Rebekah Overdorf* and Rachel Greenstadt

Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution

Abstract: Stylometry is a form of authorship attribution that relies on the linguistic information to attribute documents of unknown authorship based on the writing styles of a suspect set of authors. This paper focuses on the cross-domain subproblem where the known and suspect documents differ in the setting in which they were created. Three distinct domains, Twitter feeds, blog entries, and Reddit comments, are explored in this work. We determine that state-of-the-art methods in stylometry do not perform as well in cross-domain situations (34.3% accuracy) as they do in in-domain situations (83.5% accuracy) and propose methods that improve performance in the cross-domain setting with both feature and classification level techniques which can increase accuracy to up to 70%. In addition to testing these approaches on a large real world dataset, we also examine real world adversarial cases where an author is actively attempting to hide their identity. Being able to identify authors across domains facilitates linking identities across the Internet making this a key security and privacy concern; users can take other measures to ensure their anonymity, but due to their unique writing style, they may not be as anonymous as they believe.

Keywords: Stylometry, Machine Learning, Domain Adaptation, Privacy

DOI 10.1515/popets-2016-0021
Received 2015-11-30; revised 2016-03-01; accepted 2016-03-02.

1 Introduction

Stylometry is a linguistic analysis technique which can be used to analyze the features of a document to determine the author. This field is inherently linked to privacy and security research as the use of it can provide or deprive users of anonymity online. The more robust stylometric methods become, the greater their threat to privacy. Similarly, it also becomes a greater asset for security by serving as a verification or identification tool for digital text across the Internet.

As social media and micro-blogging sites increase in popularity, so does the need to identify the authors of these types of text. The accuracy with which stylometry can identify anonymous and pseudonymous authors has direct security implications. It can be used for verification of a person's claimed identity, or to identify the author of an anonymous threat should a suspect set be present. Conversely, stylometry can be used to eliminate a user's privacy within a domain by linking their accounts within a specific web site or forum. With improvements to cross-domain stylometry, accounts could be linked across services online regardless of the type of service (such as a forum, blog, or Twitter account).

In machine learning, it is generally assumed that the underlying distribution of the data in the target domain is similar to the data that the model is build on. In cross-domain learning, however, this is not the case. Other machine learning applications have solved this problem by labeling a small amount of target data to learn about the underlying distribution. This cannot be done in stylometry as finding labeled data by the suspect authors in the target domain is often difficult.

Take, for example, an employee who wishes to expose incriminating information about the company she works for while avoiding being discovered. She may take measures to make sure that the information cannot be traced back to any of her devices or accounts. However, with the right tools, her employer may still be able to trace the writing style of the leak back to this employee based on emails that the employee has written. It is important that the employee is aware of the privacy concerns related to stylometry, so that she can make an effort to account for this when composing the documents used to expose her company's unethical practices. Making an effort to hide her writing style while creating these documents may be her only defense against such an investigation.

***Corresponding Author: Rebekah Overdorf:** Drexel University, rjo43@drexel.edu
Rachel Greenstadt: Drexel University, greenie@cs.drexel.edu

In a real-world case[1], a White House staff member was using an anonymous Twitter account to make abrasive comments about coworkers and other government officials. The identity of the person behind the account was unknown for over two years. The staff member in question primarily made comments that, while perhaps unprofessional, did not threaten national security. However, it is not so difficult to imagine a case where important information was being leaked anonymously instead of petty comments. The White House may not have Twitter feeds for all of its employees to use as training data, as many staff members may not have Twitter accounts. Instead, they could use other data (such as email content) to perform stylometric analysis. However, Twitter text is typically more informal than these professional documents. Furthermore, Twitter's 140 character limit imposes constraints that distort the style (words may be dropped, abbreviated, etc). This paper seeks to discover how these distortions affect the accuracy of using conventional stylometric approaches to solve the problem, and how to improve upon these results.

Current state-of-the-art stylometric methods can consistently identify the author of an anonymous document with accuracies over 90% from a set of 50 authors [1]. These results, however, are achieved with a few caveats that leave a number of important problems in stylometry unsolved. This work explores one such problem, the cross-domain case. In his monograph on authorship attribution, Patrick Juola [20] cautions against blindly applying stylometry across different genres and domains. Though exploration into solving this problem has begun [27], the fact of the matter is that the performance of approaches designed for books and essays when applied to emails and tweets is undefined. Even more uncertain is how models trained on one genre or domain of document fare when applied to a different domain or genre. Can you identify which of n bloggers is the author of a given Twitter feed or which of a set of Reddit commenters is the author of a given Twitter feed?

This paper seeks to address these questions. Our contributions include:

The Difficulty of Cross-Domain Stylometry: We demonstrate high accuracy at identifying authors of documents within the same domain, including blogs, Twitter feeds, and Reddit comments from novel datasets. We then show a steep drop in accuracy when these methods that succeed in in-domain stylometry are applied to cross-domain stylometry.

Feature Selection We show that careful feature selection can improve cross-domain accuracy and present two distinct methods to better select features that translate well across specific domains.

Cross-Domain Stylometry: We analyze different cross-domain situations and propose solutions of varying success to each problem. We find that adding documents in the same domain as the test documents to the training pool increases accuracy for the cross-domain problem. We also present ensemble methods for combining the results of cross-domain classification in order to improve accuracy.

Adversarial Dataset: We apply these methods to a small adversarial dataset of cross-domain data and show, for a small number of adversarial cases, which methods apply to each situation and how well they perform.

Section 2 follows with formal definitions of the various problems this paper addresses. In Section 3, we present related work on stylometry. We then present a description of the various, blog, Reddit, and Twitter datasets we used in this work in Section 4. Section 5 presents a description of our results identifying authors of Twitter feeds, blogs, and Reddit comments as well as a number of naive approaches to solving the cross-domain classification problem. In Section 7, we present solutions to the problems outlined in Section 2. In Section 8, we look at training and testing across different domains and apply our proposed solutions to our cross-domain data. We apply these same methods to a dataset of mobile and desktop tweets in Section 9 and an adversarial dataset in Section 10. We conclude with a discussion of the implications of this work for applying stylometry to non-traditional datasets of varying domains.

2 Problem Statement

When both the training and testing documents are similar in topic and domain, stylometry can be used to identify the true author of a document. However, as documents grow further apart in domain, greater effort must be put forth to connect the works of an author. In stylometry, domain adaptation is determining the author of a document written in some target domain d_1 with a classifier trained on documents in some source

1 http://www.nytimes.com/2013/10/24/us/secret-white-house-tweeter-and-national-security-council-official-loses-job.html

domain d_2. For example, we would like to discover who authored some document, such as an essay, but we do not have any other essays by the suspect authors to train on. Instead, we have emails written by each of the suspect authors. We want to use the features from the emails that describe the writing style of each author to determine which of them wrote the essay in question.

More formally, given a document of unknown authorship D in target domain d_1 and a set of authors $\mathcal{A} = \{A_1, ..., A_n\}$ with documents written in source domain d_2, which is distinct from d_1, determine the author $A_i \in \mathcal{A}$ of D.

This problem is difficult to address, and many conventional approaches fail if applied straightforwardly. Conceptually, this makes sense as different domains are typically influenced by a wide number of factors which could affect how a person writes. This includes superficial Reddit information, such as the topic, which could influence high level features such as the words used. It is also possible that the prospective audience is different and thus causes authors to change the tone of their writings. This could also affect word choice and sentence structure throughout the document, and change the amount of data present in a given document by varying the documents' lengths between domains. There are also structural differences which could force an author to adjust their writing style, such as the 140 character limit on Twitter. Through the exploration of different methods, which we then use as baselines, we supply sufficient evidence that an author's writing style is affected by the domain in which the document is written.

