel topic cross-validation for each data set.

The results of these experiments demonstrated a degradation in performance of all evaluation measures to include a 12.3 percent decline in accuracy for the binary data set and a 27.6 percent decline in accuracy for the multi-category data set between standard and “novel topic” cross-validation. The unbalanced nature of the data across topics also appeared to affect the classifiers ability to discriminate between authors in the novel topic cross-validation. This was prevalent in the binary data set with the relatively consistent accuracy scores of the three topic categories that had roughly the same number of documents compared to the 30 percent decline in accuracy of the one category that had three times as many documents as the others. This result suggests that a balanced data set across topics would have yielded consistent

accuracy scores across all topic categories. Hence, only resulting in a more moderate degradation in performance compared to the baseline.

As for the multi-category data set, the analysis of results were complicated by the imbalance of both the topic and author classes. This imbalance probably accounted for a portion of the resulting degradation in performance; however, analysis revealed an additional factor worth consideration. The more specific the topic category, the more difficult it was for the classifier to discriminate between authors. That is, for topics such as “Dancing”, “Football”, and “Soccer”, which resulted in the lowest three accuracy scores, it was clear that the absence of content-specific words in the training model negatively impacted the classifier’s ability to predict the author of the document. This implied that the unigram model of a document included both words associated with the author’s particular style of writing as well as the topic of the document at hand. On the contrary, the classifier did a much better job discriminating among authors who had written documents about more general or broader topic areas such as “Suspensions, Dismissals, and Resignations”, “Appointments and Executive Changes”, and “Advertising and Marketing,” which resulted in highest three accuracy scores. This outcome suggests that the more general the topic, the less impact the topic had on the classifier’s ability to predict the author. In contrast, the more specific the topic, the greater impact the topic had on the classifiers ability to predict the author.

B. RECOMMENDATIONS FOR FUTURE WORK

1. Decomposition of Document Word Vector

Additional research needs to be conducted on how to discriminate between stylometric and topic spaces in the unigram word-vector model of a document. If the words used in a document are composed of those words associated with the author’s particular style of writing as well as those words associated with the topic of the document, then linearly discriminating between

these two categories of words would facilitate the use of vector operations for removing the topic feature vector in order to better discriminate author. Advances in this area would eliminate the need to control for topic when attempting to predict the author of a document and would thus allow for a more realistic authorship attribution model.

2. Testing Across Multiple Domains

Authorship attribution needs to be tested across different domains in order to determine if the stylometric vector is truly indicative of the author's particular style of writing versus the style dictated by the source. This research would require a corpus collection of documents from various news sources such as the Wall Street Journal, New York Post, Huffington Post, or Washington Post. The challenge for creating such a corpus of documents would be in identifying authors who have written articles for multiple news organizations.

3. Multiple Authors

One area of research that appears virtually unexplored is authorship detection in communications written by multiple authors. The famous work on the "Federalist Papers" done by Mosteller and Wallace eluded to this problem with twelve of the disputed papers being claimed by both Madison and Hamilton. This research would require a corpus of single and multiple author documents where models could be created and tested against documents written by multiple authors.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A: BALANCED ACCURACY CODE

```
#!/usr/bin/env python
#compute accuracy
#usage: argv[0] test*.txt

from sys import argv

intern = {}

def acc(truthFile):
    truth      = []
    prediction = []
    hsh        = {}
    f          = file(truthFile)
    for line in f.readlines() :
        line     = line.strip()
        toks     = line.split()
        t        = toks[0]
        if t not in intern : intern[t] = len(intern)
        truth.append(intern[t])
        p        = toks[1]
        if p not in intern : intern[p] = len(intern)
        prediction.append(intern[p])
        if intern[t] not in hsh : hsh[intern[t]] = 0
        hsh[intern[t]] += 1

    cCounts = [ 0.0 for i in range(len(intern)) ]
    for k in intern.keys() : cCounts[intern[k]] = hsh[intern[k]] if intern[k] in hsh else 0
    w = [ 1.0/(c *len(cCounts)) if c > 0 else 0 for c in cCounts]
    correct = 0.0
    for i in range(len(truth)) :
        correct += w[truth[i]]*(truth[i]==prediction[i])
    return correct

if __name__ == "__main__":
    for truthFile in argv[1:] : print str(acc(truthFile))
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B : RELATIONAL DATABASE DESCRIPTION

The following five tables of information, as diagrammed in Figure 6, were extracted from the subset of 871,050 XML documents of the NYT Annotated Corpus:

- Document
    - A total of 871,050 records (equates to total number of documents in the database)
    - Each document was written by anywhere from 1 to 21 different authors.
