Use of Language as a Cognitive Biometric Trait

Neeti Pokhriyal
Dept of Computer Science
University at Buffalo, SUNY
neetipok@buffalo.edu

Ifeoma Nwogu
Dept of Computer Science
University at Buffalo, SUNY
inwogu@buffalo.edu

Venugopal Govindaraju
Dept of Computer Science
University at Buffalo, SUNY
govind@buffalo.edu

Abstract

This paper investigates whether the cognitive state of a person can be learnt and used as a novel biometric trait. We explore the idea of using language written by an author, as his/her cognitive fingerprint. The dataset consists of millions of blogs written by thousands of authors on the internet. Our proposed method learns a classifier that can distinguish between genuine and impostor authors. Our results are encouraging (we report 72% Area under the ROC curve) and show that users do have a distinctive linguistic style, which is evident even when analyzing a corpora as large and diverse as the internet. When we tested on new authors that the system had never encountered before, our methodology correctly identified genuine authors with 78% accuracy and impostors with 76% accuracy.

1. Introduction

We investigate whether the cognitive state of a person can be learned and used as a soft biometric trait. Cognitive biometrics is defined as the process of identifying an individual through extracting and matching a unique signature based on the cognitive, affective, and conative state of that individual. Cognitive biometrics like many behavioral patterns, falls under the category of soft biometrics. It can be used alone or in conjunction with other biometric modalities depending on the level of security required. The use of cognitive traits for biometrics is relatively under-explored in the biometrics research community.

There is currently an increasing need for novel biometric systems that engage multiple modalities, especially because of the changing notion of privacy in today's world. As we increase the quantities of personal and identifiable data on cloud networks such as Google drives, Dropbox, iCloud, etc., passwords alone will no longer suffice as the only authentication methods. Passwords suffer from a variety of vulnerabilities including brute-force and dictionary based attacks and as a result, these highly personalized data on the cloud become vulnerable to being "stolen". Also, soft biometrics can be relatively easy to capture and process and can then be used in combination with physical biometrics, which are often more difficult and obtrusive to obtain and work with. Cognitive biometrics will become more essential as pervasive computing becomes more prevalent, where computing can happen anywhere and anytime.

We explore the idea of using language as a cognitive biometric. As studies by psycholinguist D. Carroll [5] have shown, people often develop a jargon that is more meaningful to other people in similar age groups, social backgrounds, academic settings, etc. Thus, language becomes a badge of sorts, identifying information about the user. Given the psychology-based evidence, we demonstrate that language of a user does provide distinctive characteristics that can be used as a cognitive biometric trait for authentication. While traditional biometric models use physical characteristics of individuals such as face, fingerprints, retina scans, voice, etc., using language written by an individual as an identifying characteristic has not been explored in this context before. Our results are encouraging and show that users do have distinctive writing styles, which is evident even when analyzing a corpora as large and diverse as the internet. Although language written by an individual encompasses various facets, like handwritten material, or structured text like books and other formal publications, in this research, we concentrate our effort on blogs, which are unstructured digital text authored on the internet. We use data from the blogosphere, because of its easy availability, and reflection on how people author text informally in their day-to-day settings.

Our proposed method learns a classifier that can distinguish between genuine and impostor authors. As with a typical biometric authentication system, users are initially enrolled into the system and we treat this as our training/validation dataset. We then perform different quantitative tests including an ROC analysis to show the biometric strength of using language as a cognitive biometric trait.

2. Problem Formulation

We define a cognitive fingerprint of a user as a unique signature that captures the user's mental thoughts or experiences (or cognition). Based on a psycholinguist study [5], language can be used as one such metric. As language is medium of communication and has varied connotations (spoken word, gestures, handwritings etc) attached to it, in this work we only refer to the written text of a person. We exclude hand-written texts, which have been extensively studied in literature for biometric identification [19]. We therefore use large unstructured text that is available online in the form of blogs written by many users, for this study. Blogs are extremely popular ways of communication over the web. For instance, there are around 172 million Tumblr and 75.8 million WordPress blogs.

