Cross-Topic Authorship Attribution: Will Out-Of-Topic Data Help?

Upendra Sapkota and **Thamar Solorio**
The University of Alabama at Birmingham
1300 University Boulevard
Birmingham, AL 35294, USA
{upendra,solorio}@cis.uab.edu

Manuel Montes-y-Gómez
Instituto Nacional de Astrofísica
Optica y Electrónica
Puebla, Mexico
mmontesg@ccc.inaoep.mx

Steven Bethard
The University of Alabama at Birmingham
1300 University Boulevard
Birmingham, AL 35294, USA
bethard@cis.uab.edu

Paolo Rosso
NLE Lab - PRHLT Research Center
Universitat Politècnica de València
Valencia, Spain
prosso@dsic.upv.es

Abstract

Most previous research on authorship attribution (AA) assumes that the training and test data are drawn from same distribution. But in real scenarios, this assumption is too strong. The goal of this study is to improve the prediction results in cross-topic AA (CTAA), where the training data comes from one topic but the test data comes from another. Our proposed idea is to build a predictive model for one topic using documents from all other available topics. In addition to improving the performance of CTAA, we also make a thorough analysis of the sensitivity to changes in topic of four most commonly used feature types in AA. We empirically illustrate that our proposed framework is significantly better than the one trained on a single out-of-domain topic and is as effective, in some cases, as same-topic setting.

1 Introduction

Authorship Attribution is the problem of identifying who, from a number of given candidate authors, wrote the given piece of text. The authorship attribution task can be viewed as a multi-class single-label text classification task where each author indicates a class. However, the purpose of AA is to model each author's writing style. AA methods have a wide range of applications, including Forensic Linguistics (spam filtering (de Vel et al., 2001), verifying the authorship of threatening emails), cybercrimes (identifying authors of malicious code and defending against pedophiles), and plagiarism detection (Stamatatos, 2011).

The AA methods can be useful in applied areas such as law and journalism where the identification of the true author of a piece of text (such as a ransom note) may be able to save lives or help prosecute offenders. One of the outstanding problems in AA studies is the unrealistic assumption that the samples of both known and unknown authorship are drawn from the same distribution. This assumption considerably simplifies the AA task but also limits the practical usability of the methods. In practical scenarios usually the documents under investigation are from a different domain than that of the training documents. We feel the need to advance the way AA methods are designed so that the bridge between domains will be minimized to obtain the optimum performance. Therefore, we try to improve the performance of cross-topic AA (CTAA), one of the dimensions of cross-domain AA (CDAA) where training and test data come from different topics.

In this paper, we focus on one of the outstanding research questions on AA: *Can we reliably predict the author of a document written in one topic with a predictive model developed using documents from other topics?* We hypothesize that the addition of training data even if it comes from a topic different than that of the test data improves cross-topic AA performance. To test the hypothesis, we compare the performance of our proposed model trained on documents from all available out-of-topic data with two models, one trained on single out-of-topic data and another trained on the same topic (intra-topic)

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1228–1237, Dublin, Ireland, August 23-29 2014.

data. We also compare the performance of using four widely used features in AA to demonstrate their discriminative power in intra-topic and cross-topic AA. The contributions of this study are as follows:

- We propose a new method to identify the author of a document on a topic using a predictive model trained on examples from different topics. The successful results attained indicate that authors maintain a consistent style across topics.
- This is the first comprehensive study showing empirically which widely used features in AA are effective for cross-topic AA. We demonstrate that character n-grams are a strong discriminator among authors in CTAA and that lexical features are less effective in CTAA than they are for intra-topic AA.
- We empirically illustrate that having the same amount of training documents from multiple topics is significantly better than having documents from a single topic. It shows that topic variety in training documents improves the performance of CTAA.
- We also demonstrate that across all genres, adding an extra topic to the training data gives a character n-gram model a greater boost in performance than to a stop-word, a stylistic or a lexical model. This is true regardless of the topics on which the model is trained.
- Our proposed methodology is simple to implement suggesting that our findings on cross-topic AA will be generalizable to other classification problems too.

The paper is organized as follows. Section 2 describes two cross-topic datasets while Section 3 describes the methodology for our experiments. Section 4 describes different features while Section 5 presents the experimental setup. We present the evaluation and analysis in Sections 6 and 7. In Section 8, we describe previous studies on cross-topic AA. Finally, Section 9 presents our conclusions and some future directions.

