AUTHOR ATTRIBUTION EVALUATION WITH NOVEL TOPIC CROSS-VALIDATION

Andrew I. Schein, Johnnie F. Caver, Randale J. Honaker, and Craig H. Martell
Department of Computer Science
The Naval Postgraduate School
1 University Way
Monterey, CA 93943
{*aischein,jfcaver,rjhonaker,cmartell*}*@nps.edu*

Keywords: author attribution, novel topic, cross validation, genre shift.

Abstract: The practice of using statistical models in predicting authorship (so-called author attribution models) is long established. Several recent authorship attribution studies have indicated that topic-specific cues impact author attribution machine learning models. The arrival of new topics should be anticipated rather than ignored in an author attribution evaluation methodology; a model that relies heavily on topic cues will be problematic in deployment settings where novel topics are common. We develop a protocol and test bed for measuring sensitivity to topic cues using a methodology called *novel topic cross-validation*. Our methodology performs a cross-validation where only topics unseen in training data are used in the test portion. Analysis of the testing framework suggests that corpora with large numbers of topics lead to more powerful hypothesis testing in novel topic evaluation studies. In order to implement the evaluation metric, we developed two subsets of the New York Times Annotated Corpus including one with 15 authors and 23 topics. We evaluated a maximum entropy classifier in standard and novel topic cross validation in order to compare the mechanics of the two procedures. Our novel topic evaluation framework supports automatic learning of stylometric cues that are topic neutral, and our test bed is reproducible using document identifiers available from the authors.

1 INTRODUCTION

Authorship attribution researchers build machine learning classification models or rule-based systems identifying the author of an anonymous text given undisputed knowledge of various communications written by that particular author. The earliest (as well as continuing) efforts in the field looked at the authorship of historically interesting documents. Today, interest in the field is additionally motivated by fairness and public welfare concerns: plagiarism detection and identifying authors in a criminal investigation or intelligence setting.

Several authorship attribution studies have speculated about the existence of a link between topic cues and author style features (Mikros and Argiri, 2007; Corney, 2003; Koppel et al., 2008; Madigan et al., 2005; Gehrke, 2008). We present a novel experimental protocol for measuring author attribution performance in a setting where new topics are expected to appear over time as a result of the changing statistical distribution of discussed topics. Our technique, called novel topic cross-validation, 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 result is a method for determining stylometric cues and features that remain relevant in spite of topic changes.

Novel topic cross-validation simulates a scenario where we are trying to perform the author attribution task when novel topics appear. Scenario-based motivations justify the procedure and its deviation from independent and identical distribution assumptions that typically surround the evaluation of machine learning classifiers. We can imagine being part of an organization deciding which of several competing decision rules to deploy in a system. How well can we expect these systems to perform as they confront instances with novel topics? How often must we re-train our methods to ensure performance holds up as new topics appear? Novel topic cross-validation represents an important metric to answer these questions. The scenario motivating novel topic cross-validation is analogous to many existing highly profitable modeling problems: deciding the appropriate bid for a new keyword offering at an Internet advertisement auction, or deciding how to recommend a new item to an online store (the so-called "cold-start" problem (Schein et al., 2002)). Organizations who engage in these practices develop a variety of scenario-based simulations to build their decision frameworks.

Our work developing an evaluation method that can separate topic cue influence in a classifier is additionally motivated by a desire to build robust models using stylometric features. Stylometric features are those that capture author specific word and grammar choices. Author style may vary depending on the topic, and quantifying this phenomenon is highly desirable. Many aspects of author style are likely to be relatively topic neutral, and understanding these features of style would be highly beneficial to the author attribution community. On the other hand, learning features that are topic-specific but have little to do with style should be less important to the author attribution researcher; when deploying the model we are likely to discover both new authors writing on the same topics as a target author as well as new topics that our target authors may write about. These two categories of novel entities (authors and topics) will damage performance of a method relying strictly on topic cues.

We quickly realized that existing metrics and data resources are not well developed to implement a research program that attempts to isolate stylometric cues from topic. Our development of an evaluation methodology and test bed allows the community as a whole to understand, measure, and isolate topic influence in author attribution studies. For this reason, we are making the author, topic and document IDs from our test bed available on a web site[1]. The original unfiltered data is the New York Times Annotated Corpus (Sandhaus, 2008) which is distributed by the Linguistic Data Consortium. Thus, this work introduces the methodology of novel topic cross-validation, as well as a testbed for implementing the procedure. With techniques for measurement established, researchers may begin to tackle this problem in a scientific fashion.