    - Each document was written about anywhere from 1 to 44 different topics.
    - Table attributes are as follows:
        - docID – This is the primary key for the document table and consists of a unique 6-digit numeric file name with a .xml extension. These file names are identical to the XML file names in the NYT Annotated Corpus.
        - title – A one-sentence description of the overall content of the article.
        - text – The portion of the document consisting of the text of the article with all HTML tags removed.
        - singleAuthTopicSubset – Represents a subset of 224,308 records in the database consisting of only those documents written by a single author about a single topic.
        - SATbinSubset – Represents a subset of 3,000 records in the database consisting of documents

written by a single author about a single topic where each author wrote 1,500 of the 3,000 documents and each topic appeared in at least 500 of the 3,000 documents.

        - SATmultiSubset – Represents a subset of 18,862 records in the database consisting of documents written by a single author about a single topic where each author wrote anywhere from 730 to 3,298 documents and each topic appeared in anywhere from 35 to 2,912 documents.

- Author
    - A total of 26,838 records (equates to total number of distinct authors in the database).
    - Each author wrote anywhere from 1 to 3,959 documents.
    - Table attributes are as follows:
        - authID – This is the primary key for the author table and consists of a unique sequential one-up alpha-numeric ID in the range A100000-A126837, inclusive.
        - name – This column specifies the name of the author in the form last name, first name, middle initial.
- Topic
    - A total of 1,622 records (equates to total number of distinct topics in the database).
    - Each topic was identified in anywhere from 1 to 140,830 documents.
    - Table attributes are as follows:

        - topicID – This is the primary key for the topic table and consists of a unique sequential one-up alpha-numeric ID in the range T50000-T51621, inclusive.
        - topic – A general word or phrase description of the article's subject matter.

- writtenBy