The hardest case for domain adaptation is what we call **Case 0: No labeled data in d_1.** In this case, models can only be built from the writing in the source domain (the OnlyI approach of Daumé and Marcu[17]). For example, this case might include trying to attribute a poem written with a unique meter. This case is difficult because there is no way to build a model about the ways that d_1 distorts the features in the domain.

In this paper, we will largely focus on the more tractable, common case where the domain in question (e.g. blogs, tweets, Reddit comments) has many examples available on the Internet. As a result, cross-domain stylometry as defined in this work differs from other cross-domain machine learning problems. Most researched domain adaptation applications include either:

– Labeled data in the source domain and unlabeled data in the target domain that you could label but it would be expensive. This is the case, for example, in sentiment analysis in different text domains.
– Labeled data in the source domain and unlabeled data in the target domain that cannot be labeled.

In contrast, most domain adaptation problems in stylometry include labeled data in the source domain and labeled data in the target domain, but the labels in the target domain are missing the class of interest. We have blogs for Alice, Bob, and Carol, and Twitter accounts for @dave, @ed, and @frank.

While online anonymity is relevant in many domains, the specific case of anonymity on Reddit is especially interesting. Reddit allows users to sign up for an account without providing any personal information such as email address or name. Because of this structure it is common for Reddit users to remain anonymous or to have multiple accounts. These multiple accounts are often called *throwaways* to indicate that they are only being used once to post or comment content that the user does not want linked to their main account.

We further break down the domain adaptation problem into three cases. These cases are distinguished by the amount of text in the unknown document.

– **Case 1: The document from the target domain d_1 is short.** A good example of this case is the Reddit throwaway account mentioned above. In this case, the unknown document (the text created on the account) is 500 words. This is enough to test a model, but not enough to train one. The analyst suspects that the author of the account is one of ten bloggers. In this case, there is ample data from d_2 (blogs) to train the model. Furthermore, because there are many blogs and Reddit comments on the Internet, we can use information from other authors to attempt to model the stylistic distortions that occur when writing in d_1 compared to d_2. This information can allow us to outperform Case 0.
– **Case 2: The document(s) from the target domain d_1 is larger.** In this case, we wish to attribute a somewhat larger account, for example, one in which we have approximately 2,000 words of text. An example of this might be a series of pseudonymously linked guest blog posts. In this case, the source domain data might be Twitter accounts of the various suspects. As in Case 1, we can use the copious numbers of blogs and Twitter feeds to model the distortions between different domains. However, in this case we can treat the individual posts as different test cases (or 500 word subsets of the document). By doing so, we create an ensemble learner that can reduce the error.

– **Case 3: Large Account Linking.** In this case, the unknown account contains enough words to train a model (for example, 4,500 words [3]). In this case we can train models from both domains and attempt to link them together. This is different from the other cases in that we are changing the problem definition slightly: given a set of authors $\mathcal{A}_1 = \{A_1, ..., A_n\}$ with documents written in domain d_1, and another set of authors $\mathcal{A}_2 = \{A_1, ..., A_n\}$ with documents written in domain d_2, which is distinct from d_1, link each author in $\mathcal{A}_1$ to an author in $\mathcal{A}_2$.

In this paper, we show that case 0 is difficult for authorship attribution, but present approaches to address cases 1, 2, and 3.

3 Related Work

3.1 Stylometry

Machine learning techniques have been used, to great success, in authorship attribution of documents. These methods have yielded impressive results, achieving accuracies of 90% with 50 authors [1] that can scale to 30% accuracy with 100,000 authors [32]. A variety of domains have been studied for authorship attribution, including tweets [25], blogs [24], source code [15], and emails [1].

Feature selection is an important part of any machine learning task, and this is especially true in stylometry. A popular feature that is often used in this field is top character n-grams, or a sequence of n-characters. This feature extracts the most popular character n-grams from each document in order to ascertain the writing style of the author. [20]

It is also common in stylometry to combine a number of features, as a person's style is made up of many different attributes that make them unique. One very diverse feature set utilized in this paper is the *writeprints* feature set [1]. This feature set contains a robust collection of features to be extracted from text that perform well within a variety of domains. These features are collected from previous works and include lexical [6, 13, 31], syntactic [5, 7, 8, 22], structural [13, 31], context-specific [14], and idiosyncratic [12, 23] attributes. The combination of these features yields a feature set that performs well in determining the author of a document within many domains. We utilize the writeprints feature set a number of times throughout this work. A summary of these features can be found in table 1.

Other method have focused on improving the performance of stylometry within specific domains explored by this paper. Almishari et al. design a method using unigrams, bigrams, and hashtags for Twitter in particular [4]. We use the more general writeprints method because (1) we are trying to link text across domains and (2) the results using the writeprints method are similar.

3.2 Breaking Stylometry

While in many cases stylometric methods yield impressive results, we must also look at the scenarios where this is not the case. Well-tested methods that succeed in some settings may fail in others. This is especially important in situations that relate to real world examples, to ensure that the application of stylometry behaves as expected to the problem at hand.

In the case of an author trying to circumvent these methods by changing her writing style, for example, it is very difficult to determine authorship of one of her documents [11, 19, 21]. Brennan, Afroz, and Greenstadt also show that it is possible for an author to imitate another author [11]. Imitation tricks the classifier into choosing the imitator at close to the same probability as the actually author of the document. Afroz et al. showed that if the classifier is trained on other deceptive documents, however, it can determine that obfuscation or imitation has taken place [2].

Another case where traditional stylometric methods fail is in the open world problem. In the open world problem, the author of the document in question may not be in the training set. In this case, it would be helpful for the classifier to be able to output a *none of the above* option. Stolerman et al. created a verification step to add on to the classification that verifies that the classifier chose the correct answer [29]. In the case that the verifier rejects the classification's choice, none of the above is the final result.

3.3 Cross-Domain Stylometry

There has been little work in domain adaptation in stylometry. Menon et. al [27] present the idea that careful feature selection may negate the need for domain adaptation in stylometry. They collected books written by 14 authors in a variety of different genres such that each

Feature Group	Feature Type	Description
Lexical	Word-Level	Total words, percentage of characters per word, word length distribution, vocabulary richness, frequency of 120 letter words
	Character-Level	Total chars, percentage of characters per document, count of letters, letter bigrams, letter trigrams, occurrence of special chars
	Digits	Digit frequency, Frequency of 2 digit numbers, frequency of 3 digit numbers
Syntatic	Function Words	Frequency of function words
	Punctuation	Occurrence of punctuation
	Parts of Speech (POS) tags	Frequency of POS tags, POS tag bigrams, POS tag trigrams
Content	Word-Level	Bag-of-words, word bigrams, word trigrams
Idiosyncratic	Word-Level	Misspelled words

Table 1. Summary of the writeprints feature set used in this paper. Note that the original feature set also includes a *structural* feature group, but was removed for this work as it is specific to the structure of the domain that the documents are written in.

author has at least 25 works. Although they test a number of features, they are able to achieve the best results by using stop words (or function words). When we apply this method, however, (see Section 5.4) to the more difficult cross-domain problems explored in this work, we find no significant improvement.

