BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/06 pp163–191
Volume 5, Number 3, September 2022
DOI: 10.26599/BDMA.2022.9020002

Estimating Intelligence Quotient Using Stylometry and Machine Learning Techniques: A Review

Glory O. Adebayo and Roman V. Yampolskiy*

Abstract: The task of trying to quantify a person's intelligence has been a goal of psychologists for over a century. The area of estimating IQ using stylometry has been a developing area of research and the effectiveness of using machine learning in stylometry analysis for the estimation of IQ has been demonstrated in literature whose conclusions suggest that using a large dataset could improve the quality of estimation. The unavailability of large datasets in this area of research has led to very few publications in IQ estimation from written text. In this paper, we review studies that have been done in IQ estimation and also that have been done in author profiling using stylometry and we conclude that based on the success of IQ estimation and author profiling with stylometry, a study on IQ estimation from written text using stylometry will yield good results if the right dataset is used.

Key words: stylometry; IQ estimation; authorship attribution; intelligence; IQ; author profiling; machine learning

1 Introduction

Intelligence testing has been around for many centuries, albeit known by many different names and used in many different forms[1]. For 3000 years, the Chinese have used mental tests and the Imperial courts established tests in the seventh and eighth centuries that are like the tasks on today's tests[2]. However, formal studies of the intelligence date back to the early 20th century. The first widely used intelligence test, the Simon-Binet intelligence scale, was developed by Alfred Binet and Theodore Simon in France in 1905[3]. The test was established after the French government commissioned Binet to develop an instrument to identify school kids that needed extra teaching classes. Psychologist, Lewis Terman has since revised the test with American subjects, and it is known as the Stanford-Binet intelligence scale. Also, during the First World War (WWI), the military deployed the use of aptitude tests to help the military commanders to measure the ability of their personnel.

1.1 Motivation

IQ scores have been used in the past to identify an individual's ability to succeed in educational situations. In the mid-20th century, standardized tests and IQ tests were used by schools to place students into tracks (Tracking was a way of grouping students in classes and offering them courses in academic subjects that reflected the differences in the students' prior learning). Currently, intelligence quotient is still used; IQ scores or some forms have standardized testing and have sometimes been used by employers to estimate how an individual would perform in the workplace[4] especially if the candidate has no prior work experience. Also, the military uses IQ testing to determine where to place recruits in the army as well as government intelligence agencies and law enforcement agencies[5]. In this paper we review papers that have addressed IQ estimation using machine learning and/or stylometric analysis. The review shows that there are not a lot of works done in the field of IQ estimation using either machine learning or stylometry. We also review some papers that have used stylometry to estimate age, gender, nationality,

• Glory O. Adebayo and Roman V. Yampolskiy are with the Department of Computer Science and Engineering, University of Louisville, Louisville, KY 40208, USA. E-mail: glory.adebayo@louisville.edu; roman.yampolskiy@louisville.edu.

* To whom correspondence should be addressed.

Manuscript received: 2021-06-27; revised: 2021-12-03; accepted: 2022-01-21

personality type, and native language. The aim of this paper is to identify current works done in IQ estimation using machine learning and stylometry and also to show that stylometry can yield good results when used in IQ estimation due to its success in estimating age, gender, nationality, personality type, and native language.

1.2 Stylometry

Stylometry is the statistical analysis of differences in writing style between authors. It is a study of the linguistic properties in a body of text by analyzing various text features. Stylometry combines various research fields (e.g., statistics, linguistics, and computer science) and is applied in various areas ranging from academic research to forensic evidence collection. One of the earliest examples where stylometry was implemented using computers was the identification of the disputed papers among the "Federalist papers"[6]. Tweedie et al.[6] showed that stylometric analysis when applied in the domain of authorship identification, was able to arrive at similar conclusions about the authorship of these papers as other works that have been done in the domain. Recently, stylometric analysis was used to identify chat bots in Ref. [7]. Further research was done in Ref. [8] to show that the stylometric approach becomes difficult when the bot changes behavior overtime. Also, Yampolskiy et al.[9] showed that stylometric author identification processes can be applied on a single author that can write in multiple languages.

1.3 Intelligence

Intelligence is the ability of a person to learn from experience and to adapt to, shape, and select environments[10]. The work of Charles Spearman was one of the first modern studies of intelligence. He scientifically studied intelligence and proposed that: "intelligence could be understood in terms of a general ability that pervaded all intellectual tasks, and specific abilities that were unique to each particular intellectual task"[10].

1.4 Intelligence quotient

Intelligence quotient in its early days represented a measurement concept that was used in intelligence testing[2]. Basically, it was a numeric score obtained from an intelligence test. In 1914, a German psychologist, William Stern, introduced the notion of a mental quotient by suggesting that the index of intellectual functioning derived from the Binet-Simon intelligent scales could be expressed as the ratio of the test taker's mental age to their chronological age and multiplied by 100 to eliminate decimals as seen in Eq. (1) below.

$$IQ = \frac{MA}{CA} \times 100 \tag{1}$$

where MA is mental age which is obtained from taking an intelligence test, and CA is chronological age which is the measured age of a person from birth to a given date. Over the years, IQ scores have been generally used to identify an individual's capability to succeed in educational situations. Čavojová and Mikušková[11] argued though that there is a weak correlation between cognitive abilities and final evaluation in some given courses (social psychology) but they also showed that participating in voluntary extra-curricular activities was a better predicator of academic achievement. The one major flaw of Čavojová and Mikušková[11] research was that it was focused mainly on a population that was limited to psychology students training to be future teachers and they all exhibited an average IQ at best.

IQ estimation using stylometry, machine learning, or both methods is an emerging area of study with the earliest study coming in 2015 when Wang et al.[12] proposed a model framework to estimate IQ from an MRI (magnetic resonance imaging) dataset using machine learning methods. Currently, intelligence quotient still plays a major role in the society. Employers sometimes use IQ testing in the hiring process of applicants without previous working experience. IQ scores are also sometimes used to estimate how an individual would perform in the workplace. The industrial psychology literature has also agreed on an explanation for the strong relationship between IQ and job performance: individuals who have showed high level of IQ can easily learn job relevant knowledge faster and better than others which helps them to perform better at their jobs[4].

The rest of the paper is organized as follows. We talk about the methodology of the literature review in Section 2. In Section 3 we do an overview of IQ estimation from neuroimaging data. Section 4 covers an overview of IQ estimation from written text. In Section 5 we do an overview of stylometry and its applications in author profiling. And in Sections 6 and 7, we give our discussions and conclusions on the review which also include suggestions on steps to take for future research.

2 Methodology of Research

We used a systematic methodology to search for literature for this study. Originally developed and

implemented by Badar et al.[13], we modified the methodology and came up with our own methodology. The methodology included the following stages.

- Defining the research problem.
- Building a pool of articles and papers that have done extensive or related work in our research area.
- Reading and extracting relevant data/information from our pool of articles and papers.
- Review of the quality of extracted data/information.

All the papers and articles reviewed in this study were retrieved by conducting exhaustive and extensive search on Google Scholar†.

All papers and articles including journal and conference papers that have been published in the database and related to our study area have been included in this review. Key words that have been used to perform the search include "IQ estimation", "IQ estimation using stylometry", "IQ estimation with machine learning", "IQ estimation from written texts", "stylometry", "author attribution", "author identification", "automatic IQ estimation", and "authorship identification".

2.1 Inclusion criteria

We searched for articles and papers using key words and only literature that provided information relating to IQ estimation using any form of machine learning and stylometry was selected. Also, some papers and articles that provided information on the use of stylometric methods as means of estimating social or personal traits (gender, age, nationality, character type, native language, etc.) were selected.

2.2 Exclusion criteria

One of the main focuses of this survey is to carry out a study on the use of machine learning to estimate an individual's IQ, therefore papers and articles that have done a study on estimating IQ but without using any form of machine learning or stylometric analysis have been excluded.

2.3 Selection of papers

All literature used went through an exhaustive and vigilant scrutiny of all inclusion and exclusion criteria. The papers that remained in the final pool of selected papers were downloaded in pdf format and saved. The naming convention employed depicted the title of the paper. This helped to easily index the articles and papers in the final pool of selected papers. The citations of all the articles were also downloaded and saved in a Mendeley library.

† https://scholar.google.com/.

3 IQ Estimation from Neuroimaging Data

This is also known as brain imaging. A method that uses brain imaging is an experiment technique that studies the structure or functions of a human or animal brain and should ideally produce accurate timing (usually in functional imaging) and spatial localization (in both functional and structural imaging) as it relates to cerebral functions, structure, or the changes in these brain properties[14]. These methods are usually minimally invasive and should be repeatable to allow its use in treatment monitoring and the development of therapeutic strategies. These methods are the most commonly used to find areas in the brain where either the functional response or structural measurement can be predicted by experimental or demographic variable[15]. With the emergence of MRI, there has been substantial understanding of the neurological basis of intelligence and correlations between intellectual performance and multiple neural parameters (grey matter[16], white matter[16], fractional anisotropy[17], cortical thickness[18], functional connectivity[19], and genetic effects[20]) have been published. There are three commonly used neuroimaging methods.