2 Cross-Topic Datasets

Although several corpora are available for traditional AA, we need datasets containing documents from a number of authors from different domains (different topics, different genres). We need many topics to be able to test cross-topic performance, and many genres to ensure that our findings are robust across different styles of text. Obtaining such corpora is a challenging task since most authorship attribution studies focus on a single domain. We have found two datasets that meet our criteria, one having both cross-topic and cross-genre flavor, and the other having only cross-topic flavor. The first corpus contains communication samples from 21 authors in six genres (Email, Essay, Blog, Chat, Phone Interview, and Discussion) on six topics (Catholic Church, Gay Marriage, War in Iraq, Legalization of Marijuana, Privacy Rights, and Sex Discrimination), which we call dataset 1. This dataset was obtained from Goldstein-Stewart *et al.* (2009). Using this dataset, it is possible to see how the performance of cross-topic AA changes across different genres.

Another corpus is composed of texts published in The Guardian daily newspaper written by 13 authors in one genre on four topics (dataset 2) due Stamatatos *et al.* (2013). It contains opinion articles (comments) about World, U.K., Culture, and Politics. Table 1 shows some statistics about the datasets.

Corpus	#authors	#genres	#topics	avg #docs/author	avg #sentences/doc	avg #words/doc
Dataset 1	21	6	6	36	31.7	600
Dataset 2	13	1	4	64	53	1034

Table 1: Some statistics about dataset 1 and dataset 2.

In dataset 1, the average document length is almost half the average document length in dataset 2, while the number of authors is almost twice as that in dataset 2. Also, in dataset 1, there is only one document written by an author on each topic on each genre. However, there are, on average, 16 documents per author per topic on each genre in dataset 2. Overall, dataset 1 seems more challenging and resembles more a realistic scenario of forensic investigations where very few short documents per author might be available.

3 Methodology

To answer our research question and test our hypothesis, we designed three training scenarios. First of all, to demonstrate the complexity of cross-topic tasks, we compare the performance between two training conditions: Intra-Topic (IT), and Single Cross-Topic (SCT). Once we show that it is important to solve this CTAA problem, we design one more training condition based on our proposed idea, Multiple Cross-Topics (MCT) and compare its performance with the IT and the SCT scenarios.

Intra-Topic (IT) In this scenario, all the documents in both the training and test data belong to the same topic. Although this is a strong assumption that does not hold true in most of the realistic scenarios, we examine AA under such conditions in order to be able to compare it with our proposed methods.

Single Cross-Topic (SCT) In this setting, the test data consists of documents from a single topic while the AA model is trained using documents belonging to another topic different than the topic of the test data, but from the same genre. For example, in dataset 1, for 'Chat' genre, a model could be trained on a topic 'Gay Marriage' and tested on the topic 'Legalization of Marijuana'. We experiment on all combinations of test/train topics, i.e., for each test topic, we train separately on each of the remaining topics.

Multiple Cross-Topics (MCT) Unlike in SCT and IT scenarios, here for each test topic, we train on documents from all available topics other than the one used for testing. Our assumption is that authors somehow maintain their unique writeprints across different topics. Therefore, even though the additional data comes from a topic different than that of the test data, we expect to see improvements in the performance of cross-topic AA.

In the SCT scenario, since there is a mismatch between the training and test topic, we expect to obtain experimental results worst than that of the IT scenario. However, we expect that the performance of cross-topic AA using our proposed MCT scenario will be better than SCT in all the cases.

4 Features

The choice of features depends greatly on the type of classification problem. Previous research has explored various types of features that can discriminate among the candidate authors. Stylistic features, character-level and word-level n-grams are the most frequently and successfully used features (Houvardas and Stamatatos, 2006; Zheng et al., 2006; Frantzeskou et al., 2007; Abbasi and Chen, 2008; Luyckx and Daelemans, 2011; Koppel et al., 2011). We consider four of the most widely used features. Our goal behind exploring four different types of features is to understand which features are the best for cross-topic AA.

Lexical Features. Bag-of-words is one of the commonly used document representations that uses single-content words as document features. Authorship attribution approaches using a bag-of-words representation have been found to be effective (Diederich et al., 2003; Kaster et al., 2005; Zhao and Zobel, 2005; Coyotl-Morales et al., 2006). We call bag-of-words the lexical features since we exclude stop-words.

Stop-Words. Stop-words carry no or very little semantic meaning of the texts, however, their use indicates the presence of certain syntactic structures. Although, these words are excluded in the topic-based text classification tasks due to lack of any semantic information in them, we believe these features will be effective in cross-domain AA as hinted by previous work (Goldstein-Stewart et al., 2009). Typically, words such as articles, prepositions, and conjunctions are considered as stop-words. We use a list of stop words publicly available for download (`www.webconfs.com/stop-words.php`).