Using the New York Times Annotated Corpus, we generated two sub-corpora of data with differing characteristics: one consisting of 3,000 documents cross-tabulated with 2 authors and 4 topics (the binary data set), and the other consisting of 18,862 documents cross-tabulated with 15 authors and 23 topics (the multi-category data set). From these separate sub-corpora, we perform a novel topic cross-validation comparing the results with a standard cross-validation. Our data set differs from previous test beds used to model topic/author influence in scope, balance, and classification; previous studies were limited to three or fewer topics or authors, using equally balanced data sets[2] and binary classifications. The document count of previous studies was frequently limited to several hundred. Having a larger set of documents, topics and authors combined with our innovative approach to controlling topic 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. Moreover, our analysis demonstrates that having a larger number of topics (with documents distributed as evenly as possible among them) has important and beneficial ramifications for hypothesis testing in a novel topic evaluation.

2 RELATED WORK

Recent review articles describe the state of the art in author attribution algorithms, similarity statistics, feature sets, and evaluation methods (Stamatatos, 2009; Malyutov, 2006). Since our own work focuses primarily on the influence of topic and evaluation methodology, we focus our review on these areas. When examining the previous work cited below, take note of the small number of topics and documents used; the comparison of these numbers to those found in our own data set will be highly relevant to our conclusions.

2.1 Studies of Topic Influence

Several previous efforts have made an attempt to quantify a relationship between topic and author. The first study, conducted by Mikros and Argiri, 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 (Mikros and Argiri, 2007). They tested the impact of topic on authorship attribution using the following stylometric features:

- Vocabulary richness
- Sentence length

[1] `http://faculty.nps.edu/cmartell/NPSCrossValidationSet/nps_nyt_novel_topic_cv.tar.gz`

[2] A balanced data set is one where category labels are equally represented. In this case, the balance refers to the number of articles written by each author.

- 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. The data set was completely balanced, with each author writing 100 articles, half of which were written about one of two topics. 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.

The second study, conducted by Koppel, Schler, and Bonchek-Dokow, explored the depth of difference between topic variability in authorship attribution using an unmasking technique (Koppel et al., 2008). 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. 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.

The third study, conducted by Corney (Corney, 2003), 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 156 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. 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.

In contrast to results obtained by Corney (Corney, 2003), the fourth study, by Madigan *et al.* (Madigan et al., 2005), tested the effect of topic on authorship attribution with 59 Usenet postings by two distinct authors and three distinct topics. Just as with Corney, 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 a unigram model; however, their bi-gram parts-of-speech model proved to be one of the best among the tested possibilities.

Finally, the fifth study, conducted by Baayen et al. (Baayen et al., 1996), used principal components analysis (PCA) and linear discriminant analysis (LDA) to evaluate the effectiveness of grouping text by author, using stylometric features. Their data set consisted of 576 documents written by eight students. Each student wrote a total of 24 documents in three different genres about three different topics. They found that compensating for the topic imbalance coverage led to increased performance in a cross-validation.

The recent review by Stamatatos (Stamatatos, 2009) points to a small number of additional similar studies. A key difference between our data set and those used by previous researchers is size. The number of observations in our multi-category data set is much larger than any of the previous examples. In addition, our multi-category data set has many more topics, which we will later argue is advantageous in a novel topic evaluation. The nature of our evaluation is also a bit different in that it simulates what happens when a author attribution classifier encounters a new topic. Many of the previous studies are "in-sample" analysis or examine other questions pertaining to topics.

2.2 Evaluation Methodology

Typical evaluations of author attribution divide a corpus into a train/test split. In some cases standardized train/test splits have been developed for reproducibility (Stamatatos et al., 2000). When developing an evaluation, typically researchers have attempted to control for factors that can influence outcome. In addition to topic (the focus of the current work), age, sex, or other attributes of the author may have predictive abilities that need to be controlled. In our opinion, within the literature a consensus has formed that an evaluation will ideally have a balanced number of documents per author in a test set; this greatly simplifies the interpretability of a test set accuracy. In practice, requiring data set balance limits the qualified data sets available to the author attribution researcher. In particular, it is challenging to locate a data set with many authors writing very many documents if we require these authors to write on the same topics and with the same frequency.