Magnetic resonance imaging. This is a non-invasive imaging technology that is used to produce a three-dimensional detailed image of human or animal anatomy and it is often used for treatment monitoring and disease detection and diagnosis. MRIs employ the use of power magnets with strong magnetic fields and sophisticated technology to excite and record the change in direction of the rotational axis of the protons found in the water that makes up living tissues and physicians are able to tell the difference between the various types of tissues based on the magnets. When used on the brain, MRIs can differentiate between white and grey matter and can be used to detect tumors. Wang et al.[12] and Arya and Manuel[21] employed the use of an MRI dataset, autism brain image data exchange (ABIDE), to estimate IQ. Jiang et al.[22] obtained MRI scans from subjects using a Tesla magnetic resonance scanner.

Electroencephalography (EEG). This is a specialized test that detects the electrical activity in the human brain by using small, metal discs known as electrodes that are attached to the scalp. This is since

brain cells are active all the time (even when asleep) and are always in communication with each via electrical impulses. It is one of the main diagnostic tests for epilepsy and can also aid in diagnosing other brain disorders. The brain activity is recorded as wavy lines. Firooz and Setarehdan[23] explored the use of a dataset of recorded EEG readings while taking a cognitive test.

Positron emission tomography (PET). This is an imaging test that reveals the functionality of tissues and organs in a human or animal body. It employs the use of a radioactive drug (tracer) to show activity.

3.1 Magnetic resonance imaging

The earliest application of machine learning methods to IQ estimation was when Wang et al.[12] proposed a novel framework to estimate IQ using MRI data. They used a feature selection method that was based on an extended dirty model for jointly considering element wide sparsity and group-wise sparsity. One of the major challenges they faced was the absence of a large dataset. They solved this problem by integrating multiple datasets scanned from different sources with different scanning parameters and protocols. A two-set procedure was designed for experimenting on the dataset.

The first step was to identify possible scanning source of each testing subject and the second step was to estimate the IQ of the testing subject by using a specific estimator designed for the scanning source. Two experiments to test the performance of their method were performed by using the MRI data collected from 164 typically developing children between 6 and 15 years old. For the first experiment (feature selection), they employed the use of a multi-kernel support vector regression (SVR) for estimating IQ values, and an average correlation coefficient of 0.718 and an average root mean square error of 8.695 were obtained between the true IQs and the estimated ones. The brain regions that were selected have been reported in previous studies to be highly associated with cognitive ability and memory. For the second experiment, a singlekernel SVR was used for IQ estimation, and this achieved an average correlation coefficient of 0.684 and an average root mean square error of 9.166. These results proved the effectiveness of using MRI to estimate IQ.

Arya and Manuel[21] proposed a system that classified intelligence quotient of an individual into one of the four classes of the Weschler adult intelligence scale (WAIS). The four classes used were very superior, superior, high average, and average. This assumed that very few people can get an IQ score of less than 70 and greater than 140 because IQ scores fit the normal distribution where most of the IQ values are near or around the average (100). The feature extraction and classification of features related to IQ were done using convolutional neural network (CNN). Figure 1 shows a diagram of the proposed system.

The MRI brain image dataset used was autism brain image data exchange (ABIDE) provided by neuroimaging informatics tools and resources (NITRC) which is also the dataset used by Wang et al.[12] It comprised of 3D MRI brain scans. The data preprocessing for this study was done in three steps.

Skull stripping. This step involved the process of brain tissue segmentation from the surrounding region. This was done in MRIcro using the brain extraction technique (BET). This step returned an image with the removed non-brain matter.

Slice extraction. This step involved the extraction of images of the required brain slices from the images in

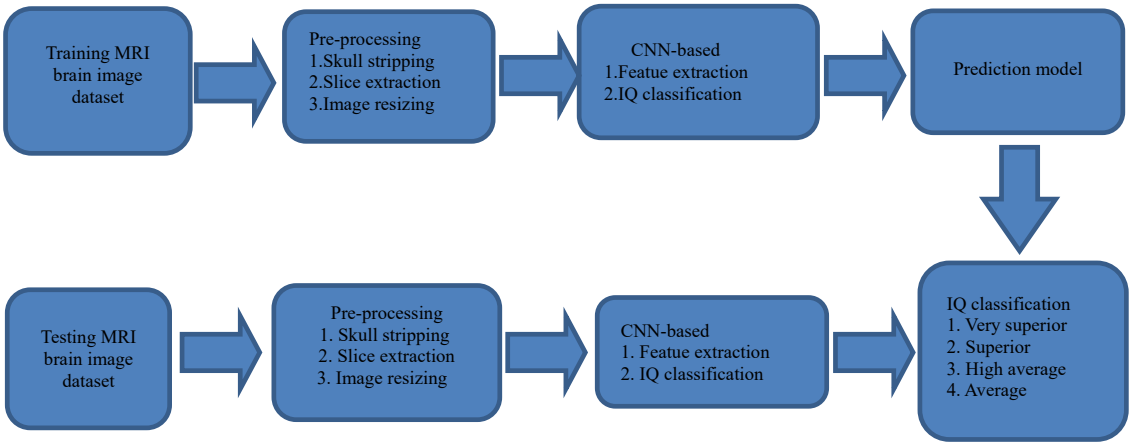

Fig. 1 Proposed IQ classification system[21].

the steps above. The brain slices were taken from three different views of the brain (sagittal view, coronal view, and transverse view).

Image resizing. The step involved resizing the images to meet the size requirement of the CNN.

As stated earlier, a novel framework was proposed for the classification of the IQ using neuroimaging features. The classification of the IQ was done using three different CNN architectures.

Smaller visual geometry group (SVGG). This CNN architecture consists of 5 convolution layers, 3 Max pooling layers, one fully connected layer, and two dense layers. The image input dimension for this architecture is 96×96×3.

Visual geometry group 16 (VGG16). This a neural network that has been proposed by Simonyan and Zisserman[24]. It has 16 layers, 13 convolutional layers, and 3 fully connected layers. It has also been pretrained. The image input dimension for this architecture is 224×224×3.

Residual network (ResNet-50). This is an architecture that is 50 layers deep (48 convolution layers, 1 Maxpool layer, and 1 average layer).

5000 bi-dimensional slices from each of the three brain views were fed into the three CNN architecture with 80% of the dataset randomly selected with each class being assigned the same number of images. The remaining 20% is used as the testing dataset. The results obtained can be seen in Table 1.

The results showed that ResNet-50 was more accurate in predicting IQ than the other two architectures with a maximum accuracy of 85.9%. Also, using the images from the sagittal view proved to yield the best results. This study showed the application of deep learning with MRI data to classify IQ by leveraging the influence of neural parameters on the level of human intelligence. The study also proved to be one of the only existing studies to detect an individual's IQ using the physiological structure of the brain.

The aim of Jiang et al.[22] was to predict IQ scores quantitatively using the functional connectivity (FC) based on brainnetome atlas (meaning) using a prediction framework that incorporated advanced feature selection and regression methods. "The brainnetome atlas will be an in vivo map, with a more fine-grained functional brain subregion and detailed anatomical and functional connection patterns for each area, which could help researchers to describe the locations of the activation or connectivity more accurately in the brain."[25]

The subjects used for this study comprised of a total of 360 healthy college students with age ranging from 17–24 and a mean age of 19.41 ± 1.09 years. There were 174 females and 186 male Han Chinese subjects. They were all interviewed using the structural clinical interview (meaning) to make sure that none of them had Axis I mental illness. They were also screened for neurological diseases (traumatic brain injuries) and for any family history of psychiatric disorder. Their IQ was measured using the Chinese version of the Wechsler adult intelligence scale-revised by China (WAIS-RC). The IQ scores obtained ranged between 74 and 132 with a mean IQ of 109.65±11.27. The MRI scans were performed on the subjects using an MR750 3.0 Tesla magnetic resonance scanner manufactured by GE healthcare. The MRI data were preprocessed using data processing assistant for resting state fMRI advanced edition (http://rfmri.org/DPARSF).

In neuroimaging data, feature dimension tends to overwhelm sample size, but feature selection helps to simplify fitted models and helps them to be easily interpreted by reducing overfitting. A nested-leave-one-out cross-validation (LOOCV), which used the outer loop to estimate the prediction accuracy and the inner loop to determine the best optimal selection number. Figure 2 shows that in the outer loop, one sample is set as the testing set and the remaining $N - 1$ samples are used as the training set (where N is the total number of samples in the dataset).

The feature selection stage was done using ReliefF algorithm which assigned a weight to every training FC feature with a weight value which statistically accounts for its relevance to the predicted measure. They derived

Table 1 Accuracy comparison of results obtained[21].