3.4 Doppelgänger Finder

The Doppelgänger Finder algorithm [3] is a method developed to link users with multiple accounts within and across cybercriminal forums. The Doppelgänger Finder algorithm works by training a n leave-one-out models, one for each of n accounts. That is, the algorithm removes each author A_i, and trains a logistic regression classifier on the remaining authors. It then tests the classifier on the documents by A_i and collects the probability scores that those documents are written by the other authors. If the probability that A_i wrote the documents by A_j is high and the probability that A_j wrote the documents by A_i is high, then A_i and A_j are likely the same person. Accounts which exceed a threshold t are considered to be doppelgängers.

With small modifications, this algorithm is applicable to Case 3 in Section 2, but not Cases 1 and 2, because it requires enough testing data to train a model. We are also able to make modifications to the algorithm because we are in a cross-domain case and therefore have more information about which authors are linked. As a result, our goal is to take the highest probability cross-domain account, rather than using a threshold. A threshold could still be useful in an open world case. In Section 8, we show that Doppelgänger Finder does work well in the account linkage scenario (Case 3).

We show that the primary reason that Doppelgänger Finder produces improvements (both within and across domains) is that it effectively produces an ensemble of classifiers. Combining these classifiers is able to reduce uncorrelated errors between them by averaging. Doppelgänger Finder produces an ensemble of two classifiers (one trained on each account to be linked), and also separates the accounts into multiple output targets and combines the results of the classifier on these outputs. We find that other aspects of the original doppelgänger finder algorithm, such as adding a constant to each feature value and applying principle component analysis, do not help and sometimes harm the accuracy of the classifier across domains and therefore should be omitted from our approach.

Doppelgänger Finder can only be applied in the case where there is enough text in the unknown account to train a model. However, in Case 1 and Case 2, there is insufficient text for this condition to apply. In this paper, we present novel methods to use ensembles to reduce the error rates in these conditions. In Case 1, we can use disjoint subsets of the feature sets to produce multiple classifiers that can be aggregated, and we can do so in a way that takes into account the changes in the feature distributions across the domains of interest. We can also use a mixed training approach to alter the feature weights and improve the classification results. In Case 2, we can aggregate the outputs of our classifier on different subsets of the test data.

4 Corpora

We utilized two novel datasets of texts, both of which were collected for the purpose of this research.

The first is a collection of users who had accounts on Twitter[2] and authored blogs on the blogging site Wordpress[3]. All of the users publicly linked their Twitter and Wordpress accounts. All of the gathered text was preprocessed; all hashtags, tags, and links were removed. Not including such factors is intended to broaden the impact of the methods used to domains beyond the ones tested in this work. We collected thousands of users' data, but only about 200 users had enough data in each domain (10,000 words) for us to consider. Our experiments utilize a random subset of these authors as experiments beyond 50 authors are computationally expensive. Previous work in stylometry (in-domain) has suggested that 4,500 words is sufficient for a good training model and 500 words for a good test document [3]. Therefore, 10,000 words allows us enough data to train a model for each domain and test it using cross-validation.

The second dataset is a collection of users who had accounts on both Twitter and Reddit[4]. We collected about 100 users who fit our criteria (10,000 words in each domains). The usernames were collected from users of the subreddit */r/twitter*[5] who also posted their Twitter handle on the site. We preprocessed the tweets in the same way as in the blog-tweet dataset. Links within the Reddit comments are filtered out as well. From this dataset, we also created a third cross-domain dataset containing tweets written on a mobile device and tweets written on a desktop.

For all of the data used in this work, we grouped the documents into 500 word documents. This creates consistency for each experiment. We use the term *Twitter feed* to mean a group of tweets that is at least 500 words long. This changes the question from *Who wrote this tweet?* to *Whose Twitter feed is this?*. This is just as practical a question, as often the author of an anonymous Twitter account comes into question and rarely is it the case that a such a Twitter feed would contain only one tweet. Practically, using Twitter feeds instead of individual tweets means that each test case has more text from which we can extract features. This has been tested in the literature [9, 26] on authorship attribution of tweets to great success.

5 Baseline and Naïve Approaches

Before exploring domain adaptation methods for stylometry, we first demonstrate the need for such methods. In this section, we not only present evidence that establishes a need for more sophisticated methods for domain adaptation in stylometry, but also set baselines. First we explore the in-domain case, in which all documents are written in the same medium. We then apply the same methods that work well in-domain to a cross-domain dataset and show that these methods perform poorly in this case.

While it is clear that the in-domain classification will provide better results than cross-domain classification, any cross-domain analysis would be incomplete without first examining the in-domain case. Not only is it helpful in better understanding the problem, but it also provides us with a clear upper-bound for how well a cross-domain solution can perform.

From our two cross-domain datasets we have four domains: blogs, Twitter feeds of bloggers, Reddit comments, and Twitter feeds of Reddit users. The Twitter feed datasets are kept separate to acknowledge that there may be a distinction between the style of tweets written by the different groups of users.

The in-domain results we obtain with this section are similar to those in related work.

5.1 The Writeprints Feature Set

As discussed in detail in Section 3.1, the writeprints feature set is an extensive and thorough set of characteristics that works well in a number of different domains. We find that it also works well, although in varying degrees, in all of the domains we considered. For consistency here and throughout this work, we partitioned all of the data into 500 word documents and guarantee that no text from any post was split between documents. This ensures that we never try to classify a document in which some of the text comes from the same post or entry as a training document, as that will artificially inflate results. In order to measure scale, we vary the number of authors, averaging the true positive rate for each set of experiments.

2 http://www.twitter.com

3 http://www.wordpress.com

4 http://www.reddit.com

5 http://www.reddit.com/r/twitter

5.2 Classification

For classification we use logistic regression as implemented in *scikit-learn*[6]. Half of the data is used as training documents to build our classifier and the other half is set aside as test documents. The same training documents used in the in-domain experiments are used in the cross-domain experiments to maintain consistency. Logistic regression is also used as the underlying classifier in the Doppelgänger Finder algorithm [3].

The classifier used in prior stylometry work has varied; Brennan et al. use a linear SVM [11], Narayanan et al. use regularized least squares classification(RLSC) [28], Abbasi and Chen use both a customized K-L transform combined with a euclidean distance measure and a linear SVM with similar results [1].

We performed some experiments using an SVM as the underlying classifier. In these cases, the results were similar, but slightly worse and not statistically significant. This is consistent with other studies comparing linear SVM and regularized logistic regression [30]. Generally speaking, for linearly separable problems, regularized linear classifiers will perform similarly. Our in-domain results and those in previous studies, show that stylometry is a linearly separable problem on which linear classifiers perform well.

There are two reasons to prefer logistic regression over a linear SVM. First of all, the performance of logistic regression is significantly faster and analysis can thus be completed more quickly. This is particularly relevant for scenarios where we are training and combining ensembles of linear classifiers. Second, logistic regression provides estimates of the posterior probabilities for the classes that are not entirely based on the discriminant function, as the estimates provided by an SVM are. Our methods which make use of ensembles depend on these probabilities and there is reason to suspect that they might be (slightly) more accurate in the logistic regression case [30].

5.3 Naïve Approach

For the naïve domain adaptation approach we replicate the previous in-domain experiments on cross-domain data. We label this technique as naïve because the naïve intuition behind this approach is that a method that works well classifying text in both domains should similarly perform well at classifying text across these domains. We use the same data to build our cross-domain classifier as the in-domain, so only the testing documents differ.