In his recent review of the author attribution methods, Stamatatos comments on evaluation: "Ideally, all the texts of the training corpus should be on exactly the same topic for all the candidate authors." (Stamatatos, 2009). This advice is important in aspects of algorithm evaluation. However, we see the field of author attribution progressing by embracing topic and social distinctions as a source of complexity with scientifically (and functionally) interesting consequences. We believe there is an important place for evaluation methodologies that focus on exploring factors that have real consequences for building deployable systems rather than neutralizing them for algorithm evaluation purposes.

3 DATA SET PREPARATION

The New York Times 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 (Sandhaus, 2008). Each XML document contains 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 contain tags for the topic, only 48.18% of the documents contain tags for the author. Therefore, we filtered the data to a subset of $871,050$ documents that are tagged with author, topic, and title.

From this corpus, we selected documents with a *single topic* and a *single author*. After our filtering steps on the NYT Annotated Corpus, we were left with a subcorpus with this property. 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 (e.g. the two authors wrote the same number of documents) and unbalanced across four topics. The multi-category data set was unbalanced across 15 authors and unbalanced across 23 topics. Due to the independent procedures used to generate the two subsets, the binary data set is not a proper subset of the multi-category data set.

3.1 Binary Data Set

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 1 is a list of authors along with the corresponding total number of documents in the subset written by each author. The average vocabulary size over all documents was 282.57 with a minimum vocabulary size of 2 and a maximum of $1,304$.

3.2 Multi-Category Data Set

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 2 lists the topic categories along with their corresponding topic identifications. Table 3 shows how the counts are distributed amongst authors and topics. 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 $2,889$.

4 EXPERIMENTAL DESIGN

4.1 Feature Extraction

We extracted the article text from each of the documents to a separate file for processing. Certain authors were duplicated in the source data with differing IDs: Stephen Holding and Stuart Elliott. These author IDs were merged in the experiments that follow. 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 reduce the dimensionality of the feature space. Finally, to facilitate use of unigram word features, data was processed into word grams by tokenizing words on whitespace.

The regular expression that extracted text from the file did not always capture the lead paragraph; we discovered that 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.

4.2 Scenario 1: 10-Fold Cross-Validation

Our baseline testing scenario is a 10-fold cross-validation of the sort that is usually applied in author attribution as well as other machine learning tasks.

Table 1: Author/Topic Data Tabulation.

Author ID	Author	Author Total Docs	Topic ID	Topic	Topic Total Docs
A100024	Dunning, Jennifer	1500	T50031	Music	1
			T50048	Motion Pictures	6
			T50050	Dancing	1,467
			T50128	Theatre	26
A100078 and A105328	Holden, Stephen	1500	T50031	Music	500
			T50048	Motion Pictures	494
			T50050	Dancing	6
			T50128	Theatre	500

Table 2: Multi-Category Topic Categories.

T50014	Books and Literature	T50187	Appointments and Executive Changes
T50031	Music	T51556	Deaths (Obituaries)
T50013	Baseball	T50172	Advertising and Marketing
T50128	Theatre	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		

A maximum entropy classifier (Berger et al., 1996; Daumé III, 2004) was trained on 90% of the documents, and then tested on the remaining 10% for each fold using a binary classification for the data set with two authors and using a multi-category 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. The 10-fold cross-validation provides an example of a typical testing framework for an author attribution evaluation, which we will use to contrast our new procedure.

4.3 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 maximum entropy 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 4 topics in the binary data set, and a total of 23 topics in the multi-category data set.

4.4 Performance Measures

Accuracy, precision, recall, and F-score were the performance measures used to evaluate the results of the experiments. Precision, recall and F-score are standard evaluation metrics in the natural language processing community (Manning and Schütze, 1999). All metrics were computed for each cross-validation fold of the binary data set. Only accuracy was computed for the multi-category data set since the other metrics are designed for binary classification. Table 4 depicts the confusion matrix used to compute the precision, recall, and F-score for the two authors in the binary data set.