Number of 3D brain images for each class	Number of MRI slices form individual images	Total number of slices for training	Total number of slices for testing	CNN architecture used	Transversal image accuracy (%)	Sagittal image accuracy (%)	Coronal image accuracy (%)
50	25	4000	1000	SVGG	51.625	61.0	58.7
50	25	4000	1000	VGG16	54.500	73.0	68.8
50	25	4000	1000	ResNet-50	66.800	85.9	76.4

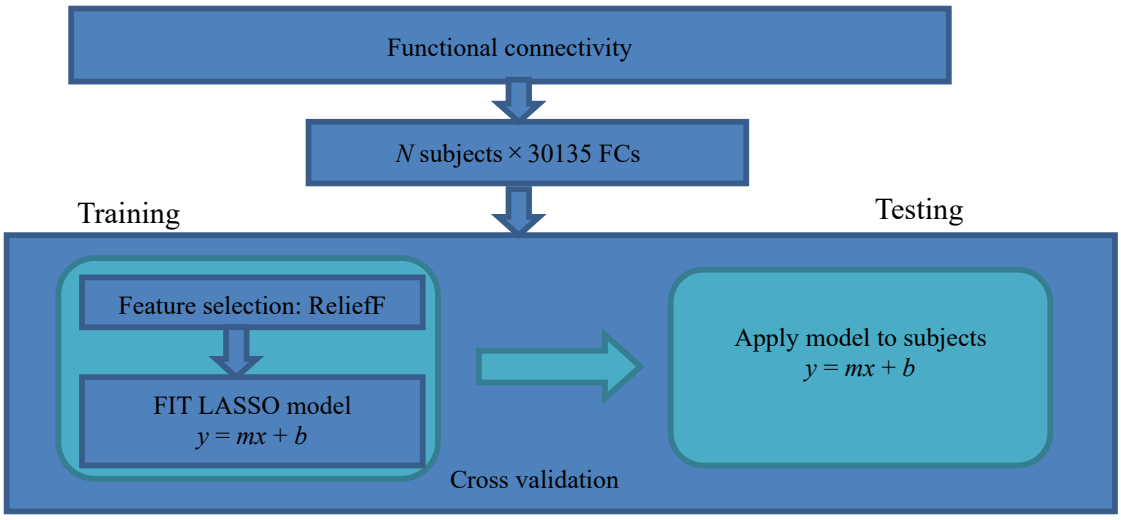

Fig. 2 Proposed framework showing feature selection and regression analysis using Brainnetome Atlas based functional connectivity[22].

a reduced number of m top weighted features by determining a certain parameter m. The IQ scores were then estimated using the selected FC features by using multiple regression methods on the training dataset and the testing data passed through the resulting regression model to generate a predictive score and test the accuracy of the resulting model. A total of five linear regression algorithms were implemented which include least absolute shrinkage and selection operator (LASSO), ridge regression, the elastic net regularization, relevance vector regression (RVR), and ordinary least regression (OLS). The loop was repeated N times to test all the subjects and, on each iteration, the predicted IQ score (for the left-out sample), the selected FC features, and the regression coefficients in the prediction model were recorded.

Additionally, the IQ scores of male and female samples were predicted and recorded because previous studies have published the sex difference in the neurobiology of functional connectivity (FC)[26]. The experiment was also repeated without using feature selection. The results obtained can be seen in Table 2 below.

The results obtained showed that ReliefF + LASSO produced the best results and yielded the most promising results when predicting the IQ of female samples. An average of 150 FC's was identified in each of the LASSO regression models because of the variable selection of sparse regression. Also, a total of 8 and 15 FC's were repeatedly by each of the loops for male and female samples, respectively.

3.2 EEG-fNIRS

The intelligence of individuals can be attributed to the structural and functional differences of the brain and intelligence is the ability to learn and understand concepts[23]. One of the very interesting areas of psychophysiology is investigating what happens in the brain when performing logical-mathematical intelligence tests. To discover the nature of the cognitive procedures going on the brain during problem solving various brain mapping systems can be used including electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS). To investigate these functional differences in the human brain as a means of estimating IQ, Firooz and Setarehdan[23] used a binomial system consisting of fNIRS and EEG. They studied the oxygenation and the electrical activity in the brain when a subject takes the Raven's progressive matrices

Table 2 Prediction results of five regression models (regression coefficients) from proposed framework[22].

Algorithm	Prediction result		
	All samples	Male samples	Female samples
ReliefF+LASSO	0.5122	0.4682	0.7212
ReliefF+ridge	0.4787	0.3010	0.4918
ReliefF+elastic net	0.4313	0.2787	0.6481
ReliefF+RVR	0.2189	0.1353	0.2359
ReliefF+OLS	0.3157	0.1468	0.3161
LASSO	0.3668	0.1678	0.6802
Ridge	0.4345	0.1815	0.4295
Elastic net	0.3449	0.1830	0.6655
RVR	0.2413	0.0859	0.2609
OLS	–0.0213	–0.1294	0.1076

(RPM) intelligence test — a nonverbal test that is usually comprised of 60 questions used in measuring the abstract reasoning of a subject. It is usually estimated as a fluid test. Positron emission tomography (PET) and EEG studies have also suggested a negative correlation between intelligence and brain activity[27,28] but fMRI showed a positive correlation according to new studies[29,30]. It is also worth noting that fNIRS is a valuable tool for illustrating the relationships between the functions of the cortical region in the brain during cognitive activities. The primary aim of the study was to examine the neuronal activities (EEG) and the local hemodynamic response of the brain (fNIRS) while it was in the process of performing a logical test. The second objective of the study was to estimate the IQ of a subject irrespective of the results the subject obtained from the RPM intelligence test. They looked to rectify the problems of previous studies which showed that stress might have been a major factor affecting test takers during the test.

They solved this problem by removing the time limitation and allowing the subjects answer as many questions as they could. Also, question difficulty was randomized as opposed to the difficulty arranged in ascending other. Fatigue level was monitored with these adjustments and the results showed that stress levels were reduced for the subjects.

The dataset used for this study was done in two phases. In the first phase, eleven (11) healthy Persian-speaking graduate students (5 males and 6 females) between the ages of 24 and 30 years old (27.8 ± 3 years) were subjected to the Cattle IQ test under normal conditions to get a base IQ and were split into two groups of low intelligence ($IQ < 120$) and high intelligence ($IQ \geqslant 120$). This revealed a balanced dataset with 6 low intelligence subjects and 5 high intelligence subjects. In the second phase, the subjects were made to take the modified RPM intelligence test and their fNIRS and EEG readings were being recorded. After this, fNIRS signal was processed signal by signal and the baseline drift was removed using linear regression analysis. The brain system is generally a non-linear system and to address the problem of applying linear approaches to a non-linear system, the brain was considered as a chaotic or quasi-chaotic system. Chaotic properties such as fractal dimensions (FD) were applied to describe non-linear time series like EEG. Linear discriminant analysis (LDA) and principal component analysis (PCA) were used for the feature selection process and PCA performed better with an information loss of less than 4%. Linear regression and support vector regressions were used as the learning methods. SVR was used because it is an optimal method for small scale regressions. The experiments were carried out with fNIRS and EEG features combined and with fNIRS features alone using various combinations of the artificial features that were generated using PCA and LDA. A leave-on-out test (one sample was exempted from the training data and used for testing) was used due to the small size of the dataset and eventually training was done on the entire dataset so that they could present an accurate model to estimate the IQ of new observations. The results showed that using a combination of fNIRS and EEG features obtained from PCA with the leave-one-out approach yielded the best results. It was also concluded that the stress level of the subjects was decreased significantly.

4 IQ Estimation from Written Text Using Stylometry

Stylometry is the statistical analysis of differences in writing style between authors by analyzing various text features. Previous studies in IQ estimation from written texts and in actual fact, authorship attribution proposed different classification of features to quantify the writing style of an individual[31]. Some of these classifications include lexical features, character features, syntactic features, and semantic features, and in some cases, certain features that are application specific have been identified. For example, an html based corpus where features like font color counts or font size counts might need to be defined[32]. However, the current review of text representation features for stylistic purposes is focused mainly on the computational requirements for measuring them[33]. For example, lexical and character features consider a text as just a sequence of word tokens and characters, respectively, while syntactic and semantic features require deeper linguistic analysis and application — specific features can only be defined in certain text domains or language domains.

Estimating IQ from written texts is an emerging new area of interest with the earliest work done in this area coming in 2017 when Hendrix and Yampolskiy[34] proposed the use of stylometric analysis to estimate an individual's IQ by analyzing the number of SAT words that are presented in a body of text. Evidence showed that the ratio of SAT words in a corpus of writing samples is roughly a bell curve and is normally distributed with an obvious left skew.

4.1 Corpus and methodology

Hendrix and Yampolskiy[34] proposed a hypothesis that hitched on comparing the curve of the collegiate word ratio (CWR) — the ratio of the total count of collegiate words (words SAT consider a part of strong vocabulary usage) used in a written text to the total count of words in the text, of sample texts with more than 100 words from the common crawl corpus as shown in Eq. (2) — with the curve of IQ scores across the entire population. It was seen that both curves where normally distributed though there is a slight left skew in the CWR curve.

$$Collegiate\ Word\ Ratio = \frac{Collegiate\ Word\ Count}{Total\ Word\ Count} \quad (2)$$

Working off of Hendrix and Yampolskiy's[34] hypothesis, Abramov and Yampolskiy[35] considered more features that could be used to expand the research into IQ estimation.

• **Lexical aptitude ration (LAR).** LAR can be defined as the equivalent of the CWR that was employed by Hendrix and Yampolskiy[34]. Given a text sample of length N, LAR can be defined as

$$LAR = \frac{CountDistinct(W)}{N}, \quad W \in D \quad (3)$$

where D = SAT vocabulary, and W = words.

• **Lexical diversity (LDMTLD).** This is the measure of the unique words that are used in a text. They identified that this measure shows high sensitivity to text length so to solve this problem they introduced the MTLD measure to reduce the effect of text length. MTLD is defined as the mean length of sequential word strings in each text that maintains, a given type-token ratio.

• **Syntactic complexity (SYNNP).** This is simply the syntactic structure of a sentence and they used the Coh-Metrix SYNNP index (measures the mean number of modifiers per noun-phrase) to measure this.

• **Meaningfulness (WRDMEAc).** This is measured by rating words based on a meaningfulness rating corpus. Words that are highly associated with other words get a high meaningfulness rating compared to words that are weakly associated with other words.