    - A total of 871,856 records representing unique author-document combinations from the author and document tables.
    - Table attributes are as follows:
        - docID – As described in document table attribute above.
        - authID – As described in author table attribute above.

- writtenAbout
    - A total of 3,130,008 records representing unique topic-document combinations from the topic and document tables.
    - Attributes are as follows:
        - docID – As described in document table attribute above.
        - topicID – As described in topic table attribute above.

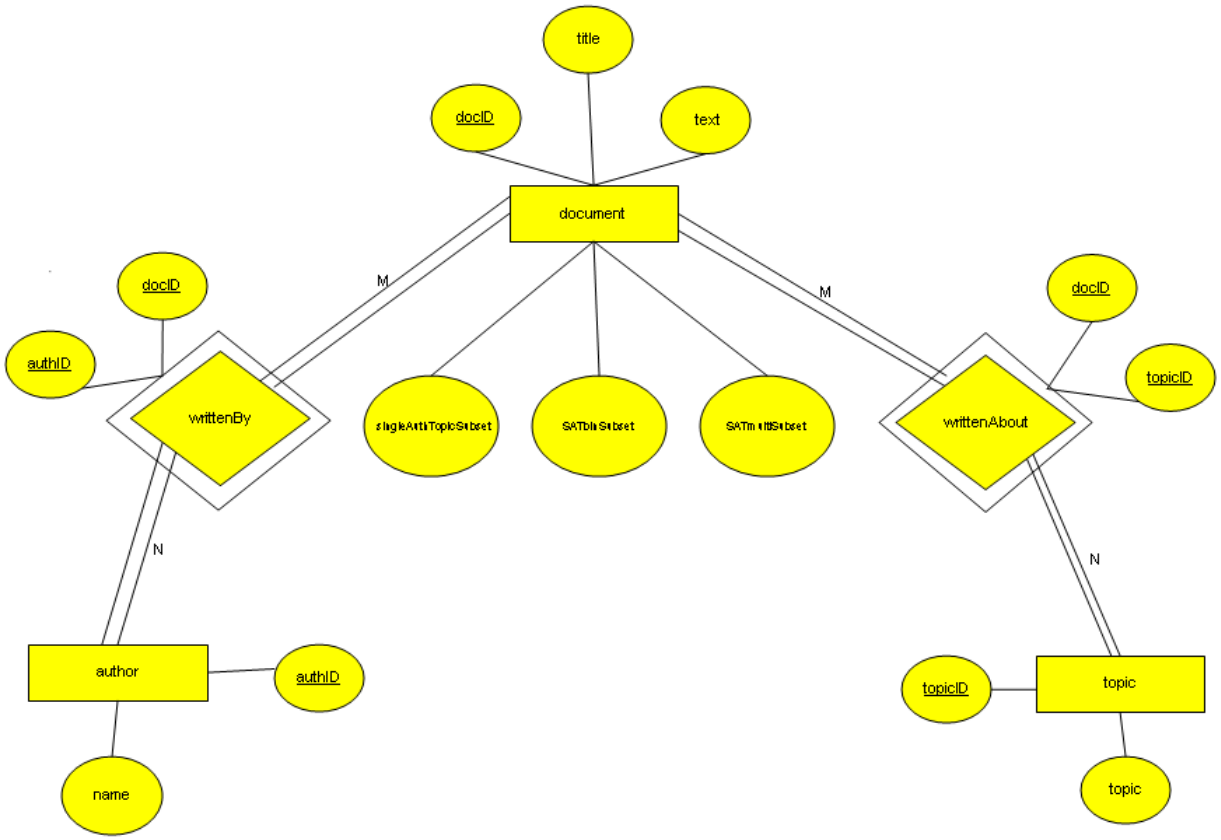

Figure 6. New York Times Sub-corpus Entity-Relationship Diagram.

APPENDIX C : AUTHOR-TOPIC TOTAL DOCUMENTS COUNT MATRIX

AUTHORS

TOPICS	A100024	A100078 A105328	A111554	A111915 A104872	A100046	A100042	A113159	A102480	A100512	A111487	A100023	A101068	A100006	A111661	A111723	TOTALS
T50014	3	4	0	4	0	1	0	0	0	3	1	354	18	1	1	390
T50031	1	1149	0	0	0	0	0	0	0	0	0	0	1	0	783	1934
T50013	0	0	491	0	12	55	1022	729	560	0	0	0	0	0	0	2869
T50128	26	509	0	0	0	0	0	0	0	0	145	1	842	0	1	1524
T50012	0	0	6	0	21	867	135	13	2	0	0	0	0	0	0	1044
T50048	6	1602	0	1	0	0	0	0	0	2	752	539	5	0	0	2907
T50015	0	1	0	0	0	0	0	0	0	764	1	0	0	1	0	767
T50097	0	0	179	0	25	10	3	6	2	0	0	0	0	0	0	225
T50050	1536	6	0	0	0	0	0	0	0	0	0	0	1	0	0	1543
T50006	9	6	0	12	0	0	0	0	0	0	3	0	2	1	2	35
T50115	0	0	781	0	780	19	0	357	0	0	0	0	0	0	0	1937
T50136	0	0	0	0	0	0	0	0	0	0	0	0	0	394	0	394
T50187	0	0	0	290	0	0	0	0	0	0	0	0	0	0	0	290
T51556	0	16	0	1	0	0	0	0	0	5	0	0	0	0	33	55
T50172	0	0	0	1487	0	0	0	0	0	0	0	0	0	0	0	1487
T50383	0	0	4	0	157	5	0	0	0	0	0	0	0	0	0	166
T50368	0	0	6	0	0	155	0	1	0	0	0	0	0	0	0	162
T50273	0	0	25	0	33	17	0	0	490	0	0	0	0	0	0	565
T50222	0	0	0	0	0	0	0	0	0	121	0	0	0	0	0	121
T50338	0	0	1	0	154	0	0	1	0	0	0	0	0	0	0	156
T50049	1	0	0	63	0	0	0	0	0	0	0	0	0	0	0	64
T50214	0	0	0	0	0	0	0	0	0	0	0	0	0	163	0	163
T50077	0	0	0	0	0	0	0	0	0	0	0	0	0	64	0	64
TOTALS	1582	3293	1493	1858	1182	1129	1160	1107	1054	895	902	894	869	624	820	18862

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D: MEGAM—CG AND LMBFGS TRAINING ALGORITHMS

The following section was derived from the unpublished notes on the conjugate gradient and limited memory BFGS optimization of logistic regression written by Hal Daume III[3].