Whether performing a standard or novel topic cross-validation, the *standard error* of the accuracy depends on the sample size. The size of the sample in a novel topic cross-validation is equal to the number of topics rather than a tunable parameter (the "n" in an n-fold cross-validation). It follows from the standard

Table 3: Topic/Author Data Tabulation for the Multi-Category Data Set.

					AUTHORS					
		A100024	**A100078** **A105328**	**A111554**	**A111915** **A104872**	**A100046**	**A100042**	**A113159**	**A102480**	**. . .**
TOPICS	**T50014**	3	4	0	4	0	1	0	0	
	T50031	1	1149	0	0	0	0	0	0	
	T50013	0	0	491	0	12	55	1022	729	
	T50128	26	509	0	0	0	0	0	0	
	T50012	0	0	6	0	21	867	135	13	
	T50048	6	1602	0	1	0	0	0	0	
	T50015	0	1	0	0	0	0	0	0	
	T50097	0	0	179	0	25	10	3	6	
	T50050	1536	6	0	0	0	0	0	0	
	T50006	9	6	0	12	0	0	0	0	
	T50115	0	0	781	0	780	19	0	357	
	T50136	0	0	0	0	0	0	0	0	
	T50187	0	0	0	290	0	0	0	0	
	T51556	0	16	0	1	0	0	0	0	
	T50172	0	0	0	1487	0	0	0	0	
	T50383	0	0	4	0	157	5	0	0	
	T50368	0	0	6	0	0	155	0	1	
	T50273	0	0	25	0	33	17	0	0	
	T50222	0	0	0	0	0	0	0	0	
	T50338	0	0	1	0	154	0	0	1	
	T50049	1	0	0	63	0	0	0	0	
	T50214	0	0	0	0	0	0	0	0	
	T50077	0	0	0	0	0	0	0	0	
	TOTALS	**1582**	**3293**	**1493**	**1858**	**1182**	**1129**	**1160**	**1107**	

	A100512	**A111487**	**A100023**	**A101068**	**A100006**	**A111661**	**A111723**	**TOTALS**
T50014	0	3	1	354	18	1	1	**390**
T50031	0	0	0	0	1	0	783	**1934**
T50013	560	0	0	0	0	0	0	**2869**
T50128	0	0	145	1	842	0	1	**1524**
T50012	2	0	0	0	0	0	0	**1044**
T50048	0	2	752	539	5	0	0	**2907**
T50015	0	764	1	0	0	1	0	**767**
T50097	2	0	0	0	0	0	0	**225**
T50050	0	0	0	0	1	0	0	**1543**
T50006	0	0	3	0	2	1	2	**35**
T50115	0	0	0	0	0	0	0	**1937**
T50136	0	0	0	0	0	394	0	**394**
T50187	0	0	0	0	0	0	0	**290**
T51556	0	5	0	0	0	0	33	**55**
T50172	0	0	0	0	0	0	0	**1487**
T50383	0	0	0	0	0	0	0	**166**
T50368	0	0	0	0	0	0	0	**162**
T50273	490	0	0	0	0	0	0	**565**
T50222	0	121	0	0	0	0	0	**121**
T50338	0	0	0	0	0	0	0	**156**
T50049	0	0	0	0	0	0	0	**64**
T50214	0	0	0	0	0	163	0	**163**
T50077	0	0	0	0	0	64	0	**64**
TOTALS	**1054**	**895**	**902**	**894**	**869**	**624**	**820**	**18862**

Table 4: Matrix Used to Construct Precision, Recall, and F-score.

Prediction	Ground Truth	
	A100024	A100078 A105328
A100024	True Positive	False Positive
A100078/A105328	False Negative	True Negative

error computation,

$$SE = \hat{\sigma}/\sqrt{n}, \qquad (1)$$

having a greater number of topics is conducive to low-error estimates of performance in a novel topic cross-validation. In ordinary *n*-cross validation, the analyst chooses *n* rather than the data set.