Stylistic Features. Previous research has shown stylistic features to be effective in AA (Stamatatos, 2006; Bhargava et al., 2013). We use 13 stylistic features: number of sentences, number of tokens per sentence, number of punctuations per sentence, number of emoticons per document, percentage of words without vowel, percentage of contractions, percentage of total alphabetic characters, percentage of two consecutive punctuations, percentage of three consecutive punctuations, percentage of upper case words,

total parenthesis count, percentage of sentence initial words with first letter capitalized, and percentage of words without vowel.

Character n-grams. An n-gram is a sequence of n-contiguous characters. These features capture both the thematic as well as stylistic information of the texts, and hence have been proven to be very effective in previous AA studies (Keselj et al., 2003; Peng et al., 2003; Escalante et al., 2011). Since these features carry stylistic choices of the authors, we believe they will be stable across domains.

5 Experimental Settings

Following the training scenarios discussed previously in Section 3, we performed a set of experiments. We used 643 predefined stop-words. We considered as lexical features all words that were not stop words, and were among the 3,500 most frequent words occurring at least twice in the training data. We used 3,500 most frequent character 3-grams occurring at least six times in the training data.

Since dataset 1 is already balanced across authors, we used all the documents from this dataset. However, dataset 2 was originally imbalanced, therefore we chose at most ten documents per author to avoid a highly skewed distribution. In order to create a corpus like in the realistic scenarios of forensic investigations such as tweets, SMS, and emails, we chunked each selected document by sentence boundaries into five new short documents. This shortening of the documents increases the complexity of the task but enhances the practical applicability of our methods. We use these chunked versions for evaluating our proposed method. Splitting the documents in this way has been used in the past to deal with the lack of more documents per author(Luyckx and Daelemans, 2011; Koppel and Winter, 2014).

We obtained the performance measures using support vector machines (SVMs) implemented in Weka (Witten and Frank, 2005) with default parameters. We considered using SVMs because preliminary results showed this algorithm outperformed other reasonable alternatives. We used prediction accuracy as the performance measure to evaluate different training scenarios. Rather than just comparing the accuracies, we make most of the decisions based on statistical significance computed using two-tailed t-tests with 95% confidence interval.

All the experiments for cross-topic settings are carried out by controlling the genre. In the IT scenario, we computed the accuracy on each test topic using stratified 10-fold cross-validation. In the SCT scenario, for each test topic, prediction accuracy was computed by training separately on each remaining topic and averaging performances. We computed the accuracy on each test topic in the MCT scenario by withholding one topic as test topic and training on all other topics. For each training scenario, we computed one single score for each genre by averaging the accuracies across all test topics belonging to that genre.

6 Experimental Results and Evaluation

In this section, we report results and analysis on different experiments we carried out. We will start by showing empirically the challenge of cross-topic AA. Then, we will show results of our proposed approach.

6.1 Is Cross-Topic AA More Difficult than Intra-Topic AA?

Genre	Lexical Features			Stop-words			Stylistic Features			Character n-grams		
	IT	SCT	IT-SCT	IT	SCT	IT-SCT	IT	SCT	IT-SCT	IT	SCT	IT-SCT
Chat	25.71	13.11	96.11*	19.21	16.54	16.14*	**41.90**	27.49	34.39*	39.21	27.56	42.27*
Essay	26.58	5.92	348.99*	16.80	11.77	42.74*	15.66	14.56	7.02	**30.90**	13.28	132.68*
Email	19.80	6.22	218.33*	16.43	12.67	29.68*	**25.29**	24.4	3.52	24.94	14.52	71.76*
Phone Interview	37.62	10.29	265.6*	33.49	18.00	86.06*	33.02	16.16	51.06*	**56.99**	25.46	123.84*
Blog	22.18	6.32	250.95*	15.37	11.25	36.62*	13.16	11.31	14.06*	**25.38**	12.03	110.97*
Discussion	23.37	11.64	100.77*	23.37	16.31	43.29*	30.99	15.8	49.02*	**40.69**	25.28	60.96*

Table 2: Comparison of AA performance on IT and SCT scenarios on dataset 1. For each feature type, the IT and SCT columns indicate the accuracy (%) while the IT-SCT column is the relative gain of IT over SCT. For each genre, bold figures represent the best accuracy. Statistical significance is indicated by * in positive direction and by $^\flat$ in negative direction.

First of all, we want to understand if the cross-topic problem is more difficult than the intra-topic problem of AA. We compared the performance of the IT and the SCT scenarios using four types of features on various genres of dataset 1 as shown in Table 2. We clearly observed that for each genre, and for each feature type, the performance of the IT scenario is better than the SCT scenario and the difference is statistically significant. The only exceptions are 'Email' and 'Essay' genres for stylistic features. This is a strong indication that irrespective of the type of domain as well as the features considered, cross-topic AA is much more difficult than intra-topic AA.