The conjugate gradient method employs an iterative process to obtain the numerical solution to a system of linear equations with a matrix that is Hermitian and symmetric. The premise behind conjugate gradient is to choose the search direction for any given iteration of the iterative process based on it's orthogonality to the search direction of the previous iteration [32]. For example, given any arbitrary direction $\boldsymbol{u}$, the vector $\boldsymbol{w}$ is updated as follows [32]:

$$\boldsymbol{w}' \leftarrow \boldsymbol{w} + \frac{\boldsymbol{g}^{\mathrm{T}}\boldsymbol{u}}{\lambda \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} + \sum_n \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n)\sigma(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n)(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_n)^2}\boldsymbol{u},$$

where σ is the logistic function and $\sigma(a) = (1 + \exp - a)^{-1}$ [32]. The gradient is given by [32]:

$$\boldsymbol{g} = -\lambda \boldsymbol{w} + \sum_n \sigma(-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n) y_n \boldsymbol{x}_n$$

The vector, $\boldsymbol{u}$, is chosen according to $\boldsymbol{u}' \leftarrow \boldsymbol{g} - \beta \boldsymbol{u}$ where β is given by the Hestenes-Stiefel formula [32]:

$$\beta = \frac{\boldsymbol{g}'^{\mathrm{T}}(\boldsymbol{g}' - \boldsymbol{g})}{\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{g}' - \boldsymbol{g})}.$$

The full training algorithm for conjugate gradient ascent implemented in the MEGAM optimization package is outlined in Figure 7.

[3] Available for download from the University of Utah School of Computing website: http://www.cs.utah.edu/~hal/megam/index.html.

Algorithm CG($\boldsymbol{x}, \boldsymbol{y}, \lambda$)
Initialize $\boldsymbol{w} \leftarrow \langle 0 \rangle_F, \mathbf{wtx} \leftarrow \langle 0 \rangle_N, \boldsymbol{g} \leftarrow \langle 0 \rangle_F, \boldsymbol{u} \leftarrow \langle 0 \rangle_F$
while not converged **do**
  $\boldsymbol{g}' \leftarrow -\lambda \boldsymbol{w}$
  for $n = 1 \ldots N$ **do**
    $\boldsymbol{g}' \leftarrow \boldsymbol{g} + \sigma(-y_n \text{wtx}[n]) y_n \boldsymbol{x}_n$
  end for
  $\beta \leftarrow (\boldsymbol{g}'^{\mathsf{T}}(\boldsymbol{g}' - \boldsymbol{g})) / (u^{\mathsf{T}}(\boldsymbol{g}' - \boldsymbol{g}))$
  $u \leftarrow \boldsymbol{g} - \beta u$
  $z \leftarrow (\boldsymbol{g}'^{\mathsf{T}} u) / (\lambda \boldsymbol{u}^{\mathsf{T}} \boldsymbol{u} + \sum_n \sigma(\text{wtx}[n]) \sigma(-\text{wtx}[n]) (\boldsymbol{u}^{\mathsf{T}} \boldsymbol{x}_n)^2)$
  $\boldsymbol{w} \leftarrow \boldsymbol{w} + z\boldsymbol{u}$
  for $n = 1 \ldots N$ **do**
    $\text{wtx}[n]) \leftarrow \text{wtx}[n] + z u^{\mathsf{T}} x_n$
  end for
  $\boldsymbol{g} \leftarrow \boldsymbol{g}'$
end while
return $\boldsymbol{w}$

Figure 7. The full training algorithm for conjugate gradient ascent [From 32].

BFGS (Broyden-Fletcher-Goldarb-Shannon) is derived from Newton's method in optimization and is used to solve nonlinear optimization problems with no constraints. In LM-BFGS, the iterative steps of the optimization process are computed in a reasonable amount of time while using only a limited amount of memory [32].

Since it is impossible to construct and invert the Hessian matrix for multiclass problems as it is done in binary class problems, an alternative presented in the LM-BFGS method is to iteratively build an approximation of the true Hessian [32]. That is, the Hessian at the ith iteration is approximated using the previous M weight vector and gradient values [32]. The full training algorithm for limited memory BFGS (less the three subroutines: ComputeGradient, ComputePosterior, and LineSearch) implemented in the MEGAM Optimization Package is presented in Figure 8.