In a standard cross validation, each fold of the cross-validation has equal sizes (modulo a small factor for unequal division), and one can compute average and standard deviations for the results of the cross-validation naively. Within a novel topic cross-validation, each division of data will have different sizes since generally there will be different numbers of documents in each topic category. For novel topic cross-validation the appropriate aggregation technique for the test statistics is to use a weighted average and standard deviation: the weights are computed as the fraction of the total document count represented within the cross-validation fold. A derivation of the unbiased variance estimate of the weighted average is provided in the appendix. Below we define the estimates of weighted mean $\hat{\mu}$ and variance $\hat{\sigma}^2$.

$$\hat{\mu} = \sum_{i=1}^{n} w_i x_i \qquad (2)$$

$$V_2 = \sum_{i=1}^{n} w_i^2 \qquad (3)$$

$$\hat{\sigma}^2 = \frac{1}{1-V_2} \sum_{i=1}^{n} w_i (x_u - \hat{\mu})^2. \qquad (4)$$

Here the x_i refer to an evaluation statistic such as an accuracy, precision, recall or F-measure. We see that having an equal proportion of documents across topics (or as close as possible) is beneficial in producing low variance estimates of performance. This close-to-equal proportion property is something we strove to accomplish in developing the document subsets for our experiments. See the appendix for further analysis of the variance term.

We report average performance scores for n-fold cross-validation and weighted averages for novel topic cross-validation. Statistical hypothesis testing comparisons of several algorithms or feature sets in novel topic cross-validation can be accomplished using weighted means and standard deviations. However, comparing performance scores between novel topic cross-validation and ordinary cross-validation is often ill-advised, particularly when topics are distributed unevenly among authors (as in our data sets). The novel topic cross-validation can lead to different author/document proportions when compared to the the training sets of a standard cross-validation, and this difference in training sets can lead to different numbers. Machine learning algorithm performance is known to vary with category imbalance (see examples in (Japkowicz, 2000)). Comparing multiple algorithms in the *same* testing scenario (e.g. novel topic cross-validation) does not suffer from this methodological flaw.

5 RESULTS

Table 5 shows the performance of the maximum entropy classifier in a standard 10-fold cross-validation and a novel topic cross-validation for the binary data set. Table 6 shows accuracies on the multi-category data set. The results are recorded as averages (over the cross-validation folds) as well as the corresponding standard deviations. We report different performance metrics between binary and multi-category data sets for the reasons outlined in the Performance Measures section. Table 7 shows the topics that had the highest and lowest accuracies in the novel topic cross-validation of the multi-category data set.

We observe that the performance metrics in a novel topic cross-validation are lower than in a standard cross-validation, accompanied in all cases by an increase in the standard deviation. Examination of the data shows that novel topic cross-validation leads to a much broader range of accuracies (or other metrics) when compared against a standard cross-validation.

For the binary data set, the standard errors of the performance measures in a novel topic cross-validation are quite large because the number of topics is only 4. For example, taking the accuracy standard deviation (0.3388) and transforming it into a standard error (Equation 1) gives: 0.1694. In contrast, the standard deviation of the multi-category accuracy is 0.1631, producing a standard error of 0.0340 after dividing by the square root of the number of topics.

Table 5: Accuracy, Precision, Recall and F-score Results for the Binary Data Set.

	10-Fold Cross-Validation		Novel Topic Cross-Validation (N=4)	
	Average	Std. Dev.	Average	Std. Dev.
Accuracy	0.9953	0.0048	0.6953	0.3388
Precision	0.9960	0.0046	0.5110	0.6037
Recall	0.9947	0.0107	0.9847	0.0254
F-Score	0.9953	0.0048	0.5072	0.5912

Table 6: Accuracy for the Multi-Category Data Set.

	10-Fold Cross-Validation		Novel Topic Cross-Validation (N=23)	
	Average	Std. Dev.	Average	Std. Dev.
Accuracy	0.9835	0.0029	0.7272	0.1631

Table 7: Topics with Highest/Lowest Accuracy in Novel Topic Cross-Validation.