6.2 Does Our Proposed Method Improve CTAA Performance?

We target to answer: *Can we reliably predict the author of a document written in one topic with a predictive model developed using documents from multiple other topics?* We carry out various experiments and compare the performance of our proposed MCT scenario with that of IT and SCT scenarios separately. Although, comparing MCT with only SCT would be enough to answer our research question and test our hypothesis, we are also interested in gaining more insights about cross-topic AA and understanding how it compares to IT, the simplest case of AA.

Genre	IT	SCT	MCT	MCT-IT	MCT-SCT
Chat	25.71	13.11	33.02	28.43^{*}	151.87^{*}
Essay	26.58	5.92	12.64	$-52.45^{\flat}$	113.51^{*}
Email	19.80	6.22	11.87	$-40.05^{\flat}$	90.84^{*}
Phone Interview	37.62	10.29	20.95	$-44.31^{\flat}$	103.6^{*}
Blog	22.18	6.32	13.15	-40.71	108.07^{*}
Discussion	23.37	11.64	25.26	8.09	117.01^{*}

(a) Lexical Features

Genre	IT	SCT	MCT	MCT-IT	MCT-SCT
Chat	19.21	16.54	33.49	74.34^{*}	102.48^{*}
Essay	16.80	11.77	22.06	31.31^{*}	97.08^{*}
Email	16.43	12.67	24.97	51.98^{*}	116.06^{*}
Phone Interview	33.49	18,00	38.89	16.12	115.67^{*}
Blog	15.37	11.25	20.43	32.92	81.6^{*}
Discussion	23.37	16.31	32.59	39.45^{*}	99.82^{*}

(b) Stop-words

Genre	IT	SCT	MCT	MCT-IT	MCT-SCT
Chat	41.90	27.49	37.62	-10.21	36.85^{*}
Essay	15.66	14.56	23.36	49.17^{*}	60.44^{*}
Email	25.29	24.4	33.12	30.96^{*}	35.74^{*}
Phone Interview	33.02	16.16	23.49	-28.86	45.36^{*}
Blog	13.16	11.31	15.67	26.29^{*}	38.55^{*}
Discussion	30.99	15.8	24.33	-21.49	53.99^{*}

(c) Stylistic Features

Genre	IT	SCT	MCT	MCT-IT	MCT-SCT
Chat	39.21	27.56	57.46	46.54^{*}	108.49^{*}
Essay	30.9	13.28	36.66	18.64	176.05^{*}
Email	24.94	14.52	36.53	46.47^{*}	151.58^{*}
Phone Interview	56.99	25.46	56.35	-1.12	121.33^{*}
Blog	25.38	12.03	33.41	31.64	177.72^{*}
Discussion	40.69	25.28	49.91	22.66^{*}	97.43^{*}

(d) Character n-grams

Table 3: Performance of lexical, stop-words, stylistic, and character $n-$gram features on dataset 1. The SCT, IT and MCT columns indicate the accuracy (%) while the MCT-SCT and MCT-IT columns present the relative gain of MCT over the other scenario. Statistical significance is indicated by * in positive direction and by $^{\flat}$ in negative direction.

MCT-SCT columns on Table 3 illustrate the statistical significance of MCT over SCT in a positive direction for all the genres. Using any type of feature in any genre, it is possible to significantly improve the performance of CTAA by training a machine learning algorithm using documents from all available out-of-domain topics. This serves as evidence to confirm our hypothesis and answer our research question that documents written in one topic can be reliably predicted with a model developed using documents from multiple other topics. This indicates that authors maintain a consistent writing style across topics.

In the MCT-IT column in Table 3(a), we can seen that the IT is significantly better than the MCT in three genres, while the MCT is better than the IT in only one. This is because lexical features directly capture the choices of authors in a certain thematic area, and hence they yield a good performance in the intra-topic setting. However, we observed contrasting and interesting patterns using stop-words, stylistic features, and character n-grams (MCT-IT column of Tables 3(b), 3(c), and 3(d)). MCT was better than IT, and the difference was significantly better, in 10 genres, while IT performance was significantly better than MCT in none of the genres. This is a very interesting finding as we observed that the cross-topic AA problem can be solved as effectively as the intra-topic AA problem using these features and a variety of topics.

Also using dataset 2, we found that for each type of feature, MCT is better than SCT, and the difference is statistically significant as shown in Table 4. This is another supporting evidence to our hypothesis. The small gain of IT over MCT suggests that our proposed approach is competitive even with the IT scenario.