To get a realistic dataset of authors and their natural writings, we looked for the blogosphere data, where people typically identify themselves with a username, and pen down their thoughts as entities called *blogs*. The dataset, used in this work is the ICWSM 2009 Spinn3r Dataset [4] where the corpus consists of 44 million entries, a snapshot of weblog activities from August 1, 2008 and October 1, 2008.

Hence, our problem is to develop a biometric authentication system that can determine whether any given individual, posing to be an enrollee, is genuine or an impostor. This involves obtaining biometric scores for each (user, enrollee) pair, where the pair is genuine if the enrollee is very similar to the user, or else is an impostor. The problem is posed as a binary classification problem, where the two classes are genuine and impostor. Each data point in this space, is a feature vector representation of the blogs written by the user, such that the feature values are small if the blogs are written by the same author, and large if written by an other user. They are dissimilarity scores, with low scores meaning low dissimilarity (or high similarity), and high scores meaning high dissimilarity. The features are differences of each of 213 stylometric features and 50 thematic features.

3. Related Work

The problem of authorship-attribution has studied in literature where given a certain author, the goal is to determine whether a not-previously-seen piece of writing can be attributed to that author. The problem has been studied in different guises with datasets of varying sizes and types. Stylometric analysis techniques have been used for attributing authorship in the past (see Federalist papers [14]) and for more recent surveys, see [11], [18]. Authorship attribution studies deal with author-identification and similarity detection. Identification involves comparing anonymous texts against those belonging to identified ones, where each anonymous text is written by one of those entities. Similarity involves comparing anonymous text, and assessing the degree of similarity. Drawing parallels between biometrics, the former can be seen as biometric verification and the latter is authentication.

A related problem formulation is in the area of online privacy and anonymity [15] where the goal is to unmask an anonymous blogger or whistleblower. In this area of work, one piece of text is compared against every piece of writing within a large corpus of text whose authors are known. Authorship of the anonymous text is thus attributed to that of the most similar text in the corpus. Plagiarism detection is yet another variant of authorship attribution where portions of writings are compared against large bodies of published writings, although this is more related to the use and arrangement of words than to other cognitive writing features (developed through mental experiences). Another direction of related study is authorship deception identification [16], which deals in identifying an author if there is a suspected author who may be trying to anonymize his/her message or is actively imitating another author's writing in order to conceal his/her true identity (in biometrics this will be related to "spoofing").

The different types of features used in authorship attribution [1] works include lexical, syntactic, structural, content features. Stylometric features have also been used for forensics and privacy assessment, and other features used include relative frequency words, character n-gram, word n-grams, part-of-speech n-grams and vocabulary richness etc. The various classification algorithms used include naive Bayes, neural networks, K nearest neighbor, etc.

4. Contributions

In our work, we perform a full biometric analysis of using language as a cognitive biometric, analyzing both the genuine and impostor signatures of the users. The following factors distinguish our work from some of past related works in this area of privacy and anonymity.

- Very recently researchers have started looking at written language usage as a biometric trait[[7, 8]. Some of the cognitive modalities reported in the literature involve the use of biological signals captured through electrocardiograms (ECG), electroencephalograms (EEG) and electrodermal responses (EDR), to provide possible individual-authentication modalities [6]. However, these are invasive and require users to have the electrodes placed on their specific body parts. It is an exciting prospect to investigate the use of language by people as a cognitive biometric trait, based on the previously reported psycholinguistic study [17].
- Our biometrics analysis are performed on a very large corpus of user real data, having several thousands of authors and writings. In general such large scale stud-

ies are not typical in biometrics although are essential in order to transition biometric systems from the lab to real-life. Because we are evaluating language as a soft biometric, it is important to have a large scale study such as this for complete results, since deductive results can only be obtained when large data is studied.

- This study also incorporates *big data into biometrics* where our dataset is characterized by high volume and high noise (veracity).

Some of the questions that we seek to answer in this study include:

- Can language be used as cognitive biometric trait? Does it satisfy the traits of a biometric system, namely: identifiability; universality; uniqueness; permanence; collectability; performance; and acceptability?
- Are unstructured blogs on the internet a good metric to capture the linguistic signatures of authors?
- How does the number of blogs per author, the total number of authors and the quality of blogs (richer, cleaner text) affect the results?