Algorithm LM-BFGS($\boldsymbol{x}, \boldsymbol{y}, \lambda$)

Initialize $\boldsymbol{w} \leftarrow \langle 0 \rangle_F$, $\text{wtx} \leftarrow \langle 0 \rangle_{NxC}$

$\boldsymbol{g} \leftarrow$ COMPUTEGRADIENT($\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}, \text{wtx}$)

$\boldsymbol{q} \leftarrow \boldsymbol{g} / \sqrt{\boldsymbol{g}^{\mathsf{T}} \boldsymbol{g}}$

$\text{qtx} \leftarrow \boldsymbol{q}^{\mathsf{T}} \boldsymbol{x}$

$\eta \leftarrow$ LINESEARCH($\lambda, \boldsymbol{y}, \text{wtx}, \boldsymbol{w}, \boldsymbol{q}, \boldsymbol{g}, \text{qtx}$)

for $n = 1 \ldots N$, $c = 1 \ldots C$ **do**

$\text{wtx}[n, c] \leftarrow \text{wtx}[n, c] + \eta \text{qtx}[n, c]^{\mathsf{T}} \boldsymbol{x}_{nc}$

end for

$\boldsymbol{w}' \leftarrow \boldsymbol{w} + \eta \boldsymbol{q}$

Initialize $\text{mem} \leftarrow 0$

while not converged **do**

$\boldsymbol{g}' \leftarrow$ COMPUTEGRADIENT($\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}, \text{wtx}$)

$\alpha \leftarrow (\boldsymbol{g}' - \boldsymbol{g})^{\mathsf{T}} (\boldsymbol{w}' - \boldsymbol{w})$

$\sigma \leftarrow (\boldsymbol{g}' - \boldsymbol{g})^{\mathsf{T}} (\boldsymbol{g}' - \boldsymbol{g})$

Push $d = (\boldsymbol{w}' - \boldsymbol{w}), \boldsymbol{u} = (\boldsymbol{g}' - \boldsymbol{g})$ and α onto mem

$\boldsymbol{q} \leftarrow \boldsymbol{g}'$

$\beta \leftarrow \langle 0 \rangle_M$

for $m = M, \ldots, 1$ **do**

$\beta[m] \leftarrow \left(\text{mem}_d[m]\right) / \left(mem_\alpha[m]\right)$

$\boldsymbol{q} \leftarrow \boldsymbol{q} - \beta[m] \left(\text{mem}_\alpha[m]\right)$

end for

$\boldsymbol{q} \leftarrow \sigma \boldsymbol{q}$

for $m = 1, \ldots, M$ **do**

$\xi \leftarrow \left(\text{mem}_u[m]\right)^{\mathsf{T}} \boldsymbol{q}$

for $f = 1, \ldots, M$ **do**

$\xi \leftarrow \left(\text{mem}_d[m, f]\right)\left(\beta[m] - \varsigma / \left(\text{mem}_a[m]\right)\right)$

$\boldsymbol{q}[f] \leftarrow \boldsymbol{q}[f] + \xi$

$\varsigma \leftarrow \varsigma + \xi$

end for

end for

$\boldsymbol{q} \leftarrow -\boldsymbol{q}$

$\text{qtx} \leftarrow \boldsymbol{q}^{\mathsf{T}} \boldsymbol{x}$

$\eta \leftarrow$ LINESEARCH $\left(\lambda, \boldsymbol{y}, \text{wtx}, \boldsymbol{w}, \boldsymbol{q}, \boldsymbol{g}, \text{qtx}\right)$

for $n = 1...N, c = 1...C$ **do**

$\text{wtx}[n,c] \leftarrow \text{wtx}[n,c] + \eta \text{qtx}[n,c]^{\mathsf{T}} \boldsymbol{x}_{nc}$

end for

$\boldsymbol{w}' \leftarrow \boldsymbol{w} + \eta \boldsymbol{q}$

$\boldsymbol{q} \leftarrow \boldsymbol{g}'$

end while

return $\boldsymbol{w}$

Figure 8. The full training algorithm for limited memory BFGS [From 32].

