Cognitive-Biometric Recognition from Language Usage: A Feasibility Study

Neeti Pokhriyal, Kshitij Tayal, Ifeoma Nwogu, and Venu Govindaraju

***Abstract*—We propose a novel cognitive biometrics modality based on written language-usage of an individual. This is a feasibility study using Internet-scale blogs, with tens of thousands of authors to create a cognitive fingerprint for an individual. Existing cognitive biometric modalities involve learning from obtrusive sensors placed on human body. Our modality is based on the characteristic pattern of how individuals express their thoughts through written language. The problems of cognitive authentication (1:1 comparison of genuine versus impostor) and identification (1:n search) are formulated. We detail the algorithms to learn a classifier to distinguish between genuine and impostor classes (for authentication) and multiple classes (for identification). We conclude that a cognitive fingerprint can be successfully learnt, using stylistic (writing *style*), semantic (*themes*), and syntactic (*grammatical*) features extracted from blogs. Our methodology shows promising results (with 79% as the area under the ROC (AUC) in case of authentication). For identification, the individual class accuracies are up to 90%. We performed stricter tests to see how our system performs for unseen user, and report accuracies of 72% (genuine), and 71% (impostor). Such a study lays the groundwork for building alternative cognitive systems. The modality, presented here, is easy to obtain, unobtrusive and needs no additional hardware.**

***Index Terms*—Soft Biometrics, Novel Cognitive Biometrics, Large-scale Biometric Datasets, Class Imbalance, Multi-class classification**

I. Introduction

This work introduces a novel cognitive biometric modality, that learns a characteristic pattern of how users express and communicate their thoughts using written language.

There is an increasing need for novel biometric systems that engage multiple modalities. The notion of privacy is continuously evolving in todayś world. Users are increasingly storing immense quantities of personally identifiable data on cloud networks such as Google drives, Dropbox, iCloud, etc., they are also continuously and simultaneously logged into a number of devices and computing environments. Even though the user demands a continuously connected environment, s/he wants minimal interruption for any authentication information, while they perform unprecedented amount of private and secure communication over the console.

Neeti Pokhriyal, Dr. Ifeoma Nwogu, and Prof. Venu Govindaraju are with the Department of Computer Science and Engineering, University at Buffalo, State University of New York. The corresponding author is neetipok@buffalo.edu.

Kshitij Tayal is with Tata Consultancy Services, Hyderabad, India. The work was done while the author was a visiting scholar at Computer Science and Engineering, University at Buffalo, State University of New York

Manuscript received *****; revised *****.

The current methods of authenticating an individual requires the user to create and manage complex passwords. This is *unnatural*, and password based authentication methods are vulnerable to brute-force and dictionary based attacks. Once the user is authenticated into the session, typical systems don't employ mechanisms to verify the identity of the person, originally authenticated at the console. Such systems are vulnerable to attacks where the password and user's identity is stolen.

Active continuous authentication systems address this problem by continuously validating and verifying the identity of the user at console (who was initially verified by a physical biometrics or password), and can flag any deviations from it. A biometric modality that learns from the characteristic pattern of written language usage of an individual can be used to build such a system. Also, cognitive-biometric modality finds application in learning authorship attribution for anonymous texts on a large-scale.

Blogs are easily accessible, popular ways for expression of thoughts and communication between individuals. Also, using blogs to learn a biometric, makes this an unconstrained study, both in the quality and quantity of data. Advantageously, a system based on such modality requires no additional hardware or set-up costs. They are unobtrusive as a method of continuous authentication, which involves listening and verifying the signature of the person at console.

A. Definition of Cognitive Biometrics

We define cognitive biometrics as the process of identifying an individual through extracting and matching a characteristic signature based on the cognitive, affective, or conative state of that individual. Cognitive biometrics, like many behavioral traits, falls under the category of *soft biometrics*. Soft biometrics (e.g. handwriting, speech, gait etc.) are the characteristics of individuals that provide *some* information about them, but are not distinctive enough to efficiently differentiate among them, in contrast to hard biometrics like fingerprints, face, iris. However, soft biometrics modalities are easy to capture and process, compared to hard biometrics which are difficult and obtrusive to obtain.

Cognition can be defined as the process of acquiring knowledge, and understanding through thought, experience, and senses. Some of the cognitive modalities reported in literature involve the use of biological signals captured through electrocardiograms (ECG), electroencephalograms (EEG) and electrodermal responses (EDR) to provide possible individual-authentication modalities [1]. These are invasive and require special sensors, electrodes, to be placed on body parts.

As language is an important form of expression of one's thoughts and communication, we explore an alternative approach to cognitive biometrics through the language usage of an individual.

Recently researchers have started looking at written language usage as a biometric trait [2], [3], [4]. Our previous work [2] was a preliminary study performed on the ICWSM Spinn3r Dataset [5], where we had extracted a subset of authors and their blogs, and learnt a model for biometric authentication.

B. Contributions

We have provided a novel formulation of cognitive biometrics based on the written language-usage of an individual. Our feasibility study is supported by extensive experimentation with tens of thousands of authors, and Internet-scale blogs. Such a study lays the groundwork for building active continuous authentication systems that are unobtrusive, transparent and require no additional hardware, and authorship-attribution models at scale. We summarize the specific contributions of this work as follows:

- Diverse subsets of blogs, each with thousands of authors, are created for the study. To the best of our knowledge, this is the first Internet-scale study on the feasibility of learning a cognitive fingerprint from language-usage. The novelty of this work lies in the conceptualization of an alternative methodology to successfully capture the cognition of a person. The datasets created as a part of this study, and source code for cognitive authentication and identification are available at https://github.com/kshitijtayal/Cognitive_Biometrics.
- We provide a novel sampling technique to deal with the problem of class-imbalance (where data-points of one class, hugely outnumber the data-points belonging to other class). This problem is exacerbated in a large-dataset for cognitive authentication. Our sampling is based on generating Bernoulli random variables, one for each of genuine and impostor classes, and choosing a data-point belonging to either class. The outcome of this sampling ensures that both classes are balanced.
- We show how cognitive identification is posed as multi-class classification problem, with large number of classes, and limited data-points within each class. To deal with the challenge of fewer data-points per class, we provide a heuristic to increase the number of data-points per class, which can be applied to large-scale texts like blogs.