Feature Type	IT	SCT	MCT	MCT-IT	MCT-SCT
Lexical Features	63.98	21.46	38.62	-39.64$^\flat$	79.96*
Stop-words	45.01	31.66	41.21	-8.44	30.16*
Stylistic Features	32.85	27.46	32.17	-2.07	17.15*
Character n-grams	75.08	45.87	64.54	-14.04$^\flat$	40.7*

Table 4: Performance of four types of features on three different training scenarios on dataset 2. For each feature type, the SCT, IT and MCT columns indicate the accuracy (%) while the MCT-SCT and MCT-IT columns present the relative gain of MCT over the other scenario. Statistical significance is indicated by * in positive direction and by $^\flat$ in negative direction.

6.3 Sensitivity of Features to Changes in Topic

We also want to demonstrate the behavior of four different feature types to changes in topic. We want to test if lexical features favor intra-topic AA and character n-grams favor cross-topic AA. Unlike lexical features, character n-grams carry stylistic choices of authors, and hence are expected to be robust across topics. In Table 2, for each genre, the relative gain of IT over SCT using lexical features is highest compared to that of stop-words, stylistic features, and character n-grams, thereby indicating that lexical features are more effective for ITAA than for CTAA. It is also apparent in Table 2 that the gain of characters n-grams is always better than that of stop-words and stylistic features. While looking at the performance on the SCT scenario using four features, it is observed that character n-grams give the best performance, while stop-words and stylistic features give the second best performance, which leaves lexical features at the bottom. This is because the first three features are topic-independent and hence were able to better discriminate among authors in cross-topic scenarios than lexical features. However, overall, character n-grams have the highest discriminative power in both IT and SCT, which confirms findings of earlier research (Stamatatos, 2013).

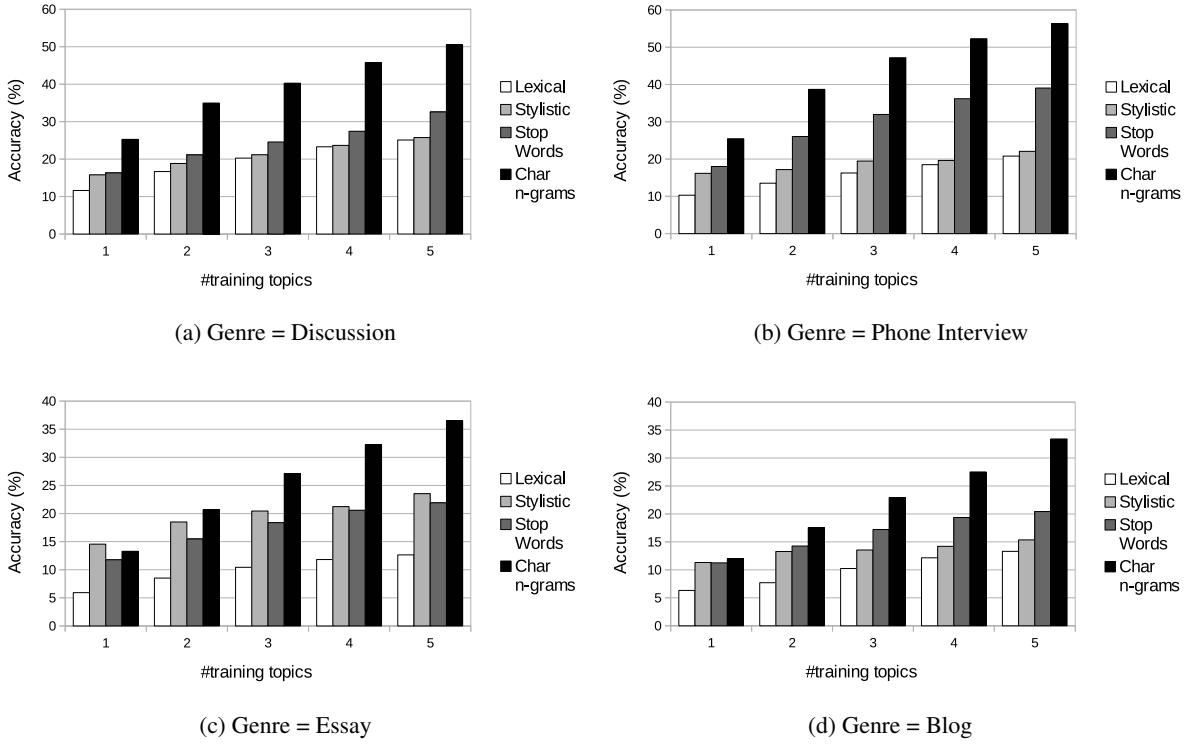
In Table 3, character n-grams, when compared to lexical features, stop-words, and stylistic features, yield the highest average relative gain on MCT over the SCT scenario (138.77%, vs 114.15% for lexical features, 97.41% for stop-words, 46.55% for stylistic features). Also, comparing the prediction accuracies of all four features separately in SCT, IT, and MCT scenarios, it is observed that character n-grams score best in most of the genres on each training scenario. This confirms that character n-grams have higher discriminative power in cross-topic AA than stop-words, stylistic features and lexical features.