APPENDIX E: BINARY DATA SET TOPIC CATEGORY STATS AND RESULTS

			Total Documents		Total Vocab Counts		Classification Results										
Fold	topicID	Topic	Test	Train	Train Vocab	Test Vocab	TP	TN	FN	FP	Correct	Incorrect	Precision	Recall	F-score	Accuracy	Balanced Accuracy
1	T50031	Music	501	2,499	62578	20169	1	486	0	14	487	14	0.0666	1.0000	0.1250	0.9721	0.9860
2	T50048	Motion Pictures	500	2,500	56801	29691	5	492	1	2	497	3	0.7142	0.8333	0.7692	0.9940	0.9146
3	T50050	Dancing	1,473	1,527	49748	38358	809	6	658	0	815	658	1.0000	0.5514	0.7108	0.5533	0.7757
4	T50128	Theater	526	2,474	57901	30114	24	486	2	14	510	16	0.6315	0.9230	0.7500	0.9696	0.9475
Average													0.6031	0.8269	0.5888	0.8722	0.9060
Std.Dev													0.3909	0.1959	0.3101	0.2129	0.0916

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX F: MULTI-CATEGORY DATA SET TOPIC CATEGORY STATS AND RESULTS

In the multi-category data set, a leave-one-topic-out 23-fold cross-validation as described in Chapter IV Section A Sub-paragraph 2 was used to compute the accuracy and balanced accuracy results.

			Total Documents		Total Vocab Counts		Classification Results			
Fold	topicID	Topic	Test	Train	Test	Train	Incorrect	Correct	Accuracy	Bal.Accuracy
1	T50014	Books and Literature	390	18,472	30,974	163,816	383	7	0.0179	0.2424
2	T50031	Music	1,934	16,928	41,289	159,156	1499	435	0.2249	0.1148
3	T50013	Baseball	2,869	15,993	31,492	162,228	2376	493	0.1718	0.0777
4	T50128	Theater	1,524	17,338	56,116	153,710	1007	517	0.3392	0.1410
5	T50012	Football	1,044	17,818	21,668	164,601	1038	6	0.0057	0.0667
6	T50048	Motion Pictures	2,907	15,955	78,651	139,519	1867	1040	0.3578	0.2229
7	T50015	Art	767	18,095	30,558	161,371	757	10	0.0130	0.2006
8	T50097	Basketball	225	18,637	9,118	166,816	44	181	0.8044	0.0800
9	T50050	Dancing	1,543	17,319	39,294	158,128	1539	4	0.0026	0.0444
10	T50006	Television	35	18,827	5,514	167,609	12	23	0.6571	0.2074
11	T50115	Hockey, Ice	1,937	16,925	24,259	162,863	1164	773	0.3991	0.0763
12	T50136	Restaurants	394	18,468	15,404	164,613	389	5	0.0127	0.0008
13	T50187	Appointments and Executive Changes	290	18,572	5,791	167,359	1	289	0.9966	0.0664
14	T51556	Deaths (Obituaries)	55	18,807	6,103	167,418	38	17	0.3091	0.1247
15	T50172	Advertising and Marketing	1,487	17,375	20,867	164,218	140	1347	0.9059	0.0604
16	T50383	Golf	166	18,696	6,745	166,804	160	6	0.0361	0.0933
17	T50368	Boxing	162	18,700	8,955	167,129	155	7	0.0432	0.0671
18	T50273	Horse Racing	565	18,297	14,800	165,648	545	20	0.0354	0.0533
19	T50222	Photography	121	18,741	7,376	167,050	117	4	0.0331	0.0022
20	T50338	Soccer	156	18,706	7,816	166,827	155	1	0.0064	0.0667
21	T50049	Suspensions, Dismissals and Resignations	64	18,798	2,860	167,683	0	64	1.0000	0.1333
22	T50214	Cooking and Cookbooks	163	18,699	6,002	166,926	129	34	0.2086	0.0139
23	T50077	Food	64	18,798	4,503	167,393	32	32	0.5000	0.0333
Average									0.3079	0.0952
StdDev									0.3431	0.0688

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] O. de Vel, A. Anderson, M. Corney and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD Rec.,* vol. 30, pp. 55–64, 2001.

[2] M. W. Corney, "Analysing E-mail Text Authorship for Forensic Purposes," March 2008.

[3] R. Zheng, J. Li, H. Chen and Z. Huang, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," *J. Am. Soc. Inf. Sci. Technol.,* vol. 57, p. 378, 2006.

[4] G. T. Gehrke, "Authorship Discovery in Blogs Using Bayesian Classification with Corrective Scaling," June 2008.