All four features listed above were computed using the NLTK and Coh-Metrix tool for all the text samples in the training set. The Open American National Corpus was used as the training set as opposed to the common crawl corpus that Hendrix and Yampolskiy[34] used as the training set. Some similarities exist between the studies carried out by Hendrix and Yampolskiy[34] and Abramov and Yampolskiy[35](Since Abramov and Yampolskiy's study was based on expanding Hendrix and Yampolskiy's study, it was inevitable). Asides from the obvious differences in corpus, corpus preprocessing (Abramov preprocessed the corpus to exclude poorly written/constructed text samples), the test dataset, and additional features that Abramov explored, the stylometric technique and hypothesis of both studies were essentially the same as would be shown later.

Hendrix and Yampolskiy[34] presented a method to automatically estimate the IQ of an individual by calculating the Z-score of each data point in the dataset as defined as

$$Z\text{-}Score = \frac{CWR\ Data\ Point - CWR\ Mean}{CWR\ Standard\ Deviation} \quad (4)$$

The Z-score represented the number of standard deviations the data point was from the mean of the entire training dataset either it was positive or negative. The IQ of the data point was then estimated by using Eq. (5) below.

$$Calculated\ IQ = (Z\text{-}Score \times IQ\ std) + IQ\ Mean \quad (5)$$

Abramov and Yampolskiy[35] employed the same technique as Hendrix and Yampolskiy[34] but instead identified that the text features listed above showed enough match between the index's and IQ score's normal distributions. Using this information, a method was proposed to calculate and estimate the IQ of an individual using each of the four indices. For example, for the SYNNP index, calculated IQ was defined as

$$Calculated\ IQ = a_{SYNNP} \times SI + b_{SYNNP} + diff_{SYNNP} \quad (6)$$

where SI = test sample SYNNP value. a_{SYNNP}, b_{SYNNP}, and $diff_{SYNNP}$ = coefficients calculated for SYNNP feature on the training set.

4.2 Result

Hendrix and Yampolskiy[34] tested the accuracy of their method on writing samples that were obtained from social media contacts with their corresponding IQ. A total of 4 samples were used for testing and the results obtained can be seen in Table 3.

Subsequently, Abramov and Yampolskiy[35] used a set of GRE text samples and the corresponding GRE analytical writing score which ranges 1−6. For this to be viable for testing, they mapped the score to an IQ range and tested for the range. Table 4 shows how this was done.

Since expected IQ is expressed as a range, the error calculation was done by calculating the error between the calculated IQ and the higher and lower boundary of the expected IQ range. If the calculated IQ falls between

Table 3 Expected IQ vs. measured IQ[34].

Sample word length	Sample collegiate word count	Sample CWR	Expected IQ	Measured IQ	Error (%)
752	94	0.1250	153	123.88	19.03
412	51	0.1238	130	123.31	5.15
136	22	0.1618	141	141.36	0.26
3279	433	0.1321	129	127.24	1.36

Table 4 Mapping of GRE writing sample scores to IQ score ranges[35].

GRE score	IQ range	GRE score	IQ range
1	70–79	4	111–120
2	80–89	5	121–130
3	90–110	6	131–160

the range, error = 0. Any value that yielded an error less than 10% from either boundary was accepted.

As shown in Table 5, the results obtained showed a high correlation between expected and estimated IQ in cases where the IQ fell within the average range with WRDMEAc feature providing the best estimation of a person's IQ 75% of the time. Their research also showed that in the cases of extremely high or low IQ, the proposed method failed. But the research proved a correlation between IQ scoring and written text and the possibility of estimating IQ score from written text.

4.3 Discussion

The studies done on estimating IQ from written text so far are mostly preliminary studies into this research area and one main problem they both faced was lack of a substantially large dataset to test the validity of their hypothesis. Although, Hendrix and Yampolskiy[34] had concluded that the premise of the research was to introduce the concept of estimating an individual's IQ using their vocabulary, it was also suggested that a larger dataset would have presented a more accurate distribution and test the accuracy of the proposed method.

Table 5 Calculated IQ scores from Abramov and Yampolskiy[35].

Exp. IQ	Sample name	SYNNP	LDMTLD	WRDMEAc	LAR
70–79	Sample 1	72.20	75.55	67.69	84.98
	Sample 2	104.98	76.11	123.90	98.12
80–89	Sample 3	131.28	76.96	72.54	95.93
	Sample 4	113.32	72.42	81.55	91.60
90–110	Sample 5	121.32	86.88	83.68	108.70
	Sample 6	114.83	84.59	93.51	125.56
111–120	Sample 7	108.01	88.98	104.92	121.93
	Sample 8	109.74	92.79	124.72	101.74
121–130	Sample 9	103.36	76.14	111.63	94.82
	Sample 10	119.37	117.02	104.47	135.67
131–160	Sample 11	118.08	95.65	121.67	89.20
	Sample 12	124.14	87.02	75.10	102.62

Similarly, Abramov and Yampolskiy[35] encountered this same problem and resolved it by using publicly available GRE sample essays and mapping the scores for the analytical writing to a range of IQ scores. What can be agreed on though was that both studies yielded promising results and it would be worth exploring the possibility of applying the hypothesis on a much larger dataset to accurately test the validity of the hypothesis. These studies provided the first in-depth attempt to estimate an individual's IQ from written text using stylometry.

5 Making Estimations from Written Text Using Stylometry

Looking at the review of the literature on both IQ estimation with machine learning and IQ estimation from written text, we would notice one common denominator: the unavailability of a large dataset either for building a model or for testing. Wang et al.[12] had to gather brain scans from multiple scanning sites and that identifies the scanning source before applying a learning algorithm while Hendrix and Yampolskiy[34] had to reach out to social media contacts before they could get a dataset to test their proposed method and this only proved to yield just 5 samples.

The equipment and time taken by Wang et al.[12] to take EEG and fNIRS readings of their subjects further showed that getting dataset required for IQ estimation could prove to be expensive and time consuming. Also, during the process of writing this review, we reached out to a lot of researchers (approximately about 50) mostly in the area of cognitive science and child psychology, to request for a dataset that comprises of a body of written text with a corresponding IQ score but the response all came back negative with one researcher, Suzanne Tyas, saying that if we could find a dataset like this, she would also like to have it for her research.

The unavailability of this kind of dataset contributes to the presence of few literature in this area of study.

Also, as stated earlier, the task of quantifying IQ using machine learning is a new area of research with the first work published in 2015. But the use of stylometry to make predictions from written texts, is a field that is thriving and the presence of a lot of literature proves that. Stylometry combines various research fields (e.g., statistics, linguistics, and computer science) and is applied in various areas ranging from academic research to forensic evidences collection.

Due to the development of computers and automation, stylometry analysis has become easier. We identify several works done in estimating education, age, gender, nationality, and language of origin from written text and we show that previous work done in this area yielded promising results.

Some studies have shown that sociolinguistic observation that different groups of people speaking or writing in a genre using different languages write using that language differently[36] while some other studies have shown that some stylistic text features such as error in writing could be used to determine the author of the written text[37]. In this survey, we have reviewed ten articles/papers that have used stylometry to predict either or all of gender[38], native language[39], age[36], and personality[36] of an anonymous author (author profiling).

5.1 Corpus

A corpus (plural corpora) is all the writings or works of a kind or on a subject especially the complete works of an author. In natural language processing (NLP) or in our case, stylometry, a corpus is a text/documents collection that serves as the dataset that is analyzed to make predictions or estimation. The emergence of social media (Twitter, Facebook, etc.), has provided access to corpus that can used in stylometry or NLP. For instance, the CLEF initiative (conference and labs of the evaluation forum) which is a self-organized body with the sole mission of promoting research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure has since 2010 extracted tweets from twitter to build a corpus that is used by researches for author profiling tasks[40–42]. There have also been other sources of corpora for NLP. Koppel et al. (2005)[37] used the International Corpus for Learner English that was created to study the English writing of non-native English speakers and Koppel et al. (2002)[43] used a genre-controlled corpus of 566 formally written text that was extracted from the British national corpus. Another thing that is worth noting is that corpora can also come in different formats either in XML[44] or in plain text[36]. Out of the 10 papers that were reviewed in this section, 7 of these papers were from PAN CLEF 2017.

PAN CLEF 2017 corpus. For PAN 2017, the task was for a group of researchers (22 teams) to tackle the author profiling task using a corpus compiled by extracting tweets from twitter. The goal was to classify the gender and language variety of a twitter user solely by their tweets (informal text). The PAN 2017 training corpus comprised of twitter profiles and tweets in four different languages (English, Spanish, Portuguese, and Arabic). The corpus was labelled with the gender and the language variety information of the tweet authors. The language varieties selected can be viewed below.

- Arabic: Egypt, Gulf, Levantine, and Maghreb,
- English: Australia, Canada, Great Britain, Ireland, New Zealand, and United States,
- Portuguese: Brazil and Portugal,
- Spanish: Argentina, Chile, Colombia, Mexico, Peru, Spain, and Venezuela.