We have performed an extensive sets of experiments under various settings to empirically obtain some answers to these questions above. We varied the number of blogs per author, the total number of authors, and the size of the subset of the entire dataset, in order to study how well language can be used as cognitive biometric trait and report our methodology and findings in the ensuing sections of the paper.

5. Data Description

The ICWSM 2009 Spinn3r data was collected in 2009, as a part of 3rd International AAAI Conference on Weblogs and Social Media sponsored Data Challenge. The raw data is obtained by crawling various blog publishing sites to get the syndicated text of the blogs and the associated contents. Thus, the corpus consists of a snapshot of weblog activities from August 1, 2008 and October 1, 2008. It consists of 44 million entries in Spinn3r.com website's XML format.

The raw format includes the RSS and the ATOM descriptors, and also several meta-data tags. The blogs are also arranged in 13 tier groups based on the blog influence. Since the data is obtained as a snapshot of the blogosphere at a certain time, a portion of the items contain only the first few hundred characters rather than the entire text of the weblog entry. Also as evident in cyber space, much of the data is not *real* blog entries, like many are posts in threaded online conversations, advertisements, blogs just posting images, or forwarding mails.texts etc. Additionally the dataset consisted of blogs written in many languages, including English.

	All Tiers	Tier 1
# Blogs	153,796	65,945
# Words	61,703,661	24,724,073
# Authors	3510	1714
Avg. # Blogs	44	38
Median # Blogs	22	21
Max. # Blogs	2477	1400

Table 1. Data characteristics for all tiers and only for tier 1 of the Spinn3r data set.

To get blogs that contained real text written by authors, we chose a subset of the ICWSM 2009 Spinn3rDataset, which is called the Personal Stories Dataset [9]. This dataset, consisted of the only the blogs which can be best characterized as a personal story. The logic, behind such decision, was our intuition that personal stories are expected to contain more distinguishing writing style markers. Thus we get blogs which are personal stories written by the blog author. Depending on the nature of internet, many of the blogs which were marked as personal stories, however had no author name associated with the ⟨authorname⟩ tag in their XML mark-up. We chose the unique identifier as the text contained in the ⟨authorname⟩ within each XML ⟨item⟩ and discarded the blogs which had no author.

The data is organized into 13 tiers, with stories spread across all tiers, but the most number of stories are found in Tier 1. The characteristics of the data for all tiers and for Tier 1 are summarized in Table 1. We report statistics for authors who have written at least 15 blogs. Since Tier 1 has the most number of stories [9], we chose tier 1 for majority of our experiments.

To find the distribution of blogs per author, we found that some authors write lots of blogs, while a lot of authors write few blogs. Thus, the number of blogs per author follows power law as shown in Figure 1.

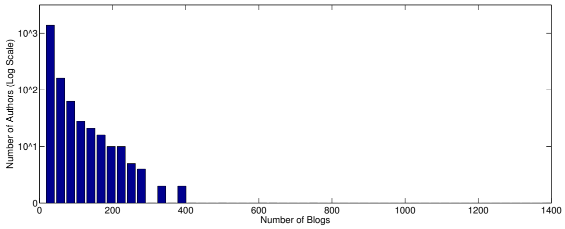

Figure 1. Distribution of number of blogs per author for all (1417) authors

Table 2 summarizes the statistics regarding authors and their corresponding blogs, who have written a minimum of 5, 15 or 30 blogs. As expected, as the minimum number of blogs per author increased, we got fewer authors, on average, they have written more number blogs.

	1	5	15	30
# Blogs	248246	153119	65945	41655
# Words	611M	58M	25M	15M
# Authors	56338	11926	1714	498
Avg. # Blogs	4	13	38	83
Median # Blogs	2	8	21	53

Table 2. Data characteristics for blogs in Tier 1

6. Methodology

Our methodology has following four steps:

1. Data Preparation:The Spinn3r dataset consists of blogs and their associated content in an XML format. We extracted personal stories from the above mentioned XML dump, using the information and code provided by the ICWSM website [9]. We extracted only those blogs whose authors chose to identify themselves with ⟨authorname⟩ tag within the XML mark-up. We removed a lot of software debris, and other XML markup in order to have only the content written by the author, which we refer to as the *description* of the blogs, i.e. everything that is written within the ⟨desc⟩ tags of the XML mark-up.