Additionally, some of the questions answered in this study are:

- How much data is needed per individual to learn his/her cognitive fingerprint ?
- Can cognitive modality, as described here, be used for identification task?
- How well do the different feature spaces perform in learning a cognitive fingerprint? What are their relative strengths and weaknesses?

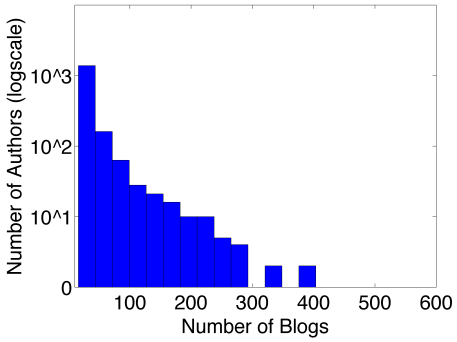


Fig. 1. The power-law distribution of blogs-per-author in our data. This portrays that lots of authors write few blogs whereas few authors write lots of blogs.

TABLE I
CHARACTERISTICS OF THE FOUR DATASETS

	Different Datasets			
Description	**Data-5**	**Data-10**	**Data-15**	**Data-20**
# Blogs	67764	30513	14884	7235
# Words	6785764	3033439	1485828	742604
# Authors	11440	3013	1043	409
# Avg words/author	593	1006	1424	1815

II. Data Description

The ICWSM Spinn3r[1] data was obtained by crawling various blog publishing sites to get a snapshot of social media, by collecting syndicated text of blogs and their associated contentss. It was collected from August to October 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.As with real datasets, much of the data is not *real* blog entries, as many are threaded online conversations, ads etc.

To get real text written by authors, we choose a subset of the Spinn3rDataset, called the Personal Stories Dataset [6]. This dataset consists of only the blogs which can be best characterized as a personal story. The logic, behind such decision, is our intuition that personal stories are expected to contain more distinguishing writing style markers.

We selected only those blogs that have an author-name in ⟨authorname⟩ tag in their XML mark-up, which acts as an identifier of the author. Next, 4 subsets of data are created. Each subset is named as Data-k, consisting of authors, who have written between k to k+5 blogs. We have Data-5, Data-10, Data-15, and Data-20 (See Table I). There are 11440 authors who have written between 5 and 10 blogs, 3013 authors who have written between 10 to 15 blogs, 1045 authors who have written 15 to 20 blogs, 409 authors who have written between 20 to 25 blogs, and 102 authors who have written about 100 blogs. This is owing to the long-tailed distribution of the number of blogs written per author within our dataset (Fig 1).

[1] http://www.icwsm.org/data/

III. Feature Descriptions

This section details how the three sets of features - Stylistic, Semantic and Syntactic, are extracted from the blogs.

*1) **Stylistic Features Description**:* The stylistic features capture the varied writing *styles* of the authors. These features are calculated quantitatively from the dataset. Table II details the 213 features extracted from the dataset. Features like the word shape, number of digits, letters, punctuation, special characters are calculated by writing regular expressions that search and count the number of their occurrences in the dataset. Vocabulary richness is calculated using a variant of the *Yule's K* function, as follows:

$$y_i = \frac{M_1}{M_2} \tag{1}$$

In Equation-1, y_i is the value of vocabulary richness of the i^{th} blog, and M_1 is the number of all unique *stemmed words* in the blog. M_2 is calculated as follows:

$$M_2 = \sum_{i=1}^{K} f(s_i)^2 \tag{2}$$

where $f(s_i)$ is the frequency of i^{th} stemmed word in the text. If a blog has rich vocabulary, its y_i is high. *Porter's stemming algorithm* [7] is used for stemming words in our code.

TABLE II
List of the 213 Stylistic features extracted from the blogs

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"	117

*2) **Semantic Feature Description**:* These features capture the context, or themes occurring in the blogs. While the stylistic features, capture *how* an author writes; semantic features capture *what* the author writes. We emphasize that authors may have different styles over varying context.

Since we wish to know the "topics" or themes pervading the blogs, we used *topic modeling* algorithms to extract them. Topic models are unsupervised algorithms that uncover hidden thematic structure in large volumes of text or documents [8].

Latent Dirichlet allocation (LDA) is a topic modeling technique, which can be thought of as mixed membership model, where each group or cluster exhibits different components in different proportions [9].

The generative process for LDA corresponds to the following joint distribution of the hidden and observed variables:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(z_{d,n}|\theta_d)(p(w_{d,n}|\beta_{1:K}, z_{d,n}))) \tag{3}$$

where $\beta_{1:K}$ are the K topics. Each β_k is a distribution over vocabulary V. The topic proportions for the d^{th} document are θ_d. The topic assignments for the d^{th} document are z_d. The observed words for document d are w_d.

As Dirichlet distribution is a conjugate prior for the multinomial, it simplifies the problem of statistical inference. The probability density of a T dimensional Dirichlet distribution over the multinomial distribution $p = (p_1, ..., p_T)$ is defined by:

$$Dir(\alpha_1, \ldots, \alpha_T) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} {p_j}^{\alpha_j - 1} \tag{4}$$

The parameters of the distribution are specified by $\alpha_1, ...\alpha_T$. The posterior for topic modeling is approximated as:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{w_{1:D}} \tag{5}$$

We used Gibbs sampling to approximate the posterior. For a given blog, w_d, the posterior topic distribution θ_d is represented as a K dimensional vector signifying the semantic features.