The language variety was pulled from the capital or more populated cities where the variety was used. From the city center, tweets within the radius of 10 km were retrieved. The tweets were labelled for language variety based on the region it was pulled from or based off of the author's profile and for gender it was labelled automatically with the help of a dictionary of proper nouns and manually, by visiting each profiles and looking at photo, description, etc. The final corpus was balanced in the number of tweets per variety and gender and in the number of tweets per author (500 authors/gender and variety and 100 tweets/authors). Also, the dataset was presented in XML format with a document containing all the tweets of one author. The dataset was divided into training and test datasets with a 60/40 proportion, i.e., 300 authors for training and 200 authors for testing. A total of 22 research teams were tasked with tackling this problem using this dataset and they all have published their results in Rangel et al.[41]

5.2 Machine learning and stylometry

The process of identifying or estimating the true writer of a given text is known as authorship attribution and it has been studied for decades[45]. Stylometry, according to Ramyaa et al.[46], in the context of author attribution, assumes that an unconscious aspect exists to an author's

style of writing that cannot be manipulated but possesses distinctive and quantifiable features. These characteristic features an author possesses should be frequent, salient, and quantified easily, and should be relatively immune to conscious control. Furthermore, these features should be able to distinguish authors especially if they write in the same genre, on similar topics, or even in the same period. But among stylometry researchers, Ramyaa et al.[46] have identified that one of the biggest problems is that there is no consensus as to what characteristic features, methodology, or techniques that could be applied in standard research. This problem has been exhibited in most studies in stylometry where most of the experiments have been directed to different authors with different techniques and there has not necessarily been a comparison of results that demonstrates which features prove to be more representative or which techniques can be considered to be more effective. Stylometry techniques like any other machine learning method, can be broken into stages: (1) preprocessing, (2) feature extraction and selection, (3) classification or analysis, and (4) testing.

Koppel et al. (2005)[37] approached the author profiling problem by showing that some stylistic text features (e.g., error in writing) could be used to determine the native language of an anonymous text. They exploited the use of several stylistic features that can be crudely classified into the following:

- Function words,
- Letter *n*-grams,
- Errors and idiosyncrasies.

To flag the errors, multiple error types were considered, and the errors were tagged automatically in each of the documents with their error types. Four error types were considered.

- Orthography — A range of spelling errors, e.g., missing letters, letter inversions, etc.
- Syntax — Non-standard usage, e.g., repeated word, missing word, etc.
- Neologisms — Creation of neologism parts-of-speech, e.g., fantabulous.
- Parts-of-speech bigrams — Rare POS bigrams.

The International Corpus for Learner English — a corpus that was created to study the English writing of non-native English speakers. All the authors in the corpus were university students mostly in their 3rd or 4th year that were taking the English as a second language class roughly in their 20's with the same proficiency in English. The nationalities that were considered were Russia, Czech Republic, Bulgaria, Spanish, and French. Also, 258 authors were considered for each of the languages. Each of the documents in the corpus was represented by a number vector of length 1035, where each vector represented the frequency of a given feature in the document. The features include:

- 400 standard function words,
- 200 letter *n*-grams,
- 185 error types,
- 250 rare POS bigrams.

A multi-class linear support vector machine was used as the learning method with a 10-fold cross validation experiment. The results showed that when all feature types were used arranged in front of each other, they obtained an accuracy of 80.2%. It is also worth noting that most of the errors were among the three Slavic languages (Russian, Czech, and Bulgarian).

The success of this methodology was highly dependent on the interaction of hundreds of features and as Koppel et al. (2005)[37] showed, there were several patterns that were unique to certain languages that they were easily able to exploit. For example, it was seen that for many authors in the Spanish corpus, there was a difficulty with doubling consonants (either they doubled unnecessarily as in fullfill or they omitted one of a double as in effect). This was also seen with a relatively huge number of the authors in the Czech corpus. This methodology also poses some questions for future research: (1) Was method precise enough to handle a lot of different candidate native languages? (2) How short can the body of text be and still permit accurate categorization?

Argamon et al.[36] exploited the sociolinguistic observation that different groups of people speaking or writing in a genre and in a language use that language differently. The main aim of the paper was to profile an author of a written text using a written text by the author. The profile dimensions that were being explored were gender, age, native language, and personality (neuroticism). They identified content-based features and style-based features and applied machine learning on the content-based features and style-based features independent of each other and combined them. A novel feature set was introduced that naturally subsumes both function and part-of-speech which has been known to be useful in linguistics. Systemic functional linguistics provided taxonomies describing meaningful distinctions among various function words and parts-of-speech. Three separate corpora were used to identify the profiles

(age and gender shared the same corpus, but they were labelled differently). The corpus for both gender and age was a full set of postings of 19 320 blog authors (each text was a full set of posts by a given author) that was written in English. The self-reported age and gender of each author was known and each age interval 13–17 (42.7%), 23–27 (41.9%), and 33–47 (15.5%) — the intermediate age groups were excluded to avoid ambiguity because many of the blogs were written across several years, in the corpus had an equal number of male and female authors. The mean length of the texts were 7250 words per author. The corpus for native language was extracted from the International Corpus of Learner English (A corpus created for the purpose of studying the English writing of non-native speakers from a variety of non-English speaking countries — Russia, Czech, Bulgaria, Spain, and France). 258 authors from each sub-corpus (languages) were selected and surpluses were randomly discarded. The resulting corpus had texts that were between 579 and 846 words long. Finally, the corpus used for personality was gathered from essays written by psychology undergraduates at the University of Texas at Hendrix which was part of their course requirements. The students were asked to write a short "stream of consciousness" essay in which they represented their thoughts and feelings over a 20-minute free-writing period. The length of the essays ranged from 251 to 1951 words. Also, each writer was required to fill out a questionnaire testing for the "Big Five" personality dimensions (neuroticism, extraversion, openness, conscientiousness, and agreeableness) but only the dimension of neuroticism (tendency to worry) was considered. "Positive" examples were defined as the participants with neuroticism scores in the upper third, and "negative" examples as those with scores in the lowest third. This was done to formulate it as a classification problem. The rest of it was discarded and this left the resulting corpus with 198 samples. They used Bayesian Multinomial Regression because it is a probalisticaly well-founded multivariate logistic regression which is resistant to over fitting. Generally, it has shown to be effective for text classification and problems relating to text classification. 10-fold cross validation was used to test the extent to which each profiling problem was solvable. The results showed that combining content based features and style based features yielded the best results for age (classes: teens, twenties+, and thirties+) with a classification accuracy of 76.1% and gender (classes: male and female) with a classification accuracy of 77.7%, content-based features yielded the best results for language (classes: Bulgarian, Czech, French, Russian, and Spanish) with a classification accuracy of 82.3%, and style-based features yielded the best results for neuroticism (classes: neurotic and non-neurotic) with a classification accuracy of 65.7%.

This study showed how the right combination of linguistic features and machine learning methods allowed them to estimate several profile aspects of an anonymous author. The study also poses two questions: (1) Can other profile components such as educational background or other personality components be extracted from texts using the techniques/methodologys given the right training corpus? (2) To what extent can a variation in genre and language affect the nature of the models that can be used to solve author profiling problems? Most of the articles reviewed (to a great extent, most of the studies that have been done) mainly focused on gender and/or native language estimation so the later question to a great degree has not been fully answered. The later question can be handled by using a genre-controlled corpus and/or a corpus with different languages.

Koppel et al. (2002)[43] presented in detail a methodology that uses a genre-controlled corpus to automatically classify formally written texts according to the gender of the author. They exploited methods used for typical text categorizations and authorship attribution to solve this problem. The genre-controlled corpus comprised of 566 documents extracted from the British National Corpus. No single author wrote more than three documents in the corpus and each document contained between 554 and 61 199 words with an average of 34 320 words each. During preprocessing, it is worth noting that there was a deviation from using a hand-selected set of features that were deemed most likely to help with distinguishing between categories (as is the norm) to begin with a very large set of lexical and quasi-syntactic features that were selected because they were more-or-less topic-dependent. Each of the documents was represented as a vector length of 1081 which is the total number of features that was used. This includes a 405 function words that was deemed to have appeared at least once in the document and also a list of n-grams part of speech (POS) using the British National Corpus' (BNC's) tag of 76 parts of speech, for example, PRP = preposition, NNI = singular noun, etc. 500 most common ordered triples, 100 most common ordered pairs, and

all single tag features were used. The use of the POS n-gram tags was identified as a relatively efficient way to capture heavier syntactic information. Using automated methods, the features were significantly reduced. These methods however made use of iterative runs of the learning algorithm to eliminate low-weighted features. Simply, the method could be described as a different approach to feature selection — the first step involved training a model using all the features and then based on the results, an automated method was used to assign weights to the features based on how well they helped the accuracy of the model. The method then selected features with the highest weights and eliminated the features with the lowest weights.

In detail, the method works by first finding a linear separator between documents authored by male authors and documents authored by a female author. This is achieved by assigning a weight vector w to each training document x, such that the vector product $w \cdot x$, exceeds a threshold T, if and only if x is written by a female author. The method that was used for finding the weights was a variant of the exponential gradient algorithm. The weights are iteratively updated using a learning formula based on a learning constant which was set to 3 throughout the entire experiment. This is done so that weights that reduce the dot product improperly are increased and vice versa. x was allowed to take on non-binary values (exponential gradient (EG)) but $s(w, x)$ was restricted to binary values (Balanced Winnow). The samples were then randomly reordered, and another cycle ran once all the samples had been used for training. This continued until all the training samples were correctly classified or 100 consecutives cycles had failed to produce an improvement in the number if samples correctly classified.