5.4 Baseline Results

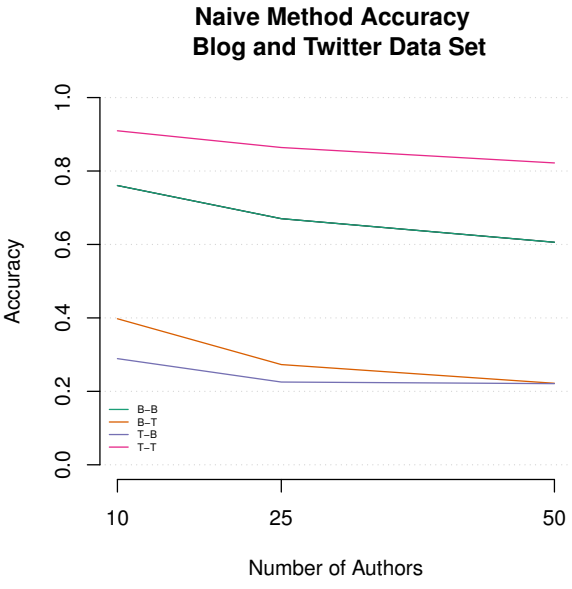

Fig. 1. Naïve results for the blog and Twitter dataset using the writeprints feature set.

Figure 1 shows the in-domain and cross-domain results for the blog and Twitter dataset. Similar to other work in in-domain authorship attribution, straightforward machine learning techniques are able to identify the authors of each document with relative ease. Twitter feeds are more accurately attributed than blogs and we do not see the same drop in accuracy as the number of classes increases. Expanding the number of classes even further would eventually cause a drop such as the one seen in the blog data, as this is generally true of classification tasks. The cross-domain accuracy when applying the same methods and training on the same data, however, fails to attribute a majority of the documents correctly.

Figure 2 shows the in-domain and cross-domain results for the Reddit and Twitter dataset. The results for this dataset reflect the results shown on the blog and Twitter dataset. While we are able to achieve a high accuracy in-domain, the cross-domain classification fails

6 http://scikit-learn.org/

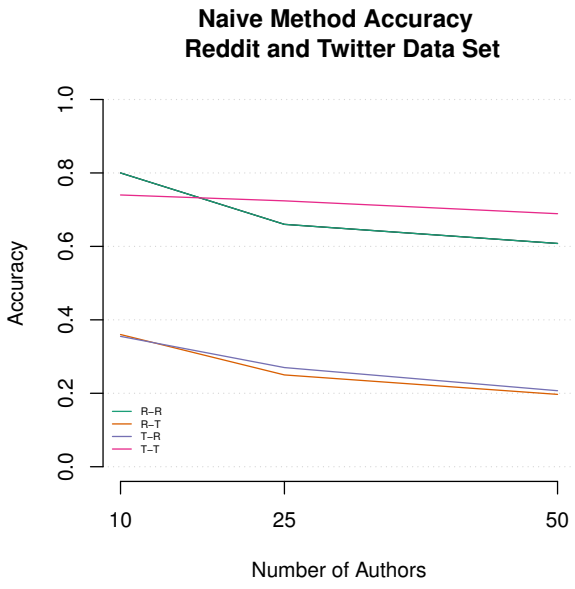

Fig. 2. Naïve results for the Reddit and Twitter dataset using the writeprints feature set.

to identify the authors of more than half of the documents.

may also be an artifact of the domains tested. Results are shown in Figure 3.

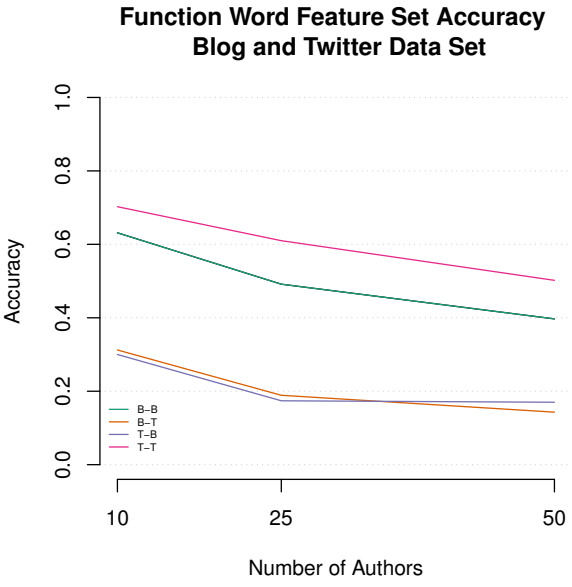

Fig. 3. In-domain and cross-domain results using only function words.

6 Feature Analysis

Feature selection is an important part of any machine learning problem. The writeprints feature set, discussed in Section 3, is a rich and diverse feature set. The question remains, though, is there a better set of features that are more resistant to cross-domain stylometry?

6.1 Function Words

We aim to remove the concept of context or subject from the feature extraction phase by using only non-lexical features. This idea comes from [27], which we describe in section 3.3. They were able to increase accuracy when classifying books written in different genres by using only function words. We use the same setup as the in-domain experiment, changing only the feature set used. Here, we obtain worse results in the in-domain case, as this is a much smaller and less representative set of features. We similarly obtain worse results than [27]. We suspect this is due to the large discrepancy in the amount of data we have to train the classifier on, but

6.2 Feature Type Comparison

While function words alone may not be the solution to cross-domain stylometry, it is possible that a set of features exists that works well for certain domain problems. Here we use different types of features, as broken down by writeprints in Table 1, in order to determine which are best for this cross-domain problem.

Figure 4 shows the results of using only one type of feature for both in-domain and cross-domain datasets. We see that some feature groups are not strong enough on their own to perform well in-domain or cross-domain, such as misspellings. We see that cross-domain accuracy does not perform as well as in-domain for any meaningful feature group. This analysis does, however, give us some insight into the types of features that do not change across domains, as part of speech (POS) features perform comparatively better in-domain.

In addition to using feature groups alone for classification, we also combine the results to form an ensemble classifier. Each feature group experiment yields a probability vector, $p_f(d)$ for each document d and feature group f. To meaningfully combine these probabil-

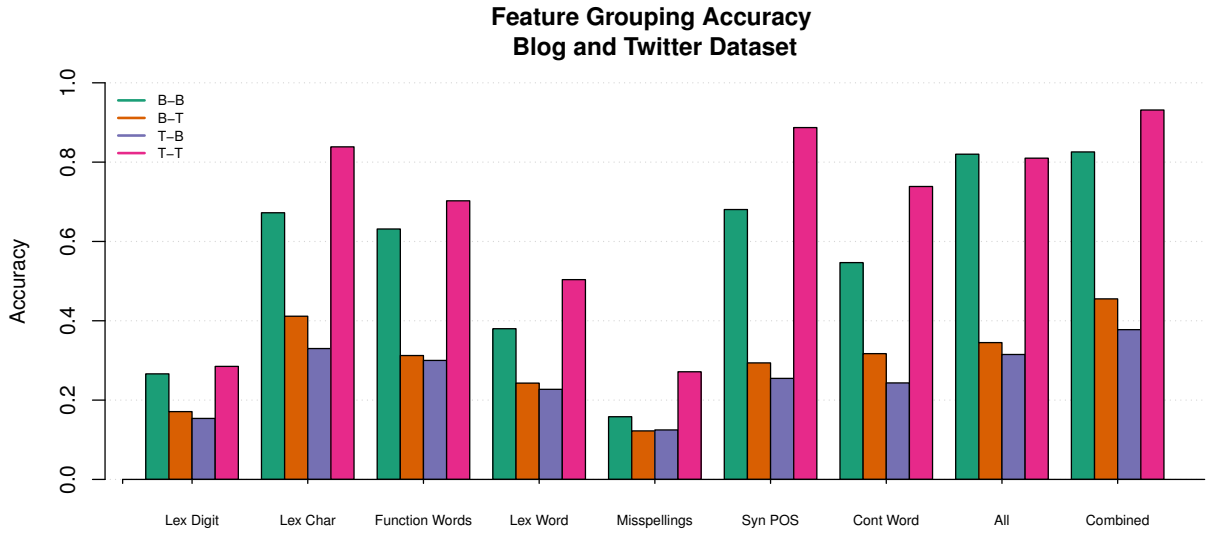

Fig. 4. In-domain and cross-domain results using a variety of feature sets on the blog and Twitter dataset. The suspect set size for these experiments is 10, making random change accuracy 0.1.

ity vectors, we sum them, effectively creating a weighted average. That is,

$$p_e(d) = \sum_f (p_f(d))$$

Since more confident feature group classifiers are weighted higher, this is essentially a weighted average. This gives us a slight improvement in-domain and a larger improvement across domains. In contrast, Lex Char n-grams, perform better alone than using all the features with an accuracy of 0.41 for cross-domain blog to Twitter experiments compared to 0.35 using all features.