*3) **Syntactic Feature Description**:* These features capture the grammatical structure of the blogs, extracted using a natural language parser [10].

In this feasibility study, we used the parser based on Probabilistic Context Free Grammars (PCFG) [11]. The parser takes text as input and outputs the part-of-speech (POS) tagged tree. We chose this parser, because it provides a robust representation for each data-point. This is important as the blogs input to the parser are grammatically *loose*. scales well with the volume of our data.

IV. Methodology for Feature Extraction

This sections details how the three features were extracted for the feasibility study.

A. Stylistic Features Extraction

All stylistic features, except for vocabulary richness, number of unique words, and number of characters in a blog, are ratios. These are calculated by dividing each feature value by the number of characters (or words) 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.

From each dataset, we constructed a matrix M, where each row, r, corresponded to a single blog written by an author, and each column, c, corresponded to each of the 213 Stylistic features extracted for each blog.

B. Semantic Features Extraction

We extract topical features using LDA, which uses the blog corpus to learn a probabilistic topic model, and outputs the topics that exists [12]. The topics are a distribution over the vocabulary of the corpus. We experimented with 10, 50, 75, 100 topics, and empirically determined 50 as the optimum number of topics as it maximized the posterior probability of the given data.

For each blog, the semantic features are characterized by a fifty dimensional feature vector, that signify the probability distribution associated with each of the 50 topics.

Tables IIIa, IIIb, IIIc, IIId present a listing of some of the topics and the top 19 words associated with them for each of Data-5, Data-10, Data-15, and Data-20 respectively. We observe similar topics emerging in all the datasets, as they are subsets of a bigger corpus.

C. Syntactic Feature Extraction

The syntactic feature extractor is described as Algorithm 1. Each blog ($B_i \in blogs$) is treated as a set of sentences ($\{s\} \in B_i$). The output of the algorithm is a count matrix F consisting of $|blogs|$ rows (one per blog) and p columns, where p is the number of possible syntactic pairs in the corpus. For each sentence, a parse tree is constructed (Algorithm Line 4). Each leaf node (words and punctuations occurring in s_{ij}) of the parse tree (Algorithm Line 6) is represented as a tuple $\langle A, B\rangle$ where A and B are syntactic categories. A is the parent of B in $\mathcal{T}$ and B is the parent of the word. The corresponding entry in F is incremented by one (Algorithm Line 7). Eventually, the entry $F[ij]$ indicates the frequency of the j^{th} syntactic pair for the i^{th} blog.

We row-normalize matrix F, to account for the varying length of the blogs written by various authors. We have around 300 unique syntactic category pairs extracted from the dataset. Since, not all pairs are frequent, F is a sparse matrix.

As an example, for given s = "Words are the only things which last forever.", parse tree produced by the Stanford Parser is depicted in Figure 2a, with tag abbreviations explained in Table 2b. Table IV shows the count for sentence s.

Algorithm 1: Extracting syntactic features from blogs

Input: $blogs$ tokenized at period(.) as sentences
Output: $\mathcal{F}$

1 $\mathcal{F} \leftarrow []$
2 **foreach** $B_i \in blogs$ **do**
3 **foreach** $s \in B_i$ **do**
4 $\mathcal{T} \leftarrow ParseTree(s)$
5 **foreach** $\mathcal{L} \in leaves(\mathcal{T})$ **do**
6 $(A, B) \leftarrow (Parent(\mathcal{L}), Grandparent(\mathcal{L}))$
7 $\mathcal{F}[i][\langle A, B\rangle] \leftarrow \mathcal{F}[i][\langle A, B\rangle] + 1$
8 **end**
9 **end**
10 **end**
11 **return** $\mathcal{F}$

Our parser is very robust to grammatical errors, i.e. for any input sentence it comes up with the most likely sentence analysis. This is important when working with diverse data as blogs. It has a computational complexity of $(O(n^3))$, where n is the number of words in the sentence.

TABLE IV
SYNTACTIC CATEGORY PAIRS WITH THEIR FREQUENCIES FOR THE SAMPLE SENTENCE

Syntactic Pairs	Frequency
(NP,NNS)	2
(VP,VBP)	2
(NP,DT)	1
(NP,JJ)	1
(WHNP,WDT)	1
(ADVP,RB)	1

V. METHODOLOGY FOR COGNITIVE AUTHENTICATION AND IDENTIFICATION

This section details how the problems of cognitive authentication and identification are formulated.

A. Cognitive Authentication

The cognitive authentication is formulated a binary classification problem, with genuine and impostor classes. If two blogs were written by the same author, they are a genuine match, else an impostor. Algorithm 2 describes authentication.

The dataset consisting of genuine and impostor samples is created as follows. First, we built a matrix, R, such that each row was a data point r of the form (A_i, B_i), where A_i is the i_{th} author, and B_i is the i_{th} blog of the A_i author. Next, we extracted stylistic, semantic and syntactic features from each r (i.e. each blog written by an author) (See Step 3 in the Algorithm). Let's say the length of corresponding features were y, m, and n. Each r in R was represented by a $l = y + m + n$ length vector.

Next, we created a distance matrix, D, from R, where each row d measured how distant each row i was pairwise to every other row j in R. The distance of i, and j was calculated as point-wise absolute difference of each of the stylistic, semantic and syntactic features, as shown in line 11 in the Algorithm 2. As i and j were both l length vectors, thus d was also a l length vector, where each entry was given as $d_n = abs(i_n - j_n)$, where $1 <= n <= l$.

If, both i and j in R, belonged to the same author, then the corresponding d belonged to genuine class, else d belonged to impostor class. The distance matrix, D, encoded the fact that if two blogs were written by the same author, they were be *more similar* to each other, signified by lower difference values, and thus a genuine match, depicted in line 13. If they were written by different authors, they are more dissimilar, and thus an impostor match, shown in line 18.