Table 6 shows the results obtained after ten separate runs of 56-fold cross validation using a feature set that includes POS only, function words (FW) only, and both function words and POS. This experiment (as seen in Table 6) showed that using a combination of FW and POS yielded the best results across genres even though using more parameters (features) than constraints (documents) could have easily led to over-fitting during training thereby affecting the testing accuracy.

Table 6 Results obtained after ten separate runs of 56-fold cross validation using a feature set that includes POS only, function words only and both function words and POS[43].

(%)

Domain	FW	POS	FWPOS
All	73.7 ± 0.86	70.5 ± 0.90	77.3 ± 0.79
Fiction	78.8 ± 1.1	77.1 ± 0.85	79.5 ± 1.1
Nonfiction	68.5 ± 1.3	67.2 ± 1.2	82.6 ± 0.99

Getting a greater accuracy overall though was affected by the difference between fiction and non-fiction. This difference was identified to harm the results. Using Winnow helped to overcome this difference because it exploited subtle dependencies between features. Less subtle learning methods (Naïve Bayes (NB) and Ripper) could not deal with this problem and they performed poorly when used for classification.

The next step was to identify how many features would contribute the most to a better classification. To achieve this, for each model that was trained in the cross validation trial, they selected 128 features that were considered to be most important (i.e., importance in a given model is the absolute value of its weight in the model multiplied by its average frequency in the training set) in each direction making it a total of 256 features and ran the cross validation again only of the selected features. This process was repeated for half of the most important features in each direction and it was iterated until 8 features in each direction. They were able to show enough differences in the writing styles between male and female authors in modern formal English articles. They exploited this difference by using a method (Winnow-like algorithm — a machine learning technique for learning a linear classifier using labelled samples) that automatically classified the documents with an accuracy of approximately 80%. This study presented convincing evidence that there was a difference between male and female writing styles and showed that some features selected were useful for classification and how the frequency distributions for these features in BNC differ for male and female. This study exemplifies the methodology used in most recent research in text categorization with the choice of features being the major difference (content-independent features). The approach used in this work should work well for other problems that involve style-based categorization.

Khan[44] followed a step by step like approach to solve the authorship profiling problem proposed: (1) Remove all XML tags, hashtags, links, and extra white spaces and extract the tweets in plain text from each file. All this text is combined into a single language variety v document D_v. (2) All extracted text was combined to find the top 100 words that appear frequently (term frequency) and

word pairs that appear frequently (word pair frequency). (3) Score each common term that occurred in the trend list T_{Li} and D_v. The trend score S_{vi} is defined in Eq. (9). (4) Steps 2 and 3 were repeated to find gender-based term frequency, word pair frequency, and trend scores calculation under each language variety. (5) Steps 1 – 4 were repeated for each Language L. These processes succeeded in creating different language variety classes C_v and gender subclasses C_{vg} where $g = gender$.

$$Term\ Frequency(TF_{v\_t}) = \frac{Total\ Term\ Occurences}{Text\ Length} \quad (7)$$

$$Word\ Pair\ Frequency(PF_{v\_t}) = \frac{Total\ Word\ Pairs}{Text\ Length - 1} \quad (8)$$

$$S_{vi} = \frac{Trend\ Score}{Text\ Length} \quad (9)$$

Finally, the system was designed to loop through each language folder and every document is processed in the same manner as a separate class C_{u_id} with unknown variety u and author identity id having its own TF_{u_t}, PF_{u_t}, and S_{ui}. The system then assigned the class C_{u_id} with a language variety while it looped through each variety class C_v. (1) For every single word and word pair that was common in each C_v and C_{u_id}, the score S_{v_id} is increased for class C_{u_id}. (2) Each trend score S_{vi} in C_v and S_{ui} in C_{u_id} and the absolute difference D_{v_id} between the both was calculated.

$$S_{v\_id} = \sum_{t=0}^{100}(TF_{u\_t} + TF_{v\_t}) + (PF_{u\_t} + PF_{v\_t}) \quad (10)$$

$$D_{v\_id} = \sum_{i=1}^{S} abs(S_{vi} - S_{ui}) \quad (11)$$

The smallest D_{v_id} is added to S_{v_id} and in this way a class C_{u_id} with unknown variety is assigned the language variety v having the highest score S_{v_id}. This system repeats the same process to predict gender, but the system decides among two classes instead. Figure 3 shows a diagrammatic representation of the system.

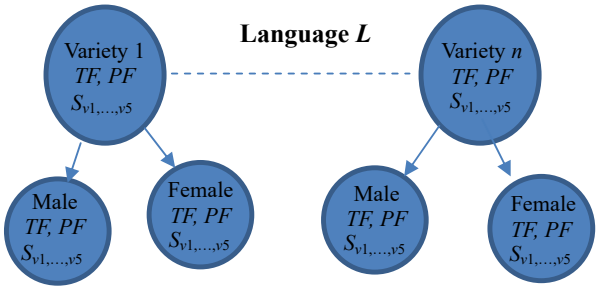

Fig. 3 Language variety classes and gender subclasses for a language L[44].

A 10-fold cross validation was used, and the same approach was used to predict both language variety and gender. The results obtained can be seen in Table 7.

The accuracy of the results obtained using this methodology shows an overall flaw in the system. One major flaw of the methodology that was identified was that as language varieties increased, there was a decrease in the prediction accuracy, and a suggested solution to this would be either increasing the number of single words and word pairs when creating the variety classes or increasing the trend.

Adame-Arcia et al.[47] implemented the use of two classification strategies, an instance-based classification and a prototype-based classification. The training dataset was preprocessed by expanding the short terms used and contractions and replacing characteristic traits (URL, Hastags, and mentions) with fixed patterns and they performed syntactic analysis using POS tagging tools for English and Spanish depending on the language of the tweet. Bag of words (BOW) was used for representation and integrate:

- The lexical terms, the lemmas of these, and the grammatical category — to differentiate the documents of each class because some of the features identify with their respective class.
- Characteristic features of the tweets — hashtags, mention of author, retweet mark, the use of URL, the use of intensification, and the use of laughter expressions, emoticons, and formal language. The position they were used was also considered.
- Features of subjectivity and opinion mining analysis — categorized as positive, high positive, negative, and high negative.

Adame-Arcia et al.[47] employed the use of two classification strategies. For the instance-based strategy, a document (set of author tweets) is used as an instance of a class to which it belongs and the similarity of the

Table 7 Results of using 10-fold cross validation to predict gender, language variety, and both by Khan[44].

Corpus	Predicion result		
	Gender	Variety	Both
Training – Arabic	0.5942	0.6079	0.3788
Training – English	0.6578	0.3017	0.2094
Training – Portuguese	0.6392	0.8975	0.5750
Training – Spanish	0.6307	0.3519	0.2193
Test – Arabic	0.5863	0.5844	0.3650
Test – English	0.6692	0.2779	0.1900
Test – Portuguese	0.6100	0.9063	0.5488
Test – Spanish	0.6354	0.3496	0.2189

new document to each sample document of the class is noted and the average similarity obtained with the class is computed. This analysis is repeated with each class of a demographic trait and the object belongs to the class which obtains the highest average of similarity. For the prototype-based strategy, the similarity between the new document and a prototype class document is calculated and this is repeated for each class. Just like the instance-based strategy, the object belongs to the class in which it obtains the highest similarity.

The classification is done independently for each.

- Author demographic trait,
- Gender classes (2 classes),
- Language variety (for English 6 classes and for Spanish 7 classes).

The result is a combination of the two classifications and accuracy was evaluated by using a 2-fold cross validation they also tested their methods on the dataset from PAN 2015 (estimating gender and age).

To compare the results, a BOW baseline was used. The results as seen in Table 8 showed that both methods performed well with gender classification but performed poorly with language variety classification. In their comparison with the BOW baseline, they concluded that the solution to the problem they encountered was that they need to analyze and reduce the features used.

Kheng et al.[48] analyzed the impact of different combinations of feature representation techniques and classification algorithms in relation to classification accuracy. Various feature extraction techniques (*n*-grams, term frequency-inverse document frequency (TD-IDF), and latent semantic analysis (LSA) were implemented using machine learning algorithms (SVMs, Naïve Bayes, and Random Forests). A variation to the methodology implemented by Kheng et al.[48] was the study of the dependency between gender and language variety. Data preprocessing was done by removing short tweets, removing the twitter handles, removing URLs, converting all hashtags to lowercase, and removing stop words. Using *n*-grams (unigrams, bigrams, and trigrams at the word level), TF-IDF, and LSA — which were also used for dimensionality reductions, they extracted features from the documents that would be used for classification. Stylometric features were also considered but not implemented. Three learning classification algorithms (SVMs, Naïve Bayes classifier, and Random Forest) were used so that their results can be compared and the best performing one could be selected. These algorithms were selected because they have been known to perform better when used for author profiling tasks. The Multinomial Naïve Bayes variant of NB was selected for the experiment because the official sklearn documentation suggested that it performed better with TF-IDF and proved to be the fastest when training and classifying the dataset that was provided. As stated above, this researcher also wanted to study the dependency between gender and language variety, i.e., could they predict gender and use the results of the classification to predict language variety (gender-then-variety strategy) and vice versa (variety-then-gender strategy)? For the sake of this work, they referred to that as a successive classification and classifying both labels independent of each other was referred to as loose classification. Only the "gender then variety" strategy was considered for successive classification in this stage of the study. To achieve this, the following steps were followed.