6.3 Estimating Domain Difference

We can also use feature vectors extracted from documents in different domains in order to understand how different domains differ by computing how much *distortion* happens between domains. We calculate the distortion between domains by computing how much an author's style varies between domains. An author's distortion is defined as the euclidean distance between the sum of the features for each instance in one domain and the other. That is,

$$dist = \sqrt{\sum_{i=0}^{len(f)} (I_{d1}[i] - I_{d2}[i])^2}$$

where I_{d1} and I_{d2} are the sum of the feature vectors for instances in distinct domains $d1$ and $d2$ and f is the list of features used. When comparing documents in the same domain, we split the data in half to create two domains, which will yield a non-zero distance and is then a measure of how varied an author's writing is in a domain. The average distortion value for all of the cross-domain datasets are expressed in Tables 2, 3, and 4.

	Wordpress	Twitter
Wordpress	722.89	2,703.59
Twitter	2,666.09	671.22

Table 2. This table shows the average per author euclidean distance between the feature vectors in the Wordpress and Twitter domains.

The distance between the feature vectors for the Wordpress and Twitter domains (Table 2) are more than double that of the in-domain data. The Twitter only feature vectors are the closest, supporting the results in the previous section that show the in-domain Twitter problem set achieving the highest results. Interestingly, there

is some difference between the Wordpress-Twitter feature vectors and the Twitter-Wordpress vectors. When choosing which features to extract, only source domain documents are parsed, as they are the documents on which we will build the model. Some of the features, for example certain *word bigrams*, will exist in the source domain, but not the target domain for any given problem set. This will cause some variation in the distance values as the feature vectors will not be the same across different problem sets, even with the same documents, if the source and target domain are switched.

	Reddit	Twitter
Reddit	**1,203.47**	**3,051.78**
Twitter	**2,526.56**	**618.04**

Table 3. This table shows the average per author euclidean distance between the feature vectors in the Reddit and Twitter domains.

We see similar results in the Reddit and Twitter dataset (Table 3). Again, the Twitter feature vectors are closer together than the Reddit only feature vectors, but both in-domain problem sets are considerably closer together than their cross-domain counterparts.

	Mobile Tweets	Desktop Tweets
Mobile Tweets	**557.03**	**783.57**
Desktop Tweets	**688.28**	**523.95**

Table 4. This table shows the average per author euclidean distance between the feature vectors in the Mobile Tweets and Desktop Tweets domains.

The mobile tweets and desktop tweets analysis tells a slightly different story. While the in-domain distortion values are lower than the cross-domain distances, the difference between them is much less stark. These domains are much closer in style than the other cross-domain problems. Additionally, the in-domain distances are slightly lower and the cross-domain distances slightly higher than the Twitter-only data in the other analysis, which includes both mobile and desktop tweets.

6.4 Top Feature Analysis

Recursive Feature Elimination [16] is a method used with linear classifiers to determine the importance of features and reduce dimensionality through recursively reducing the number of features by eliminating those with the smallest weight. We first analyze which features are most important on a dev set of authors and then use this information to improve the accuracy on a distinct test set of authors in the same domain.

Table 5 shows the results when the top features from one problem set are applied to a different problem set in the same domain. Here, we use the blog and Twitter dataset. Each experiment uses 10 authors. The fact that performance improves as more features are added suggests that high dimensionality is not the cause of poor performance. On the other hand, the fact that the improvement is so slight with added features suggests that the top 1,000 features provide most of the discrimination in the model.

In Section 6.3, we discussed the distortion of feature vectors between domains. One approach to domain adaptation is to seek out *pivot* features [10], those features which behave in the same way for discriminative learning in both domains. Table 6 shows the results when the least distorted features are used (for the same 10 author experiments as Table 5). Removing the distorted features reduces accuracy, and furthermore, it reduces accuracy further than removing the least important features. This suggests that, at least in the blog/Twitter problem, the most discriminating features are also the most distorted. While their distortion is why the cross-domain problem performs worse than the in-domain problem, removing them still does more harm than good. We also experimented with transforming the feature vectors based on the distortion and this produced even poorer results. These results suggest that careful feature selection alone is not an effective means of cross-domain stylometry.

Maximum Ranking	10	50	100	None (∼200)
Approx # Features	**1,000**	**5,000**	**10,000**	**20,000**
Accuracy	**0.364**	**0.365**	**0.375**	**0.376**

Table 5. Accuracy when the top RFE ranked features from blogs (as determined on a separate dev set) are used to classify tweets. Note that lower ranked features are more important.

Least Distorted Features	1,000	2,000	5,000	None (∼20,000)
Accuracy	0.312	0.312	0.313	0.376

Table 6. Accuracy when the least distorted features (as determined on a separate dev set) from blogs are used to classify tweets.

7 Classifier Solutions to Cross-Domain Stylometry

Recall from Section 2 that we define the domain adaptation problem as follows: given a document of unknown authorship D in some domain d_1 and a set of authors $\mathcal{A} = \{A_1, ..., A_n\}$ with documents written in domain d_2, which is distinct from d_1, determine the author $A_i \in \mathcal{A}$ of D.

7.1 Mixed-Training

In cases where there is abundant data available in the target domain, some information can be gained about the underlying distribution. While Section 6 explored how to estimate the changes in the underlying distributions by changing the feature vectors, here we explore changing the classifier itself.

The mixed-training approach removes each author A_i one time and trains a classifier on the remaining authors (from both domains). It then tests the classifier on the documents by A_i and collects the probability scores that those documents are written by the other authors, discarding results that are in the same domain. The methodology is that each iteration of mixed-train combines the data from both the target and source domains.

We provide the following theory as to why adding unlabeled instances from the target domain into the problem improves the results.

The goal of a learning system is to estimate class y from feature vector x. In all the domains we studied, the same feature set provides good in-domain results. Therefore, we hypothesize that the primary task of domain adaptation for authorship attribution is instance adaptation [18]. The different domains have different feature frequencies. Consider $p_s(x, y)$ and $p_t(x, y)$ are the true distributions of the source and target domains, where x are the features, and y are the class values. While $p_s(y|x)$ is similar to $p_t(y|x)$, the feature frequencies $p_s(x)$ and $p_t(x)$ are different, a model that learns an estimation of $p_s(y|x)$ from samples of the source domain to work well for the target domain. However, we can use the (unlabeled) target instances to bias the estimate of $p_s(x)$ to a better estimate of $p_t(x)$. This is what is happening when we add unlabeled instances of the target domain to the problem. Then a discriminant function is built for each author of d_s comparing it to a combined set of instances from d_s (the other authors in the source domain) and d_t (unlabeled authors from the target domain that we know have not authored the test document). The classifier boundary will be based on both what separates the author from the other authors in d_s and what separates the author from the authors in d_t, and will give higher weight to *pivot* features [10], those features which discriminate well in both domains.