Analyzing the computational complexity, if there were a authors, with b blogs, then there were $a * b^2$ genuine data-points. The impostor points were bounded by $a^2 * b^2$. As $a >> b$, $a^2 * b^2 >> a * b^2$. Thus impostor data-points hugely outnumbered the genuine scores, resulted in an imbalanced dataset.

TABLE III
A SAMPLE OF TOPICS INFERRED USING LATENT DIRICHLET ALLOCATION ALGORITHM FOR THE VARIOUS DATASETS. AS EACH TOPIC IS A DISTRIBUTION OVER WORDS IN THE VOCABULARY. SHOWN BELOW ARE THE TOP WORDS FOUND IN A SAMPLE OF FOUR TOPICS IN OUR 4 DATASETS.

(a) Data-5

Top words for 4 topics for Data-5
1. sick pain doctor hospital blood feeling hurt body appointment bad stomach sore surgery leg worse felt throat eye vet
2. computer back find problem found internet finally decided set laptop system problems fix online working figure issues hard due
3. weather hot rain cold water air sun day warm snow storm morning summer wind sky heat fall wet power
4. pictures love picture photos camera beautiful made great art photo amazing lovely wonderful perfect pics originally absolutely loved gorgeous

(b) Data-10

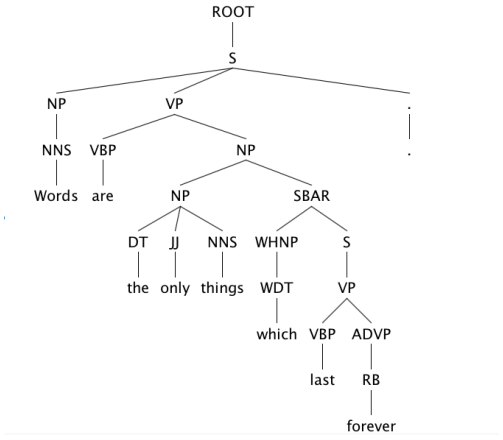
Top words for 4 topics for Data-10
1. yesterday today doctor hospital test appointment taking blood care gave problems surgery fine dr news worried pain
2. job work computer working money pay internet office company problem paid worked laptop interview bank finally find works manager
3. hot weather rain cold sun air cool warm power nice summer afternoon wind fire morning storm beautiful heat sky
4. ago time pictures years couple picture camera weeks photos long taking months days finally photo shot originally pics shots

(c) Data-15

Top words for 4 topics for Data-15
1. felt feel bad feeling sick today started yesterday tired thing makes fine horrible bit crying guess upset stomach tonight
2. computer problem internet decided find thing site check hard service laptop online set problems case completely mail reason figure
3. cool hot weather rain cold nice yesterday sun warm fair lovely wind afternoon pretty low high storm raining sunny
4. part photos full photo point group shot trip san showed experience shots level mentioned club ten st changed shoot

(d) Data-20

Top words for 4 topics for Data-20
1. yesterday today sick head pain doctor bad feeling heart worse appointment cold hospital mouth nose sore barely blood worried
2. computer post write blog real haven internet make update life posted hard business entry quick laptop word problems written
3. ride water rain weather bike sun air lake blue cold warm power tree storm beautiful rode trees clear sky
4. time pictures big made camera picture lot wedding photos photo art beautiful cute fair experience lovely hard taking married



(a) Parse Tree for a sample sentence

Abbreviations	Full Form
NP	Noun Phrase
VP	Verb Phrase
S	Simple declarative clause
SBAR	Clause introduced by subordinating conjunction
ADVP	Adverb Phrase
NNS	Noun,Plural
VBP	Verb, non-3rd person singular present
DT	Determiner
JJ	Adjective
WHNP	Wh-noun Phrase
WDT	Wh-determiner
RB	Adverb

(b) Abbreviations of the Syntactic Category Pairs for the sample sentence

1) Handling of Class Imbalance: Typically, classification algorithms are designed under the assumption that the classes are balanced (i.e. having similar number of data-points per class) for training. In an imbalanced dataset the accuracy of the classification algorithms can be high even if they misclassify all or many of the points of the minority class as majority class. This is called the **masking problem**, which leads to misleading results. The problem of class imbalance is exacerbated in big datasets, with thousands of authors.

This problem is handled by either synthetically oversampling the minority class, or under-sampling the majority class. We followed a different approach - by using a Bernoulli distribution to randomly select data points to be added to the sample. We used two different Bernoulli distributions to select the genuine and impostor samples. The parameters of the distributions were set to ensure a balance between the samples of the two classes. The probability density function associated with Bernoulli's distribution is:

$$P(n) = p^n(1-p)^{(1-n)} \tag{6}$$

where $n = 1$ (success) occurs with the probability p, and $n = 0$ (failure) occurs with probability $1 - p$.

The value of p lies between 0 and 1, and is empirically determined for each class. p is set high (0.01) for genuine class, as there are fewer genuine samples to begin with, and we want to retain more samples. For impostor class, we want more samples to be rejected, hence p was set to 0.001. This method is efficient, as we do not need to store all the samples for each class.

B. Cognitive Identification

To study the feasibility of cognitive identification, we pose it as a multi-class classification problem, where each class corresponds to an author.

The subsets of data (namely Data-5, Data-10, Data-15, and Data-20) contains a large number of classes (order of thousands), and too few data points in each class (order of tens). This makes the problem of multi classification very challenging.