2. Feature Extraction: We extracted two sets of features from the descriptions:

(i) Stylistic Features, which characterize the text and capture the writing style of the authors. These features have been used in the past in different contexts as effective discriminators of authorship. They fall in numerous categories including lexical - like word count; syntactic - like the frequency of stop words; structural - like the paragraph length; and personal - like lexical diversity. To extract these features we computed different statistics from the description of the blogs so that in total we extracted 213 stylistic features detailed in Table 3. The stylistic features were calculated quantitatively from the data. Some of the features like the number of unique words, number of digits, letters, punctuation used were calculated by writing regular expressions to search and count the number of their occurrences in the texts. Vocabulary richness was calculated using a variant of the *Yule's K* function ($= \frac{M_1}{M_2}$), where M_1 s the number of all unique *stemmed words* in the text. The stemming was done using the standard *Porter's stemming algorithm* [12]. M_2 is calculated as $\sum_{i=1}^{K} f(s_i)^2$, where $f(s_i)$ is the number of times the i^{th} stemmed word occurs in the text. If a text has rich vocabulary, the above Yule's K measure for that text will be high.

Feature #	Description	Number
Length	Number of unique words/characters in blog	2
Vocabulary Richness	Yule's K	1
Word Shape	Frequency of words with different combinations of upper case and lower case letters	5
Word Length	Frequency of words that have 1-20 characters	20
Letters	Frequency of a to z, ignoring case	26
Digits	Frequency of 0 to 9	10
Punctuation	Frequency of punctuation characters	11
Special Characters	Frequency of other non-alphabet non-digit characters	21
Function Words	Frequency of special words like "the", "of", etc.	117

Table 3. Stylistic features

Except for vocabulary richness, number of unique words, and number of characters in the blog, all other stylistic features are ratios, or frequencies, i.e. each of the feature value is divided by the number of characters in the blog. This is done so that the feature values are not biased towards the length of the blogs written by a particular author.

(ii) Semantic Features (or thematic features) capture the themes running through the blogs, i.e. contexts or topics of that the author talks about in the blogs. To get these features, we applied the *Latent Dirichlet Distribution* (LDA) algorithm [3] (also known as *topic modeling*) on the description of the blogs to extract the dominant topics, or themes that pervade our collection of large, unstructured blogs data. We used the Mallet toolbox [13] for to achieve this. The tool takes as input a collection of texts, which in our cases are the blogs written by authors, and outputs the main topics, or themes that exist in the collection. The topics are thus a distribution over the words that exist in the collection. As an additional input, the tool also requires the number of topics. We tested with 20, 50, and 100 topics and on visually inspecting them, we found that 50 topics gave a good enough representation of the themes in the blog descriptions. Table 4 presents a listing of some of the topics and the top 19 words associated with the topics. A full list of the topics and top words are included in the supplementary material. The semantic features, corresponding to each blog, were the topic proportions of each of the 50 topics, i.e. probability values associated with each of the 50 topics.

In all, we used 213 stylistic features and 50 semantic features. For our analysis, we worked with the stylistic, and the semantic features both separately and combined (the two feature vectors were concatenated to create a new vector of 263 ($213 + 50$) features), to examine how performance was affected by the different classes of features. To reduce the computational complexity of the dataset, we used some heuristics to reduce the size of the dataset; in one set of experiments, we extracted a fixed number of authors from the entire corpus, while in another, we extracted a range or a fixed number of blogs written by a particular author.

3. Training a classifier: We performed experiments both for identification and authentication. For biometric identification, the problem of authorship attribution can be formulated as a multi-class classification problem, where we have the classes as the authors. For biometric authentication, the problem of authorship attribution can be for-

Topic #	Top words
4	night bar beer drink people party drunk drinking club drinks place wine good dance dancing friends guy pretty fun
7	mom dad family house sister home brother mother parents years year aunt time father daughter remember uncle day kids
30	water beach back lake day trip park river trail mountain road miles beautiful fish island found area boat hike
44	birthday party wedding year happy day christmas cake friends dinner fun made family love time gift great pictures night
47	doctor pain hospital blood sick baby feeling day appointment back surgery nurse told started days gave dr bad feel