7.2 Averaging and Ensemble Methods

In cases where there are many documents in the target domain, for example a number of blog entries, the results for each document can be combined, or averaged, down into one result vector. That is

$$p_{ave}(a) = \frac{\sum_{i=0}^{len(D)}(p_i(a))}{len(D)}$$

where D is the list of test documents and $p_i(a)$ is the probability value for a suspect a. Since this computation is the sum of the probability values, it favors the higher confidence classifications, which will generally perform better than straight voting, where each document is instead assigned an author and the most common author choice is chosen. In voting, the lower confidence classifications are counted the same as a higher confidence classification. In addition, the averaging method also considers all of the probability values, not only the top choice author.

In the account linking case, we have additional information. We know that each author in d_1 has a corresponding author in d_2. We also have many labeled documents in the both domains. We can exploit this information by creating two classifications tasks: one built on documents in d_1 that attributes documents in d_2 and one built on documents in d_2 that attributes documents in d_1. We then average the results for each document as described above and are left with two probabilities for each author: the probability that a_1 wrote the documents by a_2 and the probability that a_2 wrote the documents by a_1. We straightforwardly multiply these two values to create a combined probability that a_1 and a_2 are the same user.

7.3 Augmented Doppelgänger Finder

A combination of mixed-train and averaging, the doppelgänger finder [3] algorithm, shown in Algorithm 1 was developed as a way to link users with multiple accounts within cybercriminal forums. Because we are not trying to link all accounts to each other, and only those across different domains, we are under more constraints which gives us a slight advantage over the situation that the doppelgänger finder was intended to solve. Doppelgänger finder is attempting to discover the author of each document from all other authors. We, however, aim to find the authors of documents in one domain given documents in another. The original doppelgänger finder algorithm also includes the addition of a constant to each feature value and applies principle component analysis to the features. We applied both of these modifications, individually and simultaneously, and observed no notable difference in the results.

It is notable that our implementation of the Doppelgänger finder algorithm has the advantage of not needing a parameterized threshold. This independence is useful for closed world situations. In open world problems, using the verification threshold regardless would allow for the algorithm to announce that it could not find an author, suggesting that the true author was not in the training set.

Algorithm 1 Augmented Doppelgänger Finder

Require: Set of authors $\mathcal{A}^\alpha = A_1, ..A_n$ and associated documents, D where each D is in the domain α; Set of authors $\mathcal{A}^\beta = A_1, ..A_n$ and associated documents, D where each D is in the domain β
Ensure: A map of authors from domains α to β where each $A_i \in \mathcal{A}^\alpha$ is mapped to an $A_j \in \mathcal{A}^\beta$, M
for $A_i \in \mathcal{A}^\alpha \cup \mathcal{A}^\beta$ **do**
▷ **Mixed-Training**
n = Number of documents written by A_i
$C \Leftarrow$ Train on all authors $\in \mathcal{A}^\alpha \cup \mathcal{A}^\beta - A_i$
$R \Leftarrow$ Test C on A_i (R contains the probability scores per author.)
▷ **Averaging**
for $A_j \in R$ **do**
$$Pr(A_i \to A_j) = \frac{\sum_{x=1}^{n} Pr(A_{jx})}{n}$$
end for
end for
▷ **Ensemble**
for $(A_i, A_j) \in \mathcal{A}^\alpha \cup \mathcal{A}^\beta$ **do**
$P = Pr(A_i \to A_j) * Pr(A_j \to A_i)$
end for
for $A_i \in \mathcal{A}^\alpha$ **do**
Find the author A_j such that $P(A_i, A_j)$ is maximum for all $A_j \in \mathcal{A}^\beta$
$M.add(A_i, A_j)$
end for
return M

8 Results

In both datasets, we, again, see that blindly applying techniques created for in-domain stylometry to cross-domain stylometry fails to correctly attribute the author most of the time. However, we do see an increase in accuracy when any of the methods described in this paper are applied.

Case 1: The document from the target domain d_1 is short. Where there is one document to be classified, is a fairly difficult case. However, since there may be data in the domain of unknown authorship, there are a few methods that can increase accuracy. Section 6 explored how feature selection can increase accuracy. In addition, the mixed-training method can also help in this case, as data from the target domain that may not be relevant to this problem can be added to the training pool to increase accuracy.

Case 2: The document(s) from the target domain d_1 is larger. This case gives us a little more information. Here, we have the ability to perform averaging on the different documents (or the larger document split up) in order to combine the results. We see an increase in accuracy when we average the probability values for each author. Figures 5 and 6 show the accuracies for problems in case 2.

Case 3: Large Account Linking. Finally in the broadest case, we can apply the augmented Doppelgänger finder. Combining the mixed-train method, averaging, and our ensemble method to combine the domains we achieve an accuracy that even outperforms the naïve in-domain accuracies. Figures 7 and 8 show the accuracies for problems in case 2.

9 Mobile vs Desktop Tweets

Another case where domain adaptation methods may be useful is identifying the authors of mobile tweets from those written on a desktop. These two domains are a

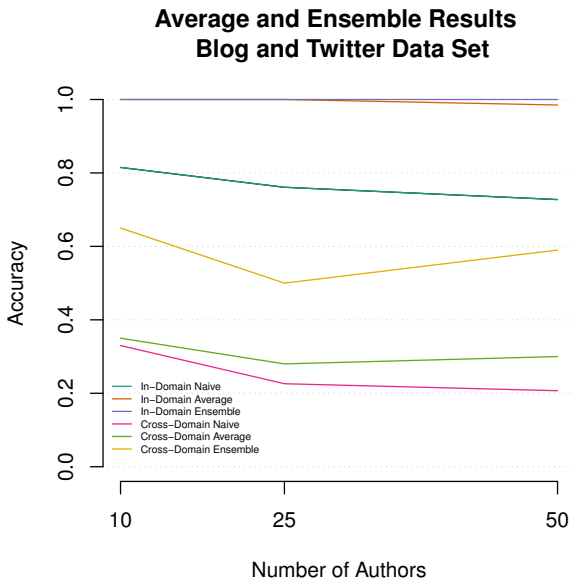

Fig. 5. Overall results for the blog and Twitter feed dataset using the writeprints feature set. We see the naive approach of blindly applying in-domain methods to this dataset fails. When we apply averaging and the ensemble method to this dataset, however, our results rebound to achieve a more acceptable accuracy.

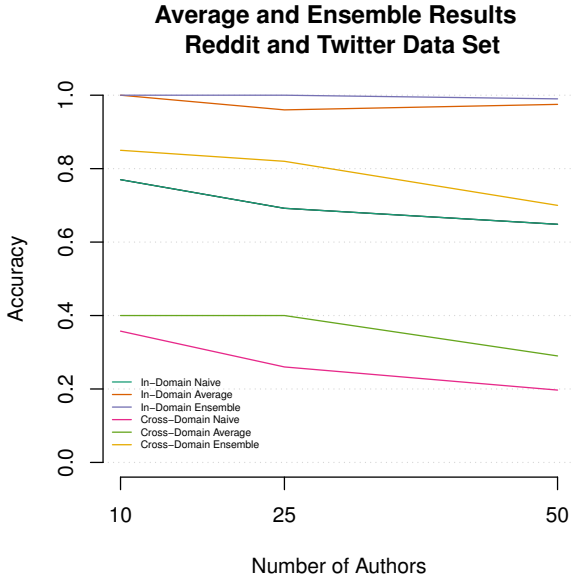

Fig. 6. Overall results for the Reddit and Twitter feed dataset using the writeprints feature set. We see the naive approach of blindly applying in-domain methods to this dataset fails. When we apply averaging and the ensemble method to this dataset, however, our results rebound to achieve a more acceptable accuracy.