Algorithm 2: Training Phase for Cognitive Authentication

Input: $authors, blogs$
Output: $\mathbf{w}$ (classifier weights)
// Feature Extraction
$F \leftarrow []$; // Feature matrix
foreach $B_i \in blogs$ **do**
  $F[i,:] = extractFeatures(B_i)$
end
// Training Data Creation
$T \leftarrow []$; // Training data matrix
$l \leftarrow []$; // Training labels vector
$cnt \leftarrow 1$
foreach $B_i \in blogs$ **do**
  foreach $B_j \in blogs$ **do**
    if $bernoulli(p) == 1$ **then**
      $T[cnt,:] = |F[i,:] - F[j,:]|$
      if $author(i) == author(j)$ **then**
        $l[cnt] = 1$
      end
      else
        $l[cnt] = 0$
      end
      $cnt \leftarrow cnt + 1$
    end
  end
end
// Logistic Regression Training
$\mathbf{w} \leftarrow logisticTrain(T, l)$
return $\mathbf{w}$

1) Creating the Dataset: We created a new subset of the blogs data, called Data-100, which consisted of authors who wrote more than 100 blogs. There are 102 such authors. The average number of blogs per author is 210.The combined feature space of stylistic, semantic and syntactic features is about 500 dimensional long. Having 210 data points per class exacerbates the problem in high dimensional space. Thus, we focused on the stylistic features, which are 225 dimensional long. Remember we still have 210 data-instances per class.

Multi-class classification with large number of classes, and few data-points per class is an active area of research. Some proposed methods work under varying assumptions, and for different types of data.

Here, we provided a technique, that worked with unstructured text data on the web. We intended to increase the number of data-points within a class. We tokenized each blog at periods(.), to get its constituent sentences. On average each blog consisted of 26 sentences. Next, each data-point could correspond to a single sentence, or to a group of sentences.

We need to find the optimal number of sentences, that should designated as a data-point, our metric is getting high classification accuracy. There was a trade-off here - assigning fewer sentences as a data-point does not provide our classifier with enough discriminatory power, while designating too many sentences as a single data-point, reduced the number of instances per class. We empirically worked with 1, 3, 5, 6, 7 sentences per data-point, and found 6 to be an optimal choice. Thus, the first data-point for an author corresponds to the first 6 sentences trimmed from the blog, the next data-point corresponds to the next six sentences, and so-on. Using this technique, we get an average of 525 data-points per class.

Algorithm 3: Training Phase for Cognitive Identification

Input: $authors, blogs$
Output: $\mathcal{T}$ (classifier)
// Training Data Creation
$T \leftarrow []$; // Training data matrix
$l \leftarrow []$; // Training labels vector
foreach $B_i \in blogs$ **do**
  $T[i,:] = extractStylisticFeatures(B_i)$
  $l[i] = author(i)$
end
// Random Forest Training
$\mathcal{T} \leftarrow randomForestTrain(T, l)$
return $\mathcal{T}$

There are 102 classes.

2) Multi-Class Classification: The procedure for identification is given in Algorithm 3. The stylistic features constituted the feature space. As known, multi-class classification is performed as one-class-versus-all, or as one-class-versus-another. In the former, the classifier learns N binary classifiers, where N is the number of classes. In the later, the classifier learns $N(N-1)/2$ binary classifiers. We performed one-versus-All classification, because one-versus-one is computationally expensive, and biometric modality is typically used in scenarios, where one is interested in distinguishing one user from the rest.

VI. Experiments

This section details the experimental set-up and the algorithms used to perform authentication and identification.

A. Cognitive Authentication

We experimented with various binary classifiers. Logistic regression classifier with a ridge estimator (set at 1.0E-8), gave the best Area Under the ROC curve (AUC). Briefly, binary logistic regression assigns the probability of the target class to be 1 given the input vector as:

$$P(Y_i = 1) = \frac{1}{1 + e^{\boldsymbol{\beta}^\top \mathbf{X}_i}} \tag{7}$$

where $\boldsymbol{\beta}$ is the weight vector and $\mathbf{X}_i$ is the set of explanatory variables associated with observation i. The weight vector $\boldsymbol{\beta}$ was learned from a training data set using a standard gradient descent based optimizer to maximize the likelihood of the data. Before training, the data was standardized with zero mean and unit variance. The experiments were conducted using Weka [13].

B. Cognitive Identification

We used various classification algorithms, like SVM, and Decision Trees. Random Forest classifier gave the best results[14]. It is an ensemble classifier, that fits a number of classifiers i on various sub-samples of the dataset [15]. When a test instance arrives, each tree gives its own prediction $p_i(x)$ for a class. The forest, then, takes the majority of the predictions of all the trees, as shown below:

$$f(x) = \sum_i p_i(x) \tag{8}$$

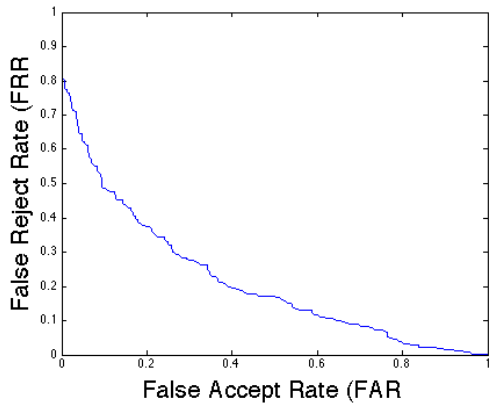


Fig. 3. ROC curve for Data-10

where $p_i(x)$ = 1 if i^{th} tree predicts true, and -1 otherwise. If f(x) $\geq$ 0, then prediction is 1, and -1 otherwise.

We experimented with 500 trees in the forest using python sklearn.They are robust to over-fitting, and run efficiently on large-scale data.

VII. Results and Discussion

This section details the results for authentication and verification experiments.

A. Cognitive Authentication

We trained a logistic regression classifier, and tested its performance on 34% of held-out data.

Owing to the diversity of the blogs and authors, we performed multiple runs of the experiments and report their average, with standard deviations across the runs mentioned with the reported accuracies for all datasets in Table V. Experiments were conducted using stylistic features alone(column 1), semantic features alone(column 2), syntactic features alone(column 3), combined stylistic and syntactic features (column 4), combined stylistic and semantic features (column 5), all features combined (column 6), the top 300 features selected using feature selection technique (column 7). Since syntactic and semantic features consistently performed poorer than stylistic features, we did not combine them for our results.