Table 4. A sample of topics inferred from the blogs

mulated as a binary classification problem, with the two classes being genuine and impostor classes. For this study, we investigated language as a biometric trait primarily in the context of biometric authentication, where we posed the problem of authorship attribution as a binary classification problem. We can visualize a square matrix, whose rows/columns are the number of authors, where each cell is a score of matching N users with N enrollees. Enrollee i and user i is the same person, and enrollee i and user j ($i \neq j$) are different persons. Each user is matched against each enrollee once. Thus the diagonal elements correspond to genuine matches, and non-diagonal elements correspond to impostor matches.

To get the biometric score between a user and an enrollee, we used pairwise difference for all the features, for all blogs. This difference is used a biometric score, such that a low values between blog A and blog B signifies that both are written by the same author (a genuine match), while a large value between blog A and blog B signifies that they are written by different authors (an impostor). Since we have multiple blogs for each author, the genuine scores consist of blocks around the diagonal (unlike the traditional case where only the diagonal entries are the genuine scores).

Thus, for each data point we take the difference of individual feature values of this point from every other point. Since we already know the labels of the data-points, we know if the difference feature vector corresponds to a genuine or an impostor score. The difference feature vector is the same size of the feature vector length, i.e. 263. Thus, we create a dataset consisting of genuine and impostor data points, with 263 length feature vectors, labeled as genuine or impostors.

4. Evaluating Performance of Classifier We chose the classifier that gave the maximized the area under the ROC curve, which is our chosen performance metric. We worked with Support Vector Machine, K-Nearest Neighbor based classifier, and Logistic Regression on a very small subset of the data (40 authors) and found that *logistic regression* [2] gave the best results. Hence all experiments were performed with this classifier using the version implemented in Weka [10], as multinomial logistic regression model with a ridge estimator (set at 1.0E-8). Before training the classifier, the data was standardized i.e. all feature values were set to have zero mean and unit variance.

For the multi-class classification approach, during testing, each blog had to be attributed to one of the classes, however, on an internet-style author attribution dataset like ours, the number of authors is very large (order of 50K in this case), and the text written by a most of the authors was small (average is 4). Thus we got a large number of classes, with very limited number of data points for each class. To alleviate this issue, we simplified the problem by using a single unique signature for each author. We computed the median of the feature values for all the blogs of a particular author as the signature. Each new blog was then compared against this signature, and we took the nearest 3 neighbors that were closest to the median of each of the authors.

Rare-class Mining: Note that in the binary classification data that we create, there are only N genuine scores, and $N^2 - N$) impostor scores, where N is the number of authors. Thus, the result is a highly imbalanced dataset, with a lot more impostor data points than genuine ones. Typically, classification algorithms are designed under the assumption of a relatively uniform/balanced distribution of classes for training. But in a problem such as our genuine/impostor identification, there is a need to better balance the number of data-points in each class. This problem is referred to as *rare-class mining*, or *masking problem* which often leads to misleading results; the accuracy of the classification algorithms can be high even though it potentially misclassifies all or many of the points of the minority class. This problem can be handled by: synthetically over-sampling the minority class; under-sampling the majority class; and generating samples so that the resulting distribution of the two classes are balanced.

The over- and under-sampling methods artificially add or remove data to achieve balance, hence we adopted the more realistic approach of generating balanced sample sets. We used a Bernoulli distribution to randomly select data points that were added to the sample. We used two different Bernoulli distributions to select the genuine and impostor samples and the parameters of the distributions were set to ensure a balance between the samples of the two classes.