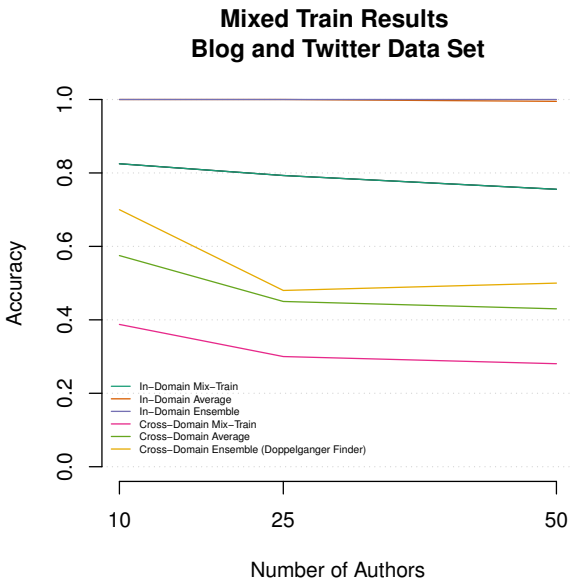

Fig. 7. Overall results for the blog and Twitter feed dataset using the writeprints feature set. We see the naive approach of blindly applying in-domain methods to this dataset fails. When we apply averaging and the ensemble method to this dataset, however, our results rebound to achieve a more acceptable accuracy.

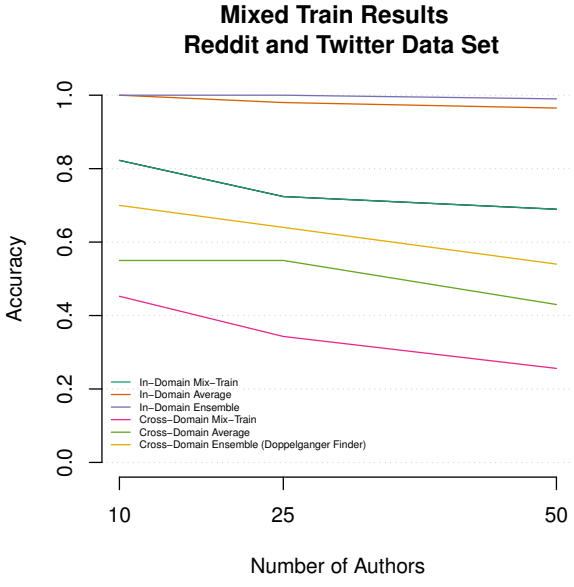

Fig. 8. Overall results for the Reddit and Twitter feed dataset using the writeprints feature set. We see the naive approach of blindly applying in-domain methods to this dataset fails. When we apply averaging and the ensemble method to this dataset, however, our results rebound to achieve a more acceptable accuracy.

subset of our Twitter dataset. While domains like the ones described in this work so far may be obviously different, here we explore a case of similar, but distinct domains: tweets written on mobile devices and tweets written on a desktop.

We utilize a subset of the Reddit/Twitter dataset described in Section 4, using only authors with more than 10,000 words of both mobile and desktop tweets, setting aside 5,000 for building the model and 5,000 for testing. Because of the relatively small number of such authors, all experiments on this dataset were limited to 30 authors. The results of both in-domain and cross-domain experiments are shown in Figure 9.

While there is a small but consistent drop in accuracy across domains, we see a similar drop in accuracy between the two in-domain datasets. The mixed-train algorithm does not show any for of the improvement that we see in the other cross-domain datasets studied in this work. As with the more obvious cross-domain data sets, averaging outputs and combining probabilities from both domains provide major improvements in both the in-domain and cross-domain cases.

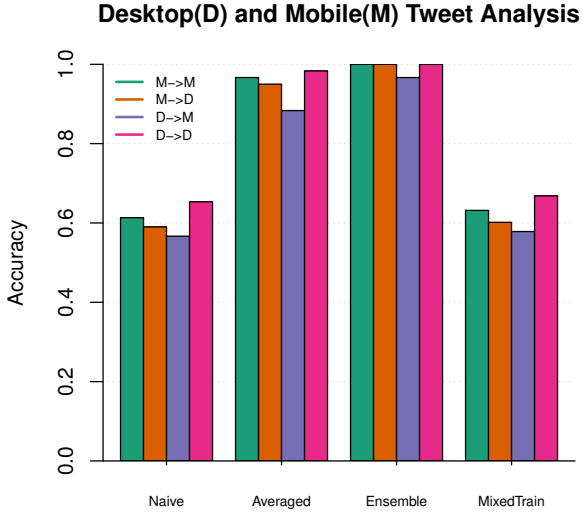

Fig. 9. The accuracy for the mobile-desktop tweet dataset. There is little difference in the in-domain and cross-domain results.

10 Adversarial Data

In most of our experiments, we consider cases in which the authors explicitly linked to their accounts in other domains. For example, our Reddit commenters disclosed their Twitter handles in /r/twitter and our bloggers linked to their Wordpress accounts in their Twitter profiles. This gives us a large data set in which we are confident about the ground truth, and avoids the ethical problems of deanonymizing users who wish to protect their identity.

However, it has the problem of not addressing the question of whether these techniques work in an adversarial case, when the author does not intend to leak her identity. To address this case, while still preserving ethical practices and maintaining confidence in our ground truth, we look at a few case studies of instances in which an account was initially created as an anonymous account, but the author later came forward and acknowledged the account. In each of these cases, the author was also creating an alternative persona or style, which is a known difficult case for stylometry [11].

Adversarial Dataset.

We consider the following real world adversarial cases:

- **Invisible Obama** [7] is a parody account representing the tweets of the empty chair addressed by Clint Eastwood at the Republican National Convention. The account remained anonymous until Ian Schafer admitted to creating it a month later. We collected tweets by @invisibleobama during the period of anonymity as well as tweets from @ischafer and posts from his blog (https://medium.com/verses-from-the-abstract).
- **Fake Steve Jobs**[8] Fake Steve Jobs is a blog written by Daniel Lyons. He was able to remain anonymous for one year until his identity was revealed by the New York Times in 2007. We collected posts from the anonymous period of iamnotstevejobs.blogspot.com as well as @realdanlyons and blog content from http://www.realdanlyons.com/.
- **Claire North**[9] is the author of *The First Fifteen Lives of Harry August*. She also writes as Kate Griffin and Catherine Webb (her real name). The

7 twitter.com/InvisibleObama

8 www.fakesteve.net and iamnotstevejobs.blogspot.com

9 www.kategriffin.net and twitter.com/clairenorth42

Claire North identity (and accompanying Twitter feed) was unlinked to her other identities for a considerable time. She revealed her identity when she felt the mystery of "Who's Claire North" was overwhelming the story. We collected tweets from the anonymous period of @clairenorth42 as well as blog content from www.kategriffin.net.

Adversarial Problem Setup.
To determine how successful our methods are for attributing fake accounts in d_{fake} to authors with text in some other domain d_{real}, we create a number of problem sets with 10 authors each where we build a classifier on the other domain, including the known real account, and test on the text found in the fake account. Table 7 includes a breakdown of the data for each adversarial case.