We use Area under the ROC Curve (AUC) (1 - FRR versus FAR) as the evaluation metric, where FRR is the False Reject Rate, and FAR is the False Accept Rate. AUC represents the probability that random genuine score is higher than random impostor score. The AUCs for each dataset: Data-5, Data-10, Data-15, Data-20 were 78.7%, 78.9%, 77.2%, and 76.3%. The ROC curve expressing the trade-off between FAR and FRR scores for the best performing dataset, Data-10's is shown in Fig 3.

1) Discussion: We observe the following:

- All features combined captured the variance in the blogs to efficiently distinguish between genuine and impostor classes, than using individual feature space(s).

TABLE VI
Top ten features selected using Information Gain for Data-10

Features	Information Gain
Topic i	0.072657
Number of Characters	0.07052
Freq of 'y'	0.065011
Freq of 'p'	0.064141
Freq of ','	0.062328
Freq of lowercase words	0.061428
Freq of 'h'	0.061223
Freq of words of length 6	0.05963
Topic j	0.059176
Freq of 'l'	0.05888

- Stylistic features performed better than either semantic or syntactic features, conforming with existing literature [16].
- Syntactic features performed poorer than other two feature spaces. We attribute this to the fact that these features are very sparse, language dependent and depends on the reliability of the parser [17].
- Semantic features performed poorer than stylistic, but better than syntactic. This stresses the fact that authors write about *similar* topics (like family, job, relationships) in the blogosphere, which are not as discriminatory as stylistic features of the writings.
- Stylistic features, when combined with other two feature spaces, boosted the performance of cognitive authentication.

Table VII describes the pros and cons of using three distinct feature spaces for cognitive authentication.

2) Feature Selection: : We ranked our features using information gain, which measures the decrease in entropy of a class, c_i when a feature, f_j is present, versus when it is absent.

$$IG(c_i, f_j) = H(c_i) - H(c_i/f_j) \tag{9}$$

where, IG refers to the information gain, and H refers to the entropy. This gives us the most discerning features, as shown in Table VI. The top feature, given as "Topic i" is the topic with *unusual/rare* words. This is intuitive as authors, who use *unusual/rare* words can be easily differentiated from authors who don't. The features dealing with the number of characters, the frequency of lowercase words, and words of length 6, as shown in Table VI, constitute another set of differentiating features. This reiterates the discriminatory power of stylistic features based on the *length* of the text. Another topical feature makes it to the top ten discerning features, mentioned as "Topic j" in Table VI. This topic consists mostly of *adult* words, stressing that authors who write about such themes in their text uses them quite often, than authors who don't. The rest of the features in Table VI are frequencies of less frequently used characters in English language. This shows that stylistic features along with a few semantic ones offer good metrics of discriminating authors.

B. Odd Man Out Analysis for Cognitive Authentication

We performed stricter tests to see how our system generalizes to users it has never seen before. We assumed that our system is trained for Biometric Authentication, and wanted

TABLE V
AUCs REPORTED FOR DIFFERENT DATASETS AND WITH VARIOUS SETTINGS OF THE FEATURE SPACE. THE HEADING FOR EACH COLUMN MARKS THE TYPE OF FEATURES USED IN PARTICULAR EXPERIMENT. THE COLUMN, TITLED - ALL, SIGNIFIES A FEATURE SPACE WITH A CONCATENATED SET OF STYLISTIC, SEMANTIC AND SYNTACTIC FEATURES, AND THE LAST COLUMN, CONSISTS OF TOP 300 FEATURES SELECTED USING INFORMATION GAIN CRITERIA. STANDARD DEVIATION ACROSS MULTIPLE RUNS IS REPORTED IN PARENTHESIS

Data (Number of Authors)	AUC's obtained with the following settings of feature space for each Dataset						
	Stylistic Features	Semantic Features	Syntactic Features	Stylistic + Syntactic Features	Stylistic + Semantic Features	All Features	Top Features
Data-5 (11440)	0.744 (0.022)	0.708 (0.011)	0.635 (0.021)	0.738 (0.021)	0.764 (0.019)	0.766 (0.021)	0.787 (0.024)
Data-10 (3013)	0.745 (0.018)	0.670 (0.004)	0.665 (0.010)	0.760 (0.020)	0.763 (0.013)	0.770 (0.012)	0.789 (.057)
Data-15 (1043)	0.705 (0.005)	0.661 (0.010)	0.641 (0.016)	0.735 (0.013)	0.729 (0.009)	0.755 (0.008)	0.772 (.015)
Data-20 (409)	0.708 (0.014)	0.639 (0.015)	0.608 (0.009)	0.725 (0.014)	0.723 (0.024)	0.748 (0.015)	0.763 (.017)

to test how effectively can it identify a new unseen user as genuine? To evaluate this, we set aside some (say k) randomly chosen authors, who were not used for training the classifier, called them *test authors*. For each blog written by a test author, we matched it with his/her other blogs, to create a genuine match. Let the count of genuine matches be denoted as n. Next, we saw how many times does our trained classifier predict each of them to be a genuine match. Let the number be denoted by m. Ideally, the fraction $\frac{m}{n}$ should be greater than 0.5. Our classifier correctly classified **72%** of such matches as genuine.

In the second scenario, we randomly selected two authors, and studied how well our system identified an impostor attack when both the users were unseen. For this, we matched all the blogs written by one test author, with the blogs written by another randomly selected test author. Let the number of impostor matches be n. Both the authors and their data had never been seen by our classifier. We then counted the number of times our classifier predicted these matches as impostors (m). Again, the fraction $\frac{m}{n}$ should be greater than 0.5. Our classifier correctly classified **71%** of such matches as impostors.