- They trained a classifier to predict gender on the whole language corpus.
- They split each language corpus into 2 sub-corpora, based on the ground truth: one for the female authors and another for the male authors.
- On each sub-corpus they trained a classifier to predict variety. This provided them with a male-variety classifier and a female-variety classifier.
- They classified each author contained within the test-dataset on gender first and sort predicted males and predicted female authors into 2 sub-test-dataset.
- They classified each author contained within the sub-test-datasets with the associated variety classifier, i.e., the female-variety classifier predicts the variety labels for the authors classified as female, same for the males.

Figures 4 and 5 show a graphical representation of how loose and successive classification workflow. They tested and optimized (tuning the parameters for the classifiers and feature extractor to achieve the best results) all

Table 8 Prediction results obtained using instance-based and prototype-based strategies by Adame-Arcia et al.[47]

Language	Prediction result					
	Instance-based			Prototype-based		
	Gender	Language variety	Joint	Gender	Language variety	Joint
Spanish	**0.60**	0.20	0.12	**0.63**	0.3	0.19
English	0.56	0.23	0.14	**0.65**	0.3	0.20

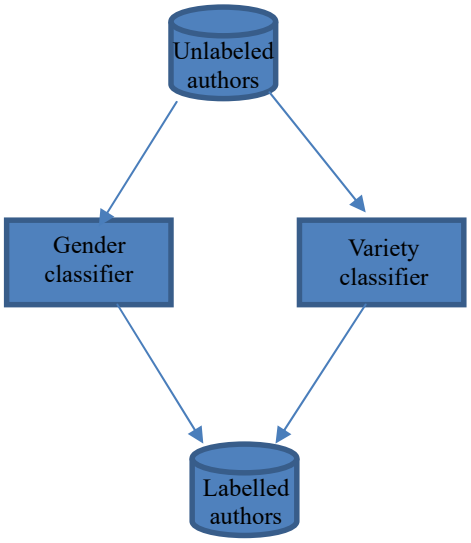

Fig. 4 Graphical illustration of loose classification[48].

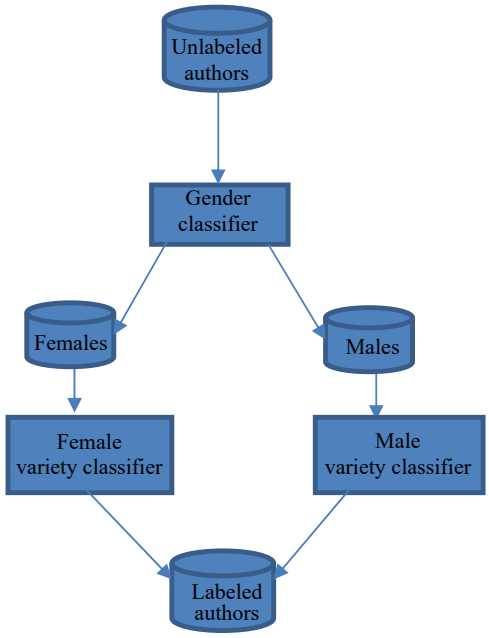

Fig. 5 Graphical illustration of successful classification[48].

the classifiers listed above using three different sets of features.

- Unigrams and bigrams at word level,
- TF-IDF based on unigrams, bigrams, and trigrams at word level,
- LSA with a combination of LSA and TF-IDF on unigrams and bigrams at word level.

Using 10-fold cross validation, they trained 24 models for each classification unit (480 models) while also testing some combinations of features with and without the stop words removal step of data preprocessing. Their evaluation of the best feature combination was done using micro-averaged and macro-averaged f-measures because the corpus had a balanced distribution over the different models. Tables 9 and 10 show the results obtained.

According to Tables 9 and 10, combining TF-IDF features on unigrams and bigrams and a Naïve Bayes Classifier (NBC) showed the best results for loose classification for gender classification and removing stop words worked better for the non-latin languages. Table 11 shows the results for successive classification. The results obtained did not look promising, so this classification method was dismissed. The eventual results obtained showed that predicting Portuguese and Spanish yielded the best results with 97.5% and 91.98%, respectively. This can be seen in Table 12.

Alrifai et al.[38] approached the author profiling by just focusing on the Arabic language corpus. The aim of the study was to consider the gender and variety (dialect) of an author as two important traits and could be used in analyzing tweets in Arabic. The corpus was cleaned by concatenating all the tweets (100 tweets) of every user into one long text. Using Farasa (a fast and accurate text processing toolkit for Arabic) they tokenized the long text into tokens.

Table 9 Gender classification results for loose classification[48].

Language	Preprocessing	Feature	Classifier	F macro	F micro
Arabic	Removal of stop words	TF-IDF (1/2-grams)	NBC	0.707	0.708
English	Removal of stop words	TF-IDF (1/2-grams)	NBC	0.669	0.669
Spanish	—	TF-IDF (1/2-grams)	NBC	0.659	0.661
Portuguese	—	TF-IDF (1/2-grams)	NBC	0.659	0.663

Table 10 Language variety classification results for loose classification[48].

Language	Preprocessing	Feature	Classifier	F macro	F micro
Arabic	—	TF-IDF (1/2-grams) & LSA	SVM	0.684	0.684
English	Removal of stop words	TF-IDF(1/2-grams)	SVM	0.669	0.669
Spanish	—	TF-IDF (1/2-grams) & LSA	SVM	0.684	0.684
Portuguese	—	TF-IDF (uni-, bi-, and tri-grams)	SVM	0.879	0.879

Table 11 Successive classification[48].

Gender	Language	Preprocessing	Feature	Classifier	F macro	F micro
Female	Arabic	—	TF-IDF (1/2-grams) & LSA	SVM	0.673	0.674
	English	Removal of stop words	TF-IDF (1/2-grams)	SVM	0.466	0.467
	Spanish	—	TF-IDF (1/2-grams) & LSA	SVM	0.518	0.520
	Portuguese	—	TF-IDF (1/2-grams) & LSA	SVM	0.880	0.880
Male	Arabic	—	TF-IDF (1/2-grams) & LSA	SVM	0.687	0.687
	English	Removal of stop words	TF-IDF (1/2-grams)	NBB	0.450	0.449
	Spanish	—	TF-IDF (1/2-grams)	SVM	0.555	0.556
	Portuguese	—	TF-IDF (1/2/3-grams)	SVM	0.859	0.859

Table 12 Results obtained from Kheng et al.[48]

Language	Score obtained		
	Gender	Variety	Joint
Arabic	0.6856	0.7544	0.5475
English	0.7546	0.7588	0.5704
Spanish	0.6968	0.9168	0.6400
Portuguese	0.6638	0.9750	0.6475

Following in the steps as some of the researchers reviewed in this work, various features that would contribute to building the best prediction model for variety and gender was considered.

- Character *n*-gram,
- Links, hashtags, and mentions usability ratios,
- Lengthened word ratios,
- Unigram, bigram, and trigram of tokens,
- Stems of tokens,
- Part of speech.

To select the best features that would yield the best models, a methodology that involved starting with a feature vector and calculating the testing corresponding accuracy was implemented. The next step was to add a new feature. If the new feature increased the accuracy, it was used otherwise, it was discarded. This was repeated for all the features. SVM classifier was used as the training algorithm.

Character *n*-gram and lengthen word ratios yielded the best accuracy and were selected for variety prediction while character *n*-gram and links, hashtags, and mentions usability ratios yielded the best accuracy for gender prediction.

Finally, the model was trained using SVM with linear, polynomial, and exponential kernels and with sequential minimal optimization (SMO) classifiers. The results obtained using SVM algorithm with the different kernels (on best feature vector of variety) showed that the polynomial kernel was the best with $F1 = 73.2\%$, compared to 67.1% for the linear and 62.7% for the exponential kernels. Also, they retrained a new model using SMO classifier instead of SVM, and the same best feature vector. The results showed an increase of 7% for variety and 3% for gender, and the testing accuracy for "both" traits together has also increased by more than 8%. This result showed that the SMO classifier led to optimum models for both traits, with testing accuracy equal to 75.5% for predicting variety only, 72.25% for predicting gender only, and 56.38% when they were predicted jointly.

5.3 Deep learning approaches

Deep learning approaches aim to create self-teaching and self-thinking machines. Recently, deep learning techniques have been presented as revolutionary methods which have surpassed most of the methods that had been referred to in the past as state-of-the-art[13]. This is mostly because of their ability to exploit simple and complex compositional features of data representations. We review some author profiling problems that were solved using deep learning methods.

Franco-Salvador et al.[49] used the Tweet NLP tokenization (specific for English tweets) but they modified it to identify Arabic, Portuguese, and Spanish punctuations. They also converted the tweets to lowercase and removed URLs. They employed the use of a recent variant of the continuous skip-gram model[50] which generates word embeddings using character *n*-gram embeddings to exploit the words' morphology. A character based embedded model helps to create robust classification models in the presence of abbreviations and typos (which is common to twitter and social media in general and helps to capture morphological nuances). The original model used the scalar product of the word vectors for scoring while the new subword model using a scoring function which represents the target words as the sum of its character *n*-gram vectors. Deep averaging networks (DANs) were used as the learning method with learning rate = 0.001 and epochs = 100. It was discovered during parameter selection that the best

performance for language variety identification was achieved using two layers but in contrast they noticed that the optimal number of hidden layers for gender identification differs depending on the language. A 10-fold cross validation was used for the training set and result obtained using DAN was compared with some baseline models (Bag of Words model classified with random forest, a model based on continuous skip-gram embedding averages classified with logistic regression, and a model based on the subword embedding classified with logistic regression).

