Blogs and Twitter Feeds: A Stylometric Environmental Impact Study

Rebekah Overdorf, Travis Dutko, and Rachel Greenstadt

Drexel University
Philadelphia, PA
{rjo43,tad82,greenie}@drexel.edu
http://www.cs.drexel.edu/

Abstract. Stylometry is the study of determining the author of a document based on the linguistic features contained in the document. Previous work in this area has yielded impressive results, but assumes that the training and testing documents are similar key attributes, namely the domain and setting in which they are written. This paper focuses on the scenario where this assumption cannot be made. We determine that standard methods in stylometry do not perform well when the training and suspect documents differ in this way. For example, when working exclusively with blogs we obtain an average accuracy of 93.30% and with Twitter feeds we obtain an average accuracy of over 98.99%. However, when we apply the same method to try to identify a twitter feed via a blog's writing, accuracy falls drastically. We provide a method to improve this cross-domain accuracy to 88.89%. Being able to identify authors across domains facilitates linking identities across the Internet, making this a key privacy concern.

1 Introduction

The need for a robust stylometric method is integral in keeping online communication anonymous as we aim to understand how supposedly anonymous communications can be attributed to an author. The accuracy and robustness with which stylometry can identify anonymous and pseudonymous authors has a direct bearing on the ability to produce anonymous speech online.

Take, for example, an employee who wishes to expose incriminating information about the company he works for while avoiding being discovered. He may take other measures to make sure that the information cannot be traced back to him, but with the right tools, his employer may still be able to trace the writing style of the leak back to this employee based on emails the employee has written. It is important, then, that the employee is aware of this possibility so he can take steps to keep himself anonymous.

While current methods achieve high accuracies within a number of domains, these methods do not take into account the possibility that the training and testing documents may be in different domains. As a result, many of them fail in this situation. We show that a document can be attributed to its author, even if no training data in the same domain is available with an accuracy of 88.89%.

For this paper, we wish to examine the scenario in which you want to identify a document of a particular domain (in our case, a Twitter feed) but do not have additional training data for that author in that specific domain. Instead, we have blogs that each of the suspect authors has written. We want to use the features from the blogs that describe the writing style of each author to determine which of them wrote the Twitter feed in question. More formally, we aim to discover the author of a document d in some target domain D_t from a set of authors A. Every author in A has sample documents in domain D_s, which is distinct from D_t. Our goal is to utilize the information we have from the documents in D_s about the authors' writing styles to discover the author of d. We supply evidence in this paper that an author's writing style changes depending on which domain she is writing in, making this problem difficult to solve via conventional means. An alternative approach is required.

Our contributions include:

- We demonstrate high accuracy at identifying authors of blogs from blogs (93%) or authors of Twitter feeds from other Twitter feeds (98%) on 15 author sets.
- However, we also show that training on blogs and testing on Twitter feeds using standard methods results in accuracy that drops to 25% on 15 author sets.
- Lastly, we apply an augmented version of Doppelgänger Finder [2], a stylometric approach for multiple account detection that can handle small stylistic changes, to the domain adaptation problem in stylometry. This provides significant improvements in the blog-to-Twitter case resulting in 88.89% accuracy on average, however requires more training data of the to-be-tested document than traditional approaches.

2 Related Work

Machine learning techniques have been used, to great success, in authorship attribution of documents. Feature selection is an important part of any machine learning task, and this is especially true in stylometry. A popular feature that is often utilized in this field is the frequency of top character n-grams [12, 17, 20], where we extract the most common n-grams. An n-gram is the number of occurrences of n characters in a row.

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 makes 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 [4, 10, 22], syntactic [3, 5, 6, 13], structural [10, 22], context-specific [11], and idiosyncratic [9, 14] attributes. The combination of these features yields a feature set that performs well in determining the author of a document in many domains.

For classification in natural language processing, sequential minimal optimization support vector machines [18] (SMO SVM) are often utilized. SVM's have had much success in authorship attribution [7, 8, 10, 12] and allow for a large feature set [19].

There has been little work in domain adaptation in stylometry. [16] makes a claim that careful feature selection may negate the need for domain adaptation in stylometry. The authors use only stop words (or function words) to identify the authors of books across different genres. They achieved an accuracy of over 97% using this method on books with from a verity of genres. In total, they collected at least 25 works from 14 distinct authors that spanned a large number of genres. Again, the authors use SVM SMO for classification.

Another recent advancement in stylometry is a method known as Doppelgänger Finder [2]. This method was introduced as a way to link users with multiple accounts within the same forum. The Doppelgänger Finder works by removing each author A_i one time and training a 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.