This indicates that our methodology efficiently learned to distinguish between users based on their language-usage, and did not just over-fit the data.

C. Results for Cognitive identification

Figure 4 reports the average accuracies for One-versus-all scenario for 102 classes. About one-third of the authors were identified with more than 70% accuracy, stressing the fact that our feature space is robust and discriminative to efficiently learn a boundary for each class. We obtained about 65% (on average), with individual class accuracies reaching up to 90%. For instance, Figure 5 shows the distribution of classification scores for one of the authors. It is evident from the distribution for the two classes that the author is distinguishable using our feature space.

Such results are data-dependent but strongly imply that our methodology learns a cognitive fingerprint of an author which is successful in discerning him/her from the others.

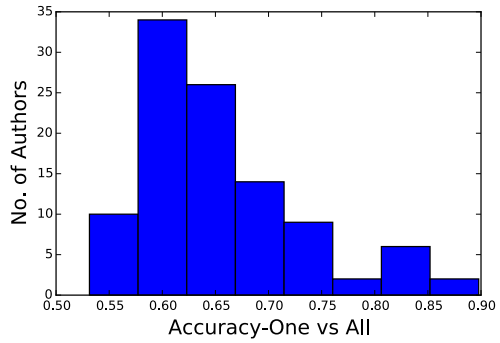


Fig. 4. Histogram of classification accuracy for 100 authors

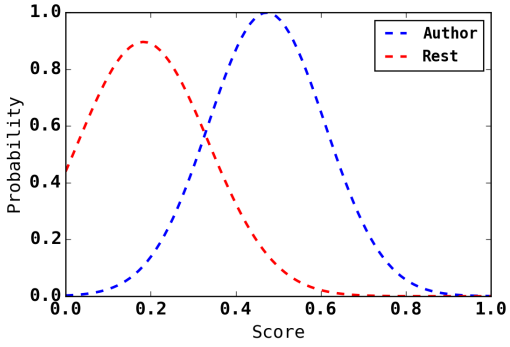


Fig. 5. Distribution of classification scores for one author vs. rest. The curves are obtained by fitting a Gaussian distribution to the scores for each class. The classification accuracy for the one randomly chosen author, displayed, here, is 80%, with a F1 score of 0.51.

VIII. RESEARCH BACKGROUND

The problem is authorship-attribution is defined as - given a certain author, how well can we attribute whether a not-previously-seen piece of writing can be attributed to that author. The problem is studied in different guises with datasets of different types. Stylometric analysis techniques have been used for attributing authorship [18], [19], [16]. The varied

TABLE VII
COMPARISON OF VARIOUS FEATURE DOMAINS

Feature Type	Pros	Cons
Stylistic	- Empirically derived. - Scale well to incorporate additional features and more data. - Easy to extract by writing rules, or formats.	- Some of the features are too simplistic, which inhibits their understanding.
Semantic	- The models used to generate these features can be updated to incorporate varied additional information, like correlation among topics, author information.	- The LDA algorithm assumes a bag of words model. However, they can be updated to include a dependence among words. - Efficient algorithms need to be used for larger datasets. - There is limited validation (using measures like perplexity) for the topics extracted.
Syntactic	- Probabilistic parsers produce the most likely grammatical structure of the sentence, and are apt to study unstructured text (like blogs).	- Dependent on the reliability of the robust parsers. - Computing and memory intensive operations.

databases for such study include: theater plays [20]; essays [21]; biblical books [22]; book chapters [23], [24]); emails, chats and SMS messages [25], [26], [27], [28], web forum messages [29], [30]; blogs [31], [32].

A related problem in the domain of online privacy or anonymity is to unmask an anonymous blogger/whistleblower [33]. Plagiarism detection is another variant, where portions of new writings are compared against large bodies of published works. These are more related to the use and arrangement of words than to extract cognitive features. Authorship deception identification is another variant, relevant in cybercrime forensic domain. It aims to detect when an author actively imitates another author's writing in order to conceal his/her true identity [34].

Various types of features explored in related works include lexical, syntactic, structural, stylometric and content features. Additional features include relative frequency of words, character n-gram, word n-grams, part-of-speech n-grams and vocabulary richness. The classification algorithms used include naive Bayes, neural-networks, K nearest neighbor [27].

Though authorship attribution traditionally dealt with few numbered texts, recently researchers have started looking into large-scale texts like blog, movie-review databases, but they report poor accuracies for these cases, while still working with ≈ 1000 authors [35].

IX. LIMITATIONS, CONCLUSION AND FUTURE WORK

We conclude that written language-usage of an individual can provide his or her cognitive fingerprint, based on our large-scale feasibility study. Ours is an open-set unconstrained study based on tens of thousands of authors and millions of blogs. We detail the methodology to extract varied feature sets from the blogs, and perform cognitive authentication (1:1 match) and verification (1:N search). We report high accuracies of 77% for authentication, and as high as 90% for verification.

Our methodology is robust, as depicted by very low values of standard deviations among the various runs of the experiments. It is scalable, as our accuracies do not degrade even when the number of authors is scaled to thousands.

Regarding the issue of time-variance, 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 capture the variability in the blogs. A longitudinal study is underway to confirm this.

An obvious use-case of cognitive-biometric is continuous authentication, where the identity of the user at console, can be actively monitored, and any deviations from it can be flagged. Continuously monitoring the text written at the console by a user, can be learnt as a cognitive fingerprint. It will be interesting to study how disparate textual modalities like emails, can be incorporated with the existing methodology to expand on the cognitive fingerprint. We intend to see how meta-data, such as Google profiles, associated with large-scale texts, can build a better cognitive fingerprint of an individual. Some interesting applications to explore include building authorship attribution systems at scale, cognitive-biometric driven captchas, and mobile friendly authentication mechanisms.

ACKNOWLEDGMENT

This work partially supported by the National Science Foundation under Grants IIP #1266183 and CNS #1314803.