Top 3 Accuracy Categories	
Topic	**Accuracy**
Suspensions, Dismissals, and Resignations	1.0000
Appointments and Executive Changes	1.000
Food	1.000
Bottom Three Accuracy Categories	
Topic	**Accuracy**
Theatre	0.4587
Baseball	0.5559
Horse Racing	0.5699

6 DISCUSSION

Our evaluation simulates an important use case for an author attribution model; predicting author identity as new topics are discovered. The scenario we have constructed is a reproducible evaluation for comparing different author attribution techniques on the novel topic problem. Ideally, we would like to have a data set where each author has written on each topic in equal numbers. If this were the case then each author would have the same amount of training data on each fold of the novel topic cross-validation and we would obtain much lower standard deviation in the performance measures. Finding such a corpus with many authors, topics, and documents is challenging, and remains a subject for future work. We have done the next best thing: taken steps to ensure that each author has at least a hundred examples in each train/test portion of the cross-validation.

We examined individual folds of the novel topic cross-validation to see which topics were particularly hard or easy to classify. The top three "easiest" and "hardest" topics are listed in Table 7. At first it appears that broader topics (such as *Suspensions, Dismissals, and Resignations*) are easiest, while very focused topics such as *baseball* were hard. However, those three "easy" topics are almost entirely written by a single author, and so it is likely that the particular author is easy to predict–independent of the topic. The ease of accurately classifying this particular author is embedded in the average accuracy of the standard 10-fold cross-validation, but does not stick out noticeably within individual folds the way it does with a novel topic cross-validation. Nonetheless, when we perform the document-weighted average of the performance metrics (*e.g.* accuracy) in the novel topic cross-validation, each document and topic has the same weight it would under an ordinary cross-validation regime.

In this paper we developed the novel topic scenario using two data sets in order to have one that is balanced in author writings (the binary data set) and another that is larger in documents, authors, and topics (the multi-category data set). In hindsight we can conclude that having a larger data set is more important for novel topic cross-validation than having perfect balance in author writings. The reason is that we would eventually like to perform machine learning method comparison with these data sets, and method comparison requires small standard errors in the metric averages. We have shown how the topic-related statistics impact standard error through the standard error calculation (Equation 1) as well as the distribution across topics (see Appendix). The results demonstrate that a hypothesis test for the superiority of a method over the maximum entropy baseline would be difficult if not impossible for the binary data set. For example, a 20% increase in accuracy in the binary data set bringing accuracy from 0.69 to 0.89 would not result in a rejection of the null hypothesis using a t-test[3].

However, for the multi-category data set the number of folds is 23, and this provides opportunity for a much more statistically powerful hypothesis test. For example, a 7% increase in accuracy will allow us to reject the null hypothesis that two models give equivalent performance (at P-value 0.05). From this line of reasoning we additionally conclude that our multi-category test bed is superior for novel topic cross-validation than the data sets described in the Related Work section (we included the number of topics in each of the data sets in part to make this point). The increase in number of authors and topics presents other advantages by providing a richer sampling of the complexity of variation that exists in actual usage.

7 CONCLUSIONS

We presented a new protocol for measuring the performance of author attribution techniques called novel topic cross-validation. The protocol is motivated by the needs of the analyst who must decide how best to deploy author attribution technology in a setting where novel topics appear. Those who deploy author attribution technologies in real world settings will want to measure the robustness of their decision rules to novel topics in order to pick among competing techniques and determining how often to retrain their methods. In order to perform a statistically powerful hypothesis test, our experiments and analysis lead us to recommend large data sets with many documents and topics, and with as close-to-equal document membership in topics as possible. We expect our protocol and data set will be valuable to those who attempt to build models that are robust to topic distribution changes.

ACKNOWLEDGMENTS

We would like to thank Constantine Perepelitsa for help tuning our evaluation scripts. Members of the Naval Postgraduate School Natural Language Processing Laboratory gave valuable feedback. We would like to thank the National Reconnaissance Office for funding that supported this work.

[3] The degrees of freedom of this test is 2 less than twice the number of topics

REFERENCES

Baayen, H., van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput.Linguist.*, 22(1):39–71.

Corney, M. W. (2003). Analysing E-mail Text Authorship for Forensic Purposes. Master's thesis, Queensland University of Technology.

Daumé III, H. (2004). Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M., and Rossi, F. (2003). *Gnu Scientific Library: Reference Manual*. Network Theory Ltd.

Gehrke, G. T. (2008). Authorship Discovery in Blogs Using Bayesian Classification with Corrective Scaling. Master's thesis, Naval Postgraduate School.