For cross-topic AA, we observed that the accuracy across the board is not high. It is because the CTAA task is harder than other single domain classification tasks since the topics of the test data are fully disjoint with the topics of the training data. On top of that, the shorter document length makes it more challenging. The current system might not be production quality, but our findings will enable better models in the future that hopefully will be accurate enough to solve CTAA problems more effectively.

6.4 Cross-Topic AA with Varying Number of Training Topics

For traditional AA, it has been shown that around 10,000 word-tokens per author suffice as a 'reliable minimum for an authorial set' (Burrows, 2007). In our study, we have as few as 600 word-tokens per author, much less than the minimum size requirement stated by previous research. In this section, we look at how performance improves with increase in amount of training data by adding additional topics.

To explore this, we experimented by training on documents from all possible combinations of topics. In dataset 1, there are a total of six topics. Therefore, for each test topic, we experiment separately using one, two, three, four, and five topics for training. When measuring performance on k training topics, we gather all possible combinations of training on k of the five topics and then average the performance across all these combinations. For example, if we use two topics for training, then for each test topic, there are $\binom{5}{2} = 10$ possible training combinations that we then average to get a final score. We illustrate the results in Figure 1 for four genres using four types of features. Irrespective of the genres, topics, and types of features used, CTAA performance improves gradually with addition of more data. In most genres, this improvement seems to be almost linear with the number of topics trained on, suggesting that gathering more out-of-topic data should continue to improve the performance. We also observed that the character



(a) Genre = Discussion

(b) Genre = Phone Interview

(c) Genre = Essay

(d) Genre = Blog

Figure 1: Effect of training on varying number of topics in CTAA using lexical, stop-words, stylistic, and character n-gram features on dataset 1.

n-grams are the most effective author discriminator in cross-topic AA.

We performed a deeper analysis of the effect of individual topics, which is shown in Table 5. We took an initial topic as training data and then paired it with each of the other topics as additional training data and measured the average performance gain from the addition of the second topic. It is shown that across all genres, adding a second topic to the training data gives a character n-gram model greatest boost in performance than to a stop word or a stylistic or a lexical model. This is true regardless of the topics on which the model is trained. We do not observe negative transfer as in transfer learning (Pan and Yang, 2010) because in cross topic AA authors maintain styles across topics.

Initial Topic	**Genre = Chat**				**Genre = Email**			
	Lexical	**Stop-words**	**Stylistic**	**Character n-grams**	**Lexical**	**Stop-words**	**Stylistic**	**Character n-grams**
Sex Discrimination	5.85	5.57	1.67	10.33	2.24	7.29	8.86	9.72
Legalization of Marijuana	7.86	7.76	1.57	12.19	2.91	3.32	5.21	7.39
Catholic Church	6.24	8.76	6.24	14.33	2.41	4.48	3.59	5.22
Privacy Rights	5.9	4.66	1.9	14.05	2.97	6.45	4.6	10.06
War in Iraq	8.1	7.95	3.48	15.57	3.96	7.58	2.99	7.79
Gay Marriage	7.19	5.85	7.19	10.29	2.57	4.31	1.98	6.82

Table 5: Average performance gain from adding an additional topic as training data across different initial topics on dataset 1. Each value is the average accuracy gain after adding the second topic.

7 Is it Just 'More Data' that is Helping or is 'Diversity' Relevant?

The quantity of training data was not controlled in the experiments presented in Section 6, therefore, we performed some additional experiments where we did control for this. In Table 6, we present the comparison of SCT and MCT scenarios using the same amount of training data to understand whether the performance improvement in the MCT scenario is due to diversity or due to the fact of adding more data. We use dataset 1 to make this comparison. For the SCT scenario, for each test topic, we averaged

performance over three random samplings, where in each sampling we randomly selected four documents per author in each training topic. For the MCT scenario, for each test topic, we averaged performance