[5] G. U. Yule, "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship," *Biometrika,* vol. 30, pp. 363–390, 1939.

[6] C. Mascol, "Curves of Pauline and Pseudo-Pauline Style I," *The Unitarian Review,* vol. 30, pp. 452–460, 1888.

[7] W. Fucks, "On Mathematical Analysis of Style," *Biometrika,* vol. 39, pp. 122–129, 1952.

[8] F. Mosteller and D. L. Wallace, "Inference in an authorship problem—A comparative study of discrimination methods applied to authorship of disputed federalist papers," *Journal of the American Statistical Association,* vol. 58, pp. 275, 1963.

[9] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *J. Am. Soc. Inf. Sci. Technol.,* vol. 60, pp. 538–556, 2009.

[10] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing.* Cambridge, Mass.: MIT Press, 1999.

[11] G. Mikros and E. K. Argiri, "Investigating topic influence in authorship attribution," in *Proceedings of the SIGIR '07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007, Amsterdam, Netherlands, July 27, 2007.*

[12] M. Koppel, J. Schler and E. Bonchek-Dokow, "Measuring Differentiability: Unmasking Pseudonymous Authors," *Journal of Machine Learning Research: JMLR.,* vol. 8, pp. 1261–1276, 2008.

[13] D. Madigan, A. Genkin, D. D. Lewis, E. G. D. D. Lewis, S. Argamon, D. Fradkin, L. Ye and D. D. L. Consulting, "Author identification on the large scale," in *In Proc. of the Meeting of the Classification Society of North America,* 2005.

[14] H. Baayen, H. van Halteren and F. Tweedie, "Outside the cave of shadows: using syntactic annotation to enhance authorship attribution," *Lit Linguist Computing,* vol. 11, pp. 121–132, September 1. 1996.

[15] A. Ratnaparkhi, "A simple introduction to maximum entropy models for natural language processing," Institute for Research in Cognitive Science, University of Pennsylvania, Tech. Rep. IRCS-97-08, 1997.

[16] T. Brants and W. Skut, "Automation of Treebank Annotation," 1998.

[17] J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," in *Proceedings of the Fifth Conference on Applied Natural Language Processing,* 1997, pp. 16–19.

[18] A. M. Harc and A. Mikheev, "Feature lattices for maximum entropy modeling," in *In Proc. of ACL-COLING,* 1998, pp. 848–854.

[19] A. Ratnaparkhi, J. Reynar and S. Roukos, "A maximum entropy model for prepositional phrase attachment," in *HLT '94: Proceedings of the Workshop on Human Language Technology,* 1994, pp. 250–255.

[20] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *University of Pennsylvania Empirical Methods in Natural Language Processing,* pp. 133–142.

[21] D. Beeferman, A. Berger and J. Lafferty, "Statistical Models for Text Segmentation," *Machine Learning,* vol. 34, pp. 177–210, 1999.

[22] V. D. Pietra, V. D. Pietra and J. Lafferty, "Inducing Features of Random Fields," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 19, pp. 380–393, 1997.

[23] R. Lau, R. Rosenfeld and S. Roukos, "Adaptive language modeling using the maximum entropy principle," in *HLT '93: Proceedings of the Workshop on Human Language Technology,* 1993, pp. 108–113.

[24] K. Nigam, J. Lafferty and A. McCallum, "Using Maximum Entropy for Text Classification."

[25] P. Adams, "Conversation Thread Extraction and Topic Detection in Text-Based Chat," 2008.

[26] A. McCallum and D. Freitag, "Maximum entropy Markov models for information extraction and segmentation," in 2000, pp. 591–598.

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

[28] A. L. Berger, V. J. D. Pietra and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computer Linguistics.,* vol. 22, pp. 39–71, 1996.

[29] R. D. De Veaux, P. F. Velleman and D. E. Bock, *Stats : Data and Models.* 2nd ed. Boston: Pearson, Addison-Wesley, 2008, p. 69.

[30] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," May 21, 1996.

[31] E. Sandhaus, "The New York Times Annotated Corpus Overview."

[32] H. Duame III. Notes on CG and LM-BFGS optimization of logistic regression [accessed September 11, 2009]. Available: http://www.cs.utah/~has/megam/index.html.

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