There are a number of tools developed recently to facilitate authorship attribution in both real-world and research settings. In this paper, we utilize the JStylo[1] [15] and JGAAP[2] authorship attribution frameworks for our experiments.

3 Corpora

For this research, we collected a dataset of authors who had accounts on Twitter and published blogs on the blogging site Wordpress. In total, we collected blogs and tweets from 57 users. To pre-process the tweets, we stripped out all hashtags, tags, and links, as they are not necessarily tied to writing style and may make the solution too specific to this situation. Not including such factors hopefully makes this solution usable outside of the blogs and tweets domains. The average number of words for each blogger is 21,153 words and the average number of tweets per user is 3,336. In each experiment, we use a random selection of 15 authors from this pool of 57, and then average the results.

4 Naïve Approaches

In this section, we aim to set a baseline for the domain adaptation problem by exploring two naïve approaches. In the first we implement no domain adaptation

[1] `https://github.com/psal`

[2] `http://evllabs.com/jgaap/`

and use a method that performs well in each domain. In the second we use only non-lexical features. We will use the results outlined in this section as a baseline for comparison.

4.1 Baseline 1

In order to create the first baseline, we must find a method that performs well in each domain independently. To this end, we use the writeprints feature set [1], as described in section 2. The writeprints method has been shown to work on a verity of data sets from many domains including emails, forum posts, chat logs, and feedback comments from peer to peer commerce websites. In addition, we use a sequential minimal optimization support vector machine for classification. Within each domain, we group the texts into similarly sized documents and divide the authors into groups of 15 for analysis. This method is able to identify the author of a 500 word document an average of 93.30% of the time for each experiment. Similarly, this method identifies the author of a group of 30 tweets an average of 98% of the time for each experiment. When we use this same method to try to cross domains, that is where we train on blogs and test on Twitter feeds (a collection of grouped tweets), our accuracy drops to an average of 31.94% for each experiment.

4.2 Baseline 2

For the second approach, we aim to remove the concept of context and subject from the feature extraction by using only non-lexical features. This idea comes from [16], which we describe in section 2. Note that the authors achieved an accuracy of 97%. We replicated these experiments on our own data set; training on 3,500 words of blogs for each author and testing on 500 word documents of both blogs and tweets. We use the same feature set as well, frequency of stop words. In addition, we use 10 authors for each experiment, less than the 14 authors used in the work that we are replicating. We randomly arranged the authors into three problem sets ten times and averaged the results. Accuracy is reported as the total number of correctly attributed documents out of the total number of documents within each domain. The average accuracy across experiments for the blog test documents was 60.50% and the average accuracy for the Twitter test documents was only 32.91%. The relatively low accuracy for the blog test documents can be attributed to the lack of training data, however the stark difference between the blog and Twitter accuracies gives us evidence that this method does not work in the general case of domain difference and only when changes in content or genre in the same medium (i.e. books) does this method appears to succeed. Note, also, that even when all of the blogs for an author are used as training data (an average of 21,153 words per author), the accuracy is similar at 31.20%.

5 Methodology

To circumvent the loss in accuracy that occurs when using the naïve approaches, we propose modifying a method previously used to connect accounts across forums. This method, named Doppelgänger Finder [2], is well suited for this problem, as we are, essentially, trying to link the accounts.

The Doppelgänger Finder method works by removing each author A_i from the set of all authors one time and training a 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.