REFERENCES

[1] L. Faria, V. Sa, and S. de Magalhaes, "Multimodal cognitive biometrics," in *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on*, June 2011, pp. 1–6.

[2] N. Pokhriyal, I. Nwogu, and V. Govindaraju, "Use of language as a cognitive biometric trait," in *IEEE International Joint Conference on Biometrics, Clearwater, IJCB, 2014, FL, USA, September 29 - October 2, 2014*, 2014, pp. 1–8.

[3] A. Fridman, A. Stolerman, S. Acharya, P. Brennan, P. Juola, R. Greenstadt, and M. Kam, "Decision fusion for multi-modal active authentication," in *IT Professional*, July 2013.

[4] A. Stolerman, A. Fridman, R. Greenstadt, P. Brennan, and P. Juola, "Active linguistic authentication revisited: Real-time stylometric evaluation towards multi-modal decision fusion," in *International Conference on Digital Forensics*, january 2014.

[5] K. Burton, A. Java, and I. Soboroff, "The ICWSM 2009 Spinn3r Dataset," in *In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA, May 2009.

[6] A. S. Gordon and R. Swanson, "Identifying personal stories in millions of weblog entries," in *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*, San Jose, CA, May 2009.

[7] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[8] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012. [Online]. Available: http://doi.acm.org/10.1145/2133806.2133826

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[10] E. Charniak, "Statistical techniques for natural language parsing," *AI Magazine*, vol. 18, pp. 33–44, 1997.

[11] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003.

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

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[14] L. Breiman, "Random forests," pp. 5–32, 2001.

[15] T. G. Dietterich, "Ensemble methods in machine learning." London, UK, UK: Springer-Verlag, 2000, pp. 1–15.

[16] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, Mar. 2009.

[17] M. Gamon, "Linguistic correlates of style: authorship classification with deep linguistic analysis features," 2004.

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

[19] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 1, pp. 9–26, Jan. 2009.

[20] T. C. Mendenhall, "The characteristic curves of composition," vol. ns-9, no. 214S, 1887, pp. 237–246.

[21] G. Yule, *On Sentence-length as a Statistical Characteristic of Style in Prose: With an Application to Two Cases of Disputed Authorship*. Biometrika Office, University College, London, 1939.

[22] D. L. MEALAND, "Correspondence analysis of luke," *Literary and Linguistic Computing*, vol. 10, no. 3, pp. 171–182, 1995.

[23] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Mining the blogosphere: Age, gender and the varieties of self-expression." *First Monday*, vol. 12, no. 9, 2007.

[24] M. Koppel, J. Schler, and E. Bonchek-Dokow, "Measuring differentiability: Unmasking pseudonymous authors," *J. Mach. Learn. Res.*, vol. 8, pp. 1261–1276, Dec. 2007.

[25] O. de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD Rec.*, vol. 30, no. 4, pp. 55–64, Dec. 2001.

[26] M. Koppel and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution," in *IN IJCAI03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003, pp. 69–72.

[27] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 7:1–7:29, Apr. 2008.

[28] S. Ishihara, "A forensic authorship classification in sms messages: A likelihood ratio based approach using n-gram," *International Journal of Speech Language and the Law*, vol. 21, no. 1, 2011.

[29] A. Abbasi and H. Chen, "Applying authorship analysis to arabic web content," in *Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics*, ser. ISI'05, 2005, pp. 183–197.

[30] T. Solorio, S. Pillay, S. Raghavan, and M. Montes-Gomez, "Modality specific meta features for authorship attribution in web forum posts," in *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 2011, pp. 156–164.

[31] M. Koppel, J. Schler, S. Argamon, and E. Messeri, "Authorship attribution with thousands of candidate authors," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06, 2006, pp. 659–660.

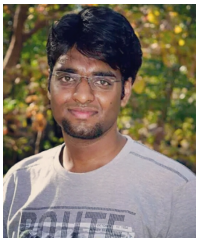
[32] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83–94, 2011.

[33] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *Security and Privacy (SP), 2012 IEEE Symposium on*, May 2012, pp. 300–314.

[34] L. Pearl and M. Steyvers, "Detecting authorship deception: a supervised machine learning approach using author writeprints," *Literary and Linguistic Computing*, vol. 27, no. 2, pp. 183–196, 2012.

[35] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship attribution with topic models," *Comput. Linguist.*, vol. 40, pp. 269–310, 2014.

Neeti Pokhriyal Neeti Pokhriyal is a Ph.D student in CSE Dept, University at Buffalo, SUNY under the guidance of Prof. Venu Govindaraju. Her research interests include Data Mining and Machine Learning. She has her Masters in CS from University of California, Riverside, where she received Dean's Distinguished Fellowship.



Kshitij Tayal Kshitij Tayal completed his Masters (Gold Medalist) in Computer Science from University of Hyderabad, India. He was also affiliated with the Institute for Development and Research in Banking Technology, Hyderabad. Currently, he works with Tata Consultancy Services, Hyderabad, India.

Dr. Ifeoma Nwogu is a Research Assistant Professor at Center for Unified Biometrics and Sensors (CUBS) and Center of Excellence for Document Analysis and Recognition (CEDAR), since Oct'11. The centers are affiliated with the CSE Dept at the University at Buffalo, SUNY. She finished PhD in 2009 at UB, and was a NSF-sponsored postdoctoral researcher at CSE Dept, University of Rochester. She has a Master's in Computer and Information Science at the University of Pennsylvania.

Prof. Venu Govindaraju , SUNY Distinguished Professor of Computer Science and Engineering, is the founding director of the Center for Unified Biometrics and Sensors. He is a Fellow of the ACM (Association of Computing Machinery), IEEE (Institute of Electrical and Electronics Engineers), AAAS (American Association for the Advancement of Science), the IAPR (International Association of Pattern Recognition), and the SPIE (International Society of Optics and Photonics).