Gough, B. J. (2010). personal communication.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000)*, volume 1, pages 111–117.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2008). Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of machine learning research : JMLR.*, 8(1):1261–1276.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., and Ye, L. (2005). Author Identification on the Large Scale. In *Proc. of the Meeting of the Classification Society of North America*.

Malyutov, M. (2006). Authorship attribution of texts: A review. Ahlswede, Rudolf (ed.) et al., General theory of information transfer and combinatorics. Berlin: Springer. Lecture Notes in Computer Science 4123, 362-380 (2006).

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass. ID: 40848647.

Mikros, G. and Argiri, E. K. (2007). 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.

Sandhaus, E. (2008). The New York Times Annotated Corpus Overview. Linguistic Data Consortium, Philadelphia.

Schein, A., Popescul, A., Ungar, L., and Pennock, D. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 253–260.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Stamatatos, E., Kokkinakis, G., and Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Comput. Linguist.*, 26(4):471–495.

APPENDIX: THE UNBIASED ESTIMATE OF THE VARIANCE OF A WEIGHTED SUM

The formula for the unbiased estimate of the variance of a weighted sample is not as widely known as might be expected. We show its derivation below to help encourage the use of this valuable computation. The exposition that follows is based on the explanation by Gough used to justify an implementation within the GNU Scientific Library (Gough, 2010; Galassi et al., 2003).

Let x_i be a sample of i.i.d. random variables from a distribution with finite second moment and expected value μ and variance σ^2. A weight $w_i > 0$ is assigned to each of the samples, and we consider the weighted sum and variance defined below:

$$\begin{aligned} W &\doteq \sum_i w_i \\ \hat{\mu} &\doteq \frac{1}{W}\sum_i w_i x_i \\ \hat{\sigma}_b^2 &\doteq \frac{1}{W}\sum_i w_i (x_i - \hat{\mu})^2. \end{aligned}$$

The b subscript indicates the variance estimate above is biased, a fact we will demonstrate shortly. Through algebraic manipulation, we simplify $\hat{\sigma}_b^2$ into primitive terms:

$$\begin{aligned} \hat{\sigma}_b^2 &= \frac{1}{W}(\sum_i w_i x_i^2 - 2\sum_i w_i x_i \hat{\mu} + \sum_i w_i \hat{\mu}^2) \\ \hat{\sigma}_b^2 &= \frac{1}{W}(\sum_i w_i x_i^2 - 2\frac{\sum_i w_i x_i \sum_j w_j x_j}{W} + \\ &\qquad \frac{\sum_i w_i x_i \sum_j w_j x_j}{W}) \\ \hat{\sigma}_b^2 &= \frac{1}{W}(\sum_i w_i x_i^2 - \frac{1}{W}\sum_i w_i x_i \sum_j w_j x_j) \\ \hat{\sigma}_b^2 &= \frac{1}{W}(\sum_i w_i x_i^2 - \frac{1}{W}\sum_{i,j} w_i w_j x_i x_j). \end{aligned}$$

Since the random variables are iid with finite second moment, we have:

$$\begin{aligned} E[x_i x_j] &= \mu^2 + \delta_{ij}\sigma^2, \text{ where} \\ \delta_{ij} &= \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

Now consider the expectation of $\hat{\sigma}_b^2$:

$$\begin{aligned} E[\hat{\sigma}_b^2] &= \frac{1}{W}\left[\sum_i w_i E[x_i^2] - \frac{1}{W}\sum_{i,j} w_i w_j E[x_i x_j]\right] \\ &= \frac{1}{W}\left[W(\mu^2 + \sigma^2) - W\mu^2 - \frac{1}{W}\sum_i w_i^2 \sigma^2\right] \\ &= \sigma^2 \frac{W - \frac{1}{W}\sum_i w_i^2}{W}. \end{aligned}$$

The solution suggests we correct $\hat{\sigma}_b^2$ with the term:

$$\frac{W^2}{W^2 - \sum_i w_i^2}.$$

When $\sum_i w_i = 1$, the unbiased estimate matches the formula used within this paper. When $w_i = \frac{1}{n}$, the formula matches the $\frac{n}{n-1}$ term commonly used to correct the maximum likelihood estimate of variance (in the unweighted setting).