	Lexical Features			Stop-words			Stylistic Features			Character n-grams		
Genre	**SCT**	**MCT**	**MCT-SCT**	**SCT**	**MCT**	**MCT-SCT**	**SCT**	**MCT**	**MCT-SCT**	**SCT**	**MCT**	**MCT-SCT**
Chat	12.24	13.94	13.89^{*}	14.37	16.35	13.78^{*}	26.52	28.52	7.54^{*}	24.39	25.17	3.2^{*}
Essay	9.11	11.3	24.04^{*}	12.43	14.12	13.6^{*}	21.35	22.93	7.4^{*}	18.37	19.58	6.59^{*}
Discussion	9.65	10.52	9.02^{*}	12.93	13.7	5.96^{*}	19.57	20.85	6.54^{*}	19.84	21.48	8.27^{*}
Email	8.84	9.98	12.9^{*}	12.48	13.91	11.46^{*}	20.89	21.92	4.93^{*}	17.91	20.76	15.91^{*}
Phone Interview	8.94	10.84	21.25^{*}	14.65	17.67	20.61^{*}	19.73	20.94	6.13^{*}	18.84	26.35	39.86^{*}
Blog	8.45	9.66	14.32^{*}	12.78	14.05	9.94^{*}	18.53	19.62	5.88^{*}	17.58	19.95	13.48^{*}

Table 6: Comparison of MCT and SCT scenarios on controlled training data using four types of features on dataset 1. For each feature type, the SCT and MCT columns indicate the accuracy (%) while the MCT-SCT columns present the relative gain of MCT over the SCT. Statistical significance is indicated by * in positive direction and by $^{\flat}$ in negative direction.

over three random samplings, where in each sampling we randomly selected four training topics. For each selection of four training topics, we averaged performance over three random samplings where in each sampling we randomly selected one document per author in each training topic. Thus, we ended up with the same number of documents for training both models. Even with the same amount of training data, training on documents from different topics is better than training on documents from a single topic, with statistically significant performance gains ranging from 3.2% to 39.86% as shown in Table 6. This demonstrates that data from a diverse set of topics will still give a boost in performance and is always significantly better than using data from the same topic.

8 Related Work

The majority of the work in authorship attribution deals with single-domain datasets. However, there have been a handful of studies that add some cross-topic flavor in the AA task (Mikros and Argiri, 2007; Goldstein-Stewart et al., 2009; Schein et al., 2010; Stamatatos, 2013). Mikros *et al.* (2007) concluded that many stylometric variables are actually discriminating topic rather than author and their use in AA should be done carefully. However, the study was performed on a single corpus containing only two authors in two topics that raises questions on reliability of their conclusions. Stamatatos (2013) illustrated the effectiveness of character n-grams in cross-topic AA. It was also shown in that study that avoiding rare features is effective in both intra-topic and cross-topic AA. However, all these conclusions came from training an SVM classifier in only one fixed topic. In contrast, in our paper, we draw our conclusions from all possible training/testing combinations rather than fixing in advance the training topic.

Goldstein-Stewart *et al.* (2009) also carried out some cross-topic experiments by concatenating the texts of an author from different genres. This experimental setting results in a corpus where each test document contains a mix of genres, which is not representative of real world AA problems. Still, to provide some comparisons to the work of Goldstein-Stewart *et al.* (2009), we concatenated all the texts in dataset 1 produced by an individual on a single topic, across all genres to produce one document per author on each topic. We compare our results with those reported in the paper under same training/testing conditions. We withheld one topic and trained on documents from the other five topics.

Test Topic	**Lexical**	**Stop-words**	**Stylistic**	**Character n-grams**	**Stop-words + Character n-grams**	**Previous Work**
Sex Discrimination	66.67	76.19	33.33	95.24	**95.24**	95
Catholic Church	76.19	95.24	38.10	95.24	**100**	95
Gay Marriage	80.95	80.95	42.86	90.48	90.48	**95**
Legalization of Marijuana	52.38	66.67	33.33	95.24	**100**	100
Privacy Rights	42.86	52.38	28.57	95.24	90.48	**100**
War in Iraq	57.14	71.43	38.10	100	**100**	81
Average	62.7	73.81	35.72	95.24	**96.03**	94.33

Table 7: Comparing performance of our work with previous work in the same training/testing setting. The results in the last column were obtained from Goldstein-Stewart *et al.*(2009). For each test topic, the bold figure represents the best performance.

The last column of Table 7 presents the results obtained by using the combination of stop-words and 88 Linguistic Inquiry and Word Count (LIWC) features as reported in Goldstein-Stewart *et al.* (2009). We observed that the combination of character n-grams and stop-words, on average, performs better than those reported in the paper. On this fixed training/testing scenario, we see better accuracies, as high as 100%, across the board. This is because, in this experiment, each training sample on average was ≈ 25 times longer than the training sample in our chunked versions. This illustrates that authorship attribution of short documents, as in our chunked versions, is a challenging task, but we believe it resembles a more realistic scenario of forensic investigations.