7. Results

7.1. Evaluating performance on different types of data

Table 5 compares the results of combining the data from all tiers. The results degrade when all the tiers of the data are used. We think the plausible reasons could be the following: The blogs are organized into various tiers by their

	AUC		
Data	Stylistic	Semantic	both
550-All tiers	0.60	0.59	0.61
550-Tier1	0.67	0.70	0.71

Table 5. AUC Scores when blogs from all tiers are used versus blogs from Tier 1

# of Authors	AUC		
(N)	Stylistic	Semantic	Both
200	0.61 (0.075)	0.65 (0.066)	0.65 (0.077)
550	0.67 (0.064)	0.70 (0.059)	0.71 (0.068)
All	0.66 (0.053)	0.68 (0.050)	0.70 (0.056)

Table 6. Classification results using logistic regression for various types of features with different number of authors (N). Standard deviation across multiples runs is reported in parenthesis

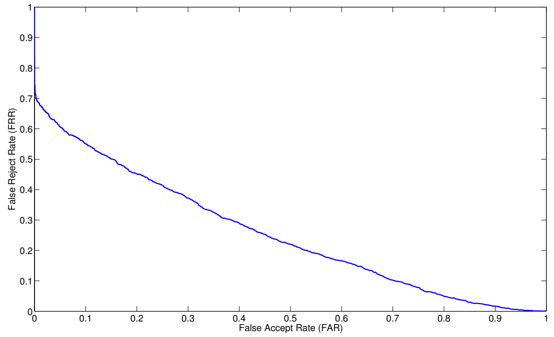

Figure 2. ROC curve (X-axis is the False Accept Rate, and Y-axis is the True Accept Rate) using logistic regression on Tier 1 data

influence score. Thus more influential blogs are in tier 1, and least in tier 13. We hypothesize that the quality of the text written by authors degrade as we go to less influential blogs, and it becomes increasingly difficult to discern authors. Additionally, we only consider the blogs written by an author who is identified with an author name in the XML dump of the dataset. It may happen that among various tiers, we may have different authors using the same author names, but in our analysis thus far, we assume that all blogs having the same author name were written by the same author.

7.2. Evaluating Performance of the Biometric Trait

For this experiment, we chose a random sample of N authors out of a total of 1714 authors (See Table 1) who had written more than 15 blogs. We created a data set of genuine and impostor samples as discussed above. We trained a logistic regression classifier on part of this data and tested the performance of the classifier on a held-out data set (34% of the original data set). This experiment was repeated 10 times to capture the variance in performance. We use *Area under the ROC Curve* (AUC) as the evaluation metric. Additionally, we experimented with three types of features for each setting: stylistic (See Table 3), semantic or topic-based (using 50 topics learned by LDA), and both combined. Table 6 shows the results obtained using different values of N (200, 550, and 1714). As evident from the table, the AUC does not vary significantly with the number of authors. With a smaller number of authors, the execution runs very fast, as the complexity of generating genuine samples is of the order of N, and the complexity of generating impostor samples is quadratic in N, but otherwise, the accuracy does not change significantly. In Figure 2, we report the ROC curve for one instance where the number of authors is 550.

7.3. Impact of Features

To gain additional insight into the impact of the types of features, we used **information gain** as a metric to measure the relative importance of the features. Figure 3 indicates that some of the most discerning features are one of the topics, blog size and frequency of some of the function words. The effectiveness of the function words is also reported in literature [14]. It is interesting to note that the topic with the most discriminating power in the feature space consisting of a topic which is mostly *adult* words, which co-occurred together a lot of times in the text. This signifies that there are authors who chose to write blogs on adult themes and make use of such words quite a bit. This buttresses the fact that stylistic features along with a semantic ones offer good metrics of discriminating authors. Notice from the AUC results in Table 6 that the accuracies are higher when both the feature sets are combined.

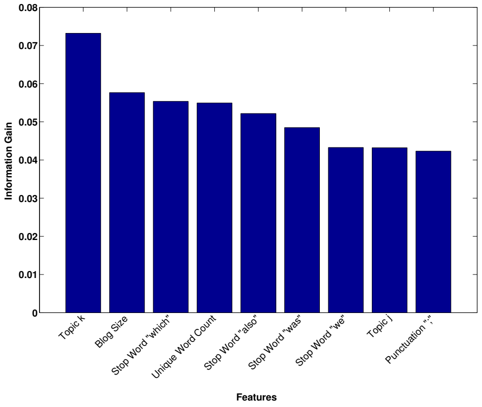

Figure 3. Top features using information gain

7.4. Using Authors with Different Minimum Number of Blogs

In this section, we investigate the question: *What is the minimum number of words (text) required to efficiently learn*