Fake	d_f	Words	Real	d_r	Words
Steve Jobs	Blog	53,000	Daniel Lyons	Twitter	34,500
Invisible Obama	Twitter	11,000	Ian Schafer	Blog	2,500
Claire North	Twitter	30,000	Kate Griffin	Blog	171,000

Table 7. Adversarial dataset.

Adversarial Results.
Results for the adversarial dataset are shown in Table 8. The results for each dataset vary greatly. This may be due in part to the lack of data for averaging, but it also illustrates the point that each of these problems, which differ in domains, amount of data, and the level in which the author hid their style, are extremely different. In some cases, domain adaptation may not be needed, in others our approach may work, and in others, the style may be so changed as to be almost indistinguishable.

	Naive Approach	Averaged/ Ensemble	Mixed-Training
Fake Steve Jobs	0.70	1.00	0.87
Invisible Obama	0.03	0.00	0.22
Clair North	0.00	0.00	0.05

Table 8. Accuracies for the adversarial dataset.

11 Discussion and Future Work

Cross-domain stylometry is challenging and applying standard methods to these problems results in misclassifications. Our results show that it is possible, when the problem permits, to accurately classify accounts across domains.

Why Cross-Domain Stylometry is Hard.
Cross-domain stylometry is difficult when the features that separate the authors in the source domain are distorted in the target domain. Section 6 demonstrates how these differences manifest in the different domains studied by calculating the level of distortion and the discriminating power of the features.

Approaches to Cross-Domain Stylometry that Work.
We present several approaches that improve the situation. These improvements come from making changes that reweight the features to reduce distortion or by using ensemble learning to improve results by combining multiple classifiers with uncorrelated errors. For case 1, we can improve results by (1) using the fact that many writeprints features are redundant to build an ensemble of feature subspace classifiers and (2) we can weight features in a way that takes distortion and importance into account by including documents from the target domain (even when not by suspect authors) in our training set (the mixed-training approach). We show that attempting to select only features that are less distorted may be counterproductive as these may be the most discriminating features (as is true is the blog-tweet case). In case 2, we can further improve results by breaking up the target document into 500 word chunks and aggregating the results. Case 3 allows us to utilize the Doppelgänger Finder algorithm to provide further improvements. In Section 6, we show that the distortions in features across domains are not purely symmetric. Therefore, the errors in cross-domain classifiers from d_1 to d_2 and d_2 to d_1 are not highly correlated. As a result, aggregating the results of the two cross-domain classifiers, as Doppelgänger Finder does, produces large gains in accuracy.

Feature subspace aggregation, output averaging, and Doppelgänger Finder all have in common that they create ensembles that combine weaker classifiers to produce a stronger one. These techniques also apply to in-domain stylometry problems, and improve the results

there. One avenue for further improvement would be to aggregate not only using the classifier probabilities, but also using the performance of the feature subsets on a cross-domain training set to improve the aggregation. We also have not studied combining feature subspace aggregation with mixed training, and using this algorithm as the base classifier in Doppelgänger Finder. The degree to which these techniques improve results will depend on whether their improvements are correlated or not.

Currently, cross-domain stylometry is limited to closed-world problems, which have a specific suspect set. It is very conceivable that an open-world cross-domain stylometry problem would come up in real world settings. Recall from the Introduction the example of the White House staffer. Ten staffers could have their emails passed through the mixed-training approach, but if none of them were the author, it will still classify one of them. Further work on authorship verification should be done in applying this algorithm to open world scenarios. The augmented doppelgänger finder could be adjusted to incorporate this by use of a authorship verification threshold. The mixed-train approach may be applied along with methods in [29].

Finally, we would like to apply these domain-adaptation methods to more domains so as to test it more rigorously as a cross-domain classification tool. Some future domains to explore are poems, essays, reviews, and scholarly articles. Along the same lines, we would like to expand the adversarial dataset and more extensively understand how this case differs from the non-adversarial data.

12 Conclusion

Stylometric analysis has continued to advance over recent years, each improvement yielding an increase in the domains it can be applied to, its accuracy, or other increases in the robustness of the methods. This paper furthers the field by providing multiple contributions to in-domain and cross-domain stylometry. We establish high accuracies with in-domain classification of blogs, Twitter feeds, and Reddit comments in Section 5. We demonstrate that those same methods don't perform well when applied to cross-domain stylometry, as well as investigate other domain adaptation algorithms used in other contexts that do not work for authorship attribution. We then present a number of methods which perform well on the non-adversarial datasets we collected. Finally, we explore applying these methods to a small adversarial dataset.

Advances in authorship attribution offer both positive and negative repercussions for security and privacy. However, it is important to understand the assumptions that underlie these good results. We caution against blind application of stylometric methods when crossing stylistic gaps, such as domain and topic, as misclassifications are likely.

References

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

[2] S. Afroz, M. Brennan, and R. Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*. IEEE, 2012.

[3] Sadia Afroz, Aylin Caliskan-Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelgänger finder: Taking stylometry to the underground. In *IEEE Security and Privacy*, 2014.

[4] Mishari Almishari, Mohamed Ali Kaafar, Ekin Oguz, and Gene Tsudik. Stylometric linkability of tweets. In *Workshop on Privacy in the Electronic Society (WPES)*, 2014.

[5] Shlomo Argamon, Moshe Koppel, and Galit Avneri. Routing documents according to style. In *First International workshop on innovative information systems*, pages 85–92. Citeseer, 1998.

[6] Shlomo Argamon, Marin Šarić, and Sterling S Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480. ACM, 2003.

[7] Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. An experiment in authorship attribution. In *6th JADT*, pages 29–37. Citeseer, 2002.

[8] Harald Baayen, Hans Van Halteren, and Fiona Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996.

[9] Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics*, pages 37–47. Springer, 2013.

[10] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

[11] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3):12:1–12:22, November 2012.

[12] Carole E Chaski. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8:1–65, 2001.

[13] Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

[14] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied intelligence*, 19(1-2):109–123, 2003.

[15] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, Carole E Chaski, and Blake Stephen Howald. Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *International Journal of Digital Evidence*, 6(1):1–18, 2007.

[16] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[17] Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

[18] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Association for Computational Linguistics (ACL)*, 2007.

[19] P. Juola and D. Vescovi. Empirical evaluation of authorship obfuscation using jgaap. In *Proceedings of the 3rd ACM workshop on Artificial intelligence and security*, pages 14–18. ACM, 2010.

[20] Patrick Juola. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334, 2008.

[21] G. Kacmarcik and M. Gamon. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451. Association for Computational Linguistics, 2006.

[22] Moshe Koppel, Navot Akiva, and Ido Dagan. Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57(11):1519–1525, 2006.

[23] Moshe Koppel and Jonathan Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, pages 72–80. Citeseer, 2003.

[24] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.

[25] Robert Layton, Paul Watters, and Richard Dazeley. Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*, pages 1–8. IEEE, 2010.

[26] Robert Layton, Paul Watters, and Richard Dazeley. Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*, pages 1–8. IEEE, 2010.

[27] Rohith Menon and Yejin Choi. Domain independent authorship attribution without domain adaptation. In *RANLP*, pages 309–315. Citeseer, 2011.

[28] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin, and D. Song. On the feasibility of internet-scale author identification. In *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*. IEEE, 2012.

[29] Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. Classify, but verify: Breaking the closed-world assumption in stylometric authorship attribution. *International Conference on Digital Forensics*, 2013.

[30] Tong Zhang and Frank J. Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4, April 2001.

[31] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.

[32] Sven Meyer Zu Eissen, Benno Stein, and Marion Kulig. Plagiarism detection without reference collections. In *Advances in data analysis*, pages 359–366. Springer, 2007.