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 $\alpha$; Set of authors $\mathcal{A}^\beta = A_1, ..A_n$ and associated documents, $D$ where each $D$ is in the domain $\beta$
Ensure: A map of authors from domains $\alpha$ to $\beta$ where each $A_i \in \mathcal{A}^\alpha$ is mapped to an $A_j \in \mathcal{A}^\beta$, $M$
  $F \Leftarrow$ Add weight k with every feature frequency (default k=10)
  $F' \Leftarrow$ Features selected using PCA on $F$
  ▷ Calculate pairwise probabilities
  for $A_i \in \mathcal{A}^\alpha \cup \mathcal{A}^\beta$ do
      $n =$ Number of documents written by $A_i$
      $C \Leftarrow$ Train on all authors $\in \mathcal{A}^\alpha \cup \mathcal{A}^\beta$ using $F'$
      $R \Leftarrow$ Test $C$ on $A_i$ ($R$ contains the probability scores per author.)
      for $A_j \in R$ do
          $Pr(A_i \rightarrow A_j) = \frac{\sum_{x=1}^{n} Pr(A_{jx})}{n}$
      end for
  end for
  ▷ Find highest probabilities
  for $(A_i, A_j) \in \mathcal{A}^\alpha \cup \mathcal{A}^\beta$ do
      $P = Combine(Pr(A_i \rightarrow A_j), Pr(A_j \rightarrow A_i))$
  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 $M$
```

We modify the Doppelgänger Finder algorithm for use in domain adaptation in stylometry. Because we are not trying to link all accounts to each other, we are under more constraints. These constraints give us a distinct advantage over the situation that the Doppelgänger Finder was intended to solve. The Doppelgänger Finder is attempting to discover the author of each document from all other authors. However, we aim to find the author of a single Twitter feed from a selection of blogs. We do not have to compute all of the pairwise probabilities between authors, just the pairwise probabilities between each blogger and the Twitter feed in question. Our modification to this algorithm is described in Algorithm 1.

6 Results

To test our Augmented Dopplegänger Finder algorithm, we used blog documents of 500 words and groups of 30 tweets. We found that our method outperformed the baseline accuracies outlined in section 4, achieving an accuracy of 88.89% across three tests with random sets of authors. Table 1 summarized the results found in this paper.

Train	Test	Method	Authors	TP Rate
Twitter Feed	Twitter Feed	Writeprints/SMO	15	98.99
Blogs	Blogs	Writeprints/SMO	15	93.30
Blogs	Twitter Feed	Writeprints/SMO	15	31.94[1]
Blogs	Twitter Feed	Function Words/SMO	10	32.91[2]
Blogs	Twitter Feed	Augmented Doppelgänger	15	88.89

[1] Baseline 1

[2] Baseline 2

Table 1. This table shows a summary of the domain adaptation the results outlined in this paper. The best results are achieved when the testing and training documents are in the same domain. When that is not available, however, we can still attain a relatively high accuracy using the augmented Dopplegänger Finder

7 Conclusions and Future Work

Stylometric analysis has continued to advance over recent years; each improvement acquires a new domain in which it is usable or improves a domain's accuracy. This paper contributes to this advance via taking a step towards bridging the documents of different domains. The methods which work well in a single domain need not apply when attempting to cross the boundary. This problem can be solved using an alternative approach-in the situation of identifying a Twitter feed from a blog with the augmented Dopplegänger Finder. However, there is still room for this idea to be expanded upon.

One scenario we would like to explore in the future is the situation where training documents in both the same domain as the test document and a different domain than the test document are available. That is, the situation where you would like to determine the author of a document in domain D_A, but posses training documents of some proportion in both D_A and a distinct domain D_B. Depending on which proportion of the training data is in each domain, it may be beneficial for you to omit one of the domains and train on only one of them. Alternatively, there may be a method that extracts certain features from each set individually and combines them into a model to be tested on. Finally, it may be possible to find a training system robust enough to use any type of training data and still attain a high accuracy. We would also like to see which other domains the Augmented Doppelgänger Finder works well in. For example, can you find the author of an email based on a collection of blogs or can you find the author of a something as structured as a poem based on a Twitter feed.

Advances in authorship attribution make retaining privacy online more difficult, however, it is important to understand the assumptions that underly these good results. Blind application of these methods across domains may yield poor results. However, the Augmented Doppelgänger Finder approach makes these difference in style less significant. As documents are linked more easily to identities in a particular domain, the privacy of users in that domain (such as Twitter) decreases. With increased research into cross-domain classification, and further refinement of the algorithms applied in this paper, the advances in one domain may affect the privacy in others. Users who create a private account on one service but one linked to their real identity on another may find their privacy threatened by cross-domain stylometry.

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. Sadia Afroz, Aylin Caliskan-Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelgänger finder: Taking stylometry to the underground. 2014.
3. 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.
4. 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.
5. Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. An experiment in authorship attribution. In *6th JADT*, pages 29–37. Citeseer, 2002.
6. 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.
7. Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics*, pages 37–47. Springer, 2013.
8. 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.
9. Carole E Chaski. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8:1–65, 2001.

10. 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.
11. Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied intelligence*, 19(1-2):109–123, 2003.
12. Patrick Juola. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334, 2008.
13. 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.
14. 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.
15. Andrew McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. Use fewer instances of the letter "i": Toward writing style anonymization. In *Privacy Enhancing Technologies Symposium (PETS)*, 2012.
16. Rohith Menon and Yejin Choi. Domain independent authorship attribution without domain adaptation. In *RANLP*, pages 309–315. Citeseer, 2011.
17. Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 267–274. Association for Computational Linguistics, 2003.
18. J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.
19. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.
20. Efstathios Stamatatos et al. Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, pages 41–46. Citeseer, 2006.
21. 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.
22. 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.