9 Conclusions and Future Work

In this research, we presented the first comprehensive study with rigorous analysis on cross-topic AA. Although previous work had hinted some of our findings, it was based on very limited experiments (using only one fixed topic for training). We investigated CTAA using all possible combinations of topics to draw more robust and stable conclusions. We first illustrated the difficulty of cross-topic AA by comparing its performance with intra-topic AA using different types of features. We demonstrated that a framework trained on documents belonging to thematic areas different than that of the documents under investigation statistically improves the performance of cross-topic AA. This improves the ability of the model to find the authors of documents belonging to a new topic not present during the training of the model. By controlling the training data, we demonstrated that training on diverse topics is better than training on a single topic confirming that MCT not only benefits from more data but also from a thematic variety. We also showed a statistical analysis that lexical features are closer to the thematic area and hence were an effective author discriminator in intra-topic attribution. Similarly, character n-grams prove to be a very powerful feature especially in a condition where training and test documents come from different thematic areas. Although intra-topic AA is easier than cross-topic AA, our proposed model for CTAA achieves performance close or in some cases, better than that of an intra-topic AA model. Another interesting conclusion of our study is that addition of more training data from any topic, no matter how distant or close it is with the topic of documents under investigation, improves the performance of CTAA for all types of features. We believe that our contribution to cross-topic AA will be generalizable to other classification problems too.

In the future, we plan to explore the cross-genre problem of AA that is critical for tasks like linking user accounts across emails, blogs, and other social media. Our proposed CTAA approach can be directly applied to the cross-genre problem but we may discover different feature behavior in this scenario. We also plan to explore domain adaptation and transfer learning techniques to solve CDAA problems.

Acknowledgements

This research was partially supported by ONR grant N00014-12-1-0217, NSF award 1254108, and NSF award 1350360. It was also supported in part by the CONACYT grant 134186 and the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie.

References

A. Abbasi and H. Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):7:1–7:29, April.

M. Bhargava, P. Mehndiratta, and K. Asawa. 2013. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics*, volume 8302 of *Lecture Notes in Computer Science*, pages 37–47. Springer International Publishing.

J. Burrows. 2007. All the way through: Testing for authorship in different frequency strata. *Literary & Linguistic Computing*, 22:27 – 47.

R. María Coyotl-Morales, L. Villaseñor Pineda, M. Montes-y Gómez, and P. Rosso. 2006. Authorship attribution using word sequences. In *Proceedings of the 11th Iberoamerican Conference on Progress in Pattern Recognition, Image Analysis and Applications*, CIARP'06, pages 844–853, Berlin, Heidelberg. Springer-Verlag.

O. de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Multi-topic e-mail authorship attribution forensics. In *Proceedings of the Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security*.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. 2003. Authorship attribution with support vector machines. *Applied Intelligence*, 19:109–123, May.

H. J. Escalante, T. Solorio, and M. Montes-y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 288–298, Portland, Oregon, USA, June. Association for Computational Linguistics.

G. Frantzeskou, E. Stamatatos, S. Gritzalis, and C. E. Chaski. 2007. Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method. *Journal of Digital Evidence*, 6(1).

J. Goldstein-Stewart, R. Winder, and R. Evans Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 336–344, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Houvardas and E. Stamatatos. 2006. N-gram feature selection for authorship identification. In J. Euzenat and J. Domingue, editors, *AIMSA 2006*, volume 4183 of *LNAI*, pages 77–86.

A. Kaster, S. Siersdorfer, and G. Weikum. 2005. Combining text and linguistic document representations for authorship attribution. In *SIGIR Workshop: Stylistic Analysis of Text for Information Access*, pages 27–35.

V. Keselj, F. Peng, N. Cercone, and C. Thomas. 2003. N-gram based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 255–264.

M. Koppel and Y. Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

M. Koppel, J. Schler, and S. Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45:83–94.

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

G. K. Mikros and E. K. Argiri. 2007. Investigating topic influence in authorship attribution. In *Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, pages 29–35.

S. Jialin Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October.

F. Peng, D. Schuurmans, V. Keselj, and S. Wang. 2003. Language independent authorship attribution using character level language models. In *10th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 267–274.

A. I. Schein, J. F. Caver, R. J. Honaker, and C. H. Martell. 2010. Author attribution evaluation with novel topic cross-validation. In *The 2010 International Conference on Knowledge Discovery and Information Retrieval*, Valencia, Spain, October.

E. Stamatatos. 2006. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence tools*, 15(5):823–838.

E. Stamatatos. 2011. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, 62(12):2512–2527.

E. Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law & Policy*, 21(2):421 – 439.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kauffmann, 2nd edition.

Y. Zhao and J. Zobel. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of 2nd Asian Information Retrieval Symposium*, volume 3689 of *LNCS*, pages 174–189, Jeju Island, Korea.

R. Zheng, J. Li, H. Chen, and Z. Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.*, 57(3):378–393, February.