Min # of Blogs	AUC		
(k)	Stylistic	Semantic	Both
5	0.62 (0.003)	0.67 (0.006)	0.64 (0.018)
15	0.67 (0.069)	0.68 (0.063)	0.71 (0.072)
30	0.62 (0.076)	0.64 (0.092)	0.65 (0.091)

Table 7. AUC with different thresholds on minimum number of blogs written per author (k). Standard deviation across multiple runs is reported in parenthesis

# of Blogs	AUC		
(p)	Stylistic	Semantic	Both
5	0.73 (0.002)	0.75 (0.008)	0.75 (0.005)
15	0.65 (0.007)	0.67 (0.015)	0.69 (0.006)
30	0.60 (0.049)	0.62 (0.000)	0.64 (0.044)

Table 8. Classification results using logistic regression for different types of features with different number of blogs per author (p) used for training. Standard deviation across multiple runs is reported in parenthesis

to distinguish between genuine and impostor authors? To study this, we took authors who wrote at least 5, 10, or 15 blogs in Tier 1. In Table 7, we report the results using other values of k, where k is the minimum number of blogs written. The experiments were repeated 3-5 times to capture the variance in performance. The results indicated that 15 was the best number of blogs with which we could efficiently learn to distinguish between genuine and impostor authors. However, this heuristic is data dependent. The results may be poor at 30, because more number of blogs add more variance, and thus more confusion for the classifier. We require additional experiments to obtain a conclusive answer.

7.5. Varying the Number of Blogs Per Author

In this section, we investigate the question: *Given a fixed number of users (or authors), what is the minimum number of text required to train a biometric system such that it efficiently learns to distinguish between genuine users and impostors*. For this, we fixed the minimum number of blogs as 5, 15, and 30 for a comparative study. We calculated the number of authors who had written at least 30 blogs in Tier 1, which is 498 authors (out of 1714). We then randomly sampled p blogs ($p \leq 30$) for each author and generated the biometric data, where p was 5, 15, or 30.

The results in Table 8 suggest that if we have a fixed number of authors, then we may achieve a good AUC value, even with a smaller number of blogs per author. One possible reason is that increasing the number of blogs per author increases the variability per author, and thus tends to give poorer results. This is also data dependent. As we use more blogs per author for the same author, we add more variance in the blog writings, which may extend to both the style of writing and the thematic contents of the blogs.

7.6. Odd Man Out Analysis

We performed two types of analyses to evaluate how well the developed system generalizes to data from users who were not included in the training data.

We assume that we have trained our system to distinguish between genuine and impostor users. If we have new users that our system has never seen before, how effectively can the system detect whether the user is a genuine author (these authors were not included in the training data, but have been successfully enrolled in the system)? To evaluate this, we set aside a set of authors who were not used for training the classifier (*test authors*). For each blog written by a test author, we compared it with all the other blogs written by the same author Let the count of other blogs be denoted as n. We then counted the number of times our classifier successfully predicted these pairs as a genuine match, denoted as m. Ideally, the fraction $\frac{m}{n}$ should be greater than 0.5. Our classifier correctly classified **78%** of such matches as genuine.

Similarly, we compared each blog written by a test author with blogs written by another randomly selected test author where both sets of authors and their data have never been seen by the classifier. We then counted the number of times our classifier predicted these matches as impostors (m). Again, the fraction $\frac{m}{n}$ should be ideally greater than 0.5. Our classifier correctly classified **76%** of such matches as impostors.

This strongly indicates that our methodology has not just overfit to the styles of the authors used for training, but in general, has learned to a large extent, how to distinguish between an pair of writing styles, whether or not the writings originate from the authors that have been seen before by the classifier during training.

8. Discussions

This is also a study of biometrics in big data, i.e., when we have unprecedented amount of data available for biometric authentication and verification tasks. Big data implies that we not only need to address the challenges in volume, but also in diversity and uncertainty of the data. Our study deals with millions of blogs written by tens of thousands of authors. Here are some of our lessons-learned while working with such data:

1. We use a random sample of genuine and impostor data points, since the actual number of data points become prohibitively large to store and analyze,
2. Given that the analysis requires computing distance between every pair of blogs, the complexity of this authorship attribution is $O(N^2)$ in the number of blogs. To address this challenge, we used efficient data structures (such as *associative maps* in Java and Python) and scalable procedures involving large matrices.

3. To reduce the uncertainty in the data, we constrained the dataset, by taking only the blogs which are personal stories written by an identifiable author.

9. Conclusions and Future Work

We conclude that language can indeed be used a soft biometric, as it does hold some biometric fingerprint of the author. We report reasonable performance (72% AUC), even when the data consisted of unstructured blogs collected from across the internet. Our study indicates that blogs provide a diverse and convenient way to study about authorship on the internet. We found that better results are obtained with cleaner, high quality texts. We found that if number of authors are known, then even few texts per author would suffice to build a good classifier. However, the accuracy of the classifier is independent of the number of authors for the study. We also performed stricter testing, where our classifier was to correctly classify an unseen author. When classified genuine authors 78% of the time, and impostors 76% of the time. Obviously these conclusions are data-dependent, but provide an encouraging lead.

Regarding the issue of permanance, as long as the author maintains a specific writing style, this methodology will work. As our features are canonical in nature, they should be resistant to moderate changes in writing style and are expected to capture the variability in the nature of blogs. More work needs to be done to better understand permanence and spoof-ability. For the data set used in this study, we verified that multiple persons have not authored with same author name. It is difficult to ensure in the blogs data set, when a single person has written as multiple author names (she created profiles with different names). In that sense, our results are for the worst-case scenario.

The problem of author attribution can also be formulated as a multi-class classification approach, such that during testing, each blog has to be attributed to one of the known classes (authors). However, on an internet-style author attribution like ours, the number of authors is very large (order of 50K in this case), and the text written by most of the authors is usually small (average is 4). Thus we get a large number of classes, with very limited number of data points for each class. A simplistic solution can be devised in which each author is characterized by a signature which is obtained by combining the blogs written by that author. A new blog is then compared to all the available signatures and assigned to the author with most similar signature. Given that fusion of data instances of an enrollee into a signature is an open-area in biometrics, this is definitely an area of future work for our language biometrics paradigm.

An interesting extension to this research would be to work more closely with psycholinguistic community to investigate additional language-based features to more effectively capture the cognitive fingerprint of a person. With a large set of features to work with, we can employ feature selection algorithms to reduce the feature spaces and increase the area under the ROC curve. So far, we have only performed our evaluative study on Tier 1 and compared this with all the other tiers combined. However there needs to be a more detailed study on the other tiers individually to see how the statistics regarding authorship attribution vary with the tiers.

Acknowledgements

This work was supported by NSF Grant 1314803.

References

[1] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 2008.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* 2006.

[3] D. M. Blei, et al. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[4] K. Burton, et al. The ICWSM 2009 Spinn3r Dataset. In *In Proceedings of ICWSM*, 2009.

[5] D. Carroll. *Psychology of Language*. Cengage Learning, 2007.

[6] L. Faria, et al. Multimodal cognitive biometrics. In *CISTI*, 2011.

[7] A. Fridman, et al. Decision fusion for multi-modal active authentication. In *IT Professional*, 2013.

[8] A. Stolerman, et al. Active linguistic authentication revisited: Real-time stylometric evaluation towards multi-modal decision fusion. In *IFIP WG 11.9 ICDF*, 2014.

[9] A. S. Gordon and R. Swanson. Identifying personal stories in millions of weblog entries. In *ICWSM*, 2009.

[10] M. Hall, et al. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 2009.

[11] M. Koppel, et al. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 2009.

[12] C. D. Manning, et al. *Introduction to Information Retrieval*. 2008.

[13] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[14] F. Mosteller and D. Wallace. *The Federalist: Inference and Disputed Authorship*. Addison-Wesley series in behavioral science quantitative methods. 1964.

[15] A. Narayanan, et al. On the feasibility of internet-scale author identification. In *IEEE SSP*, 2012.

[16] L. Pearl and M. Steyvers. Detecting authorship deception: a supervised machine learning approach using author writeprints. *Literary and Linguistic Computing*, 2012.

[17] J. Pennebaker, et al. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 2003.

[18] E. Stamatatos. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 2009.

[19] Y. Zhu, et al. Biometric personal identification based on handwriting. In *ICPR*, 2000.