The results as seen in Table 13 show that the embedding-based models outperform the BOW model which is the only purely lexical approach. DAN with subword embeddings shows the best result and this proves that deep averaging networks perform well with author profiling to magnify the most discriminant values contained in an embedding average and this demonstrates that it is a competitive alternative. The methodology implemented by Franco-Salvador et al.[49] yielded really good results especially when used to predict language variety. This method opens the door for investigation into how semantic representations and deep learning techniques can be employed in author profiling.

The use of deep learning methods to tackle the author profiling task set out by CLEF 2017 so has yielded good results so it would be nice to see if deep learning techniques, when applied by another researcher, gives the same or better result.

Kodiyan et al.[39] implemented a solution that is based on a bidirectional recurrent neural network (bi-RNN) using gated recurrent units (GRUs) in combination with an attention mechanism. In recent times, the success of RNNs has been achieved through LSTM and GRUs and they have achieved excellent results when used with various NLP tasks.

Kodiyan et al.[39] preprocessed the CLEF 2017 corpus by

- Converting every tweet to lowercase,
- Replacing the urls and usernames with tokens,
- Converting all hashtags to plain texts,
- Using vocabulary to map tokens with a tokenID which point to a vector representation for later use.

Each token in a tweet was represented by a pre-trained word embedding. A graphical representation of the bi-RNN with attention model that was used can be seen in Ref. [39].

Embedded layer: This layer was used to map the tokenIDs with their vector representation. This is used to look up the word-vector in the embeddings. The result is the matrix: $S \in O^{d \times n}$. O is the set of all vectors in each layer of the bi-RNN with GRUs, d is the dimension of the word vector, and n is the size of the input and was calculated by getting the tweet with the biggest number of tokens rounded up the nearest 10. This resulted in a maximum input size (n) of 60. Also, shorter inputs were padded with zeros to match the size and to reduce the effect of unknown and padded words masking was used. This way the model only used known words and skips zero-values.

GRU layer: It consists of two GRUs with u number of units and a GRU was used for each direction and this results in two vectors Formulas (12) and (13). The vectors were concatenated into a resulting matrix Formula (14).

$$R_F \in O^{u \times n} \tag{12}$$

$$R_B \in O^{u \times n} \tag{13}$$

$$R \in O^{2u \times n} \tag{14}$$

for the model $u = 50$.

Attention layer: This layer was used to weight the most important parts of the GRU encoded input to deliver

Table 13 Prediction results obtained using multiple models by Franco-Salvador et al.[49]

Task	Model	Prediction result				
		Arabic	English	Portuguese	Spanish	**Average**
Language variety	Random	25.0	16.7	50.0	14.3	26.5
	BOW	71.2	59.4	88.7	75.1	73.6
	Skip-gram emb.	73.0	62.4	98.6	80.6	78.7
	Subword emb.	70.7	68.3	98.5	79.6	79.3
	DAN	**80.6**	**76.5**	**98.9**	**91.0**	**86.8**
Gender	Random	50.0	50.0	50.0	50.0	50.0
	BOW	66.4	66.7	71.0	63.4	66.9
	Skip-gram emb.	71.2	78.4	76.5	73.3	74.8
	Subword emb.	73.7	78.8	72.6	74.5	74.9
	DAN	**74.5**	**80.8**	**78.8**	**75.5**	**77.4**

a simplified matrix input. The hidden state h_t:

$$h_t = \tanh(W_a R + b) \tag{15}$$

where $W_a \in O^{2u \times 2n}$ is the weight matrix, $b \in \{R^{2u}\}$ is the bias, and R is the output matrix from the previous layer.

The hidden state and the weight vector were then used to calculate the final attention a for each word:

$$a = \text{softmax}(h_t W_u) \tag{16}$$

The attention vector is then multiplied with the vector output from the previous step and the result is summed together. The result is a summarized representation of the sentence as vector $s_a \in O^{2u}$.

Softmax layer: This is the final layer that they used. It is a fully connected layer that used softmax as the activation function. The number of output nodes was dependent on the number of classification possibilities (gender prediction: two nodes and language variety: between 2 and 7 nodes depending on the language).

For optimization, the model was trained using AdaDelta optimizer $\epsilon = 10^{-5}$ and the default values for the other hyper-parameters.

The model was trained to classify single tweets, i.e., the tweets were classified separately to get the classification if an author. Let us consider gender classification for an author u with three tweets t_1, t_2, and t_3. The tweets were first classified individually $t_1 = [x_1, y_1]$, $t_2 = [x_2, y_2]$, and $t_3 = [x_3, y_3]$, where x_n is the probability of the tweet being written by a female and y_n is the probability the tweet was composed by a male. The outputs are summed, and the highest probability is selected as the gender classification of the author, i.e., gender classification = female, if $(x_1 + x_2 + x_3) > (y_1 + y_2 + y_3)$.

Also, the model was trained with 10-fold cross validation with 80% training data, 10% validation data, and 10% for testing for each fold. It is worthy to note that the test data did not influence the training and were only used to evaluate the model. This model was evaluated using an $F1$ score as shown below.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{17}$$

The $F1$ score was evaluated on the validation and a test dataset. Whenever the model achieved a higher $F1$ score on the validation data than the previous one, the model and its weights were saved. This also means that the model has a higher chance of performing well on the test set. They referred to this as the model checkpoints. During the process of working with the attention mechanism they developed a tool to represent how the different words in a tweet are weighted. The tool helped them to understand which words were more important for the model. For example, words like "color" and "Walmart" were marked as very important because they are common words in American English.

Also, words like "bloody" and "cheeky" were marked as significant for British English, which are also common usage in British English.

To evaluate their results, they used a 2-layered CNN architecture with a fully connected softmax layer at the end as a baseline. Tables 14 and 15 show the results that were obtained.

The results show that the highest score for gender prediction was achieved while Arabic achieved the lowest accuracy. For the variety prediction, the RNN model achieved great results with the lowest score in Arabic. Their results proved that their approach gave excellent results compared to the CNN model.

Schaetti[51] evaluated a strategy for the author profiling task by employing the use of TF-IDF and a deep learning model that was based on convolution neural networks. They also showed how the strategy can be applied to the four different languages in the task. Basically, they aimed at implementing two separate models and evaluate the results gotten from the models.

Before Schaetti[51] selected the features from the tweets, they went through a cleanup process using the following steps.

Table 14 Prediction results obtained for gender classifcation using bi-GRU + Attention and CNN by Kodiyan et al.[39]

Language	Prediction result (%)	
	bi-GRU+Attention	CNN
English	79.03	73.24
Spanish	72.57	72.93
Portuguese	79.50	79.83
Arabic	71.58	70.88
Average	75.67	74.22

Table 15 Prediction results obtained for language variety classifcation using bi-GRU + Attention and CNN by Kodiyan et al.[39]

Language	Prediction result (%)	
	bi-GRU+Attention	CNN
English	79.03	70.90
Spanish	92.05	89.67
Portuguese	98.76	98.75
Arabic	78.71	78.38
Average	87.11	84.22

• URL removal,
• Removing of the twitter usernames,
• Removal of special characters (#, $, and S),
• Clean numbers (100.00→ 100 S 00, 100,00 →100 $ 00, 100'000 →100 # 000) — This step was used to introduce three tokens (S, $, and #) which indicated the way a user represented numbers (points or comma for float, comma or apostrophe for thousands),
• Tokenizing punctuation (????→ " ? ? ? ? "),
• Removing the new line character,
• Replacing the useless characters with space (–, . . . , *, /, +, and \),
• Removing multiple spaces.

It is worth noting that they considered the fact that each the alphabets used would vary across the languages in the corpus so for each collection they kept in the text only letters and punctuation that corresponded to the language's alphabet. They also kept accented letters that are usually not used in a language to represent the author's usage. For example, accented letters in an English language text were kept representing the usage of French words by the English author because it could help as information to profile the author. This was based on a hypothesis that authors may use French or Spanish words depending on their country of origin. Hashtags were computed as normal words. Finally, words were separated by space and used as tokens.

Term Frequency — Inverse Documents Frequency (TF-IDF) is a weighting method that has often been used in information retrieval and text mining. It is a statistical measure that makes it possible to see the importance of term in a document in relationship to a corpus or collection. Term frequency is the number of times a term occurs in a specific document while the inverse document frequency is the measure of a term in the entire collection. The TF-IDF model is therefore a model that gives more weight to terms that appear less frequently in a document and are more discriminatory. It does this by calculating the base-10 logarithm of the inverse of the proportion of documents in the corpus where the term is contained. Therefore, term frequency for a term t and a document d is

$$tf_{d,t} = \frac{n_{d,t}}{|d|} \tag{18}$$

where $n_{d,t}$ = number of times t occurs in d.

The inverse document frequency of t in the whole collection was defined as

$$idf_t = \log \frac{|D|}{|\{d : t \in d\}|} \tag{19}$$

where $|D|$ = number of classes in the classification problem, $|\{d : t \in d\}|$ = number of documents where t appears.

Therefore, the final *tfidf* value for d and t was defined as

$$tfidf_{d,t} = tf_{d,t} \times idf_t = \frac{n_{d,t}}{|d|} \times \log \frac{|D|}{|\{d : t \in d\}|} \tag{20}$$

For each document d, a vector $tfidf_d$ is computed with each *tfidf* values for every term t in the collection. The value zero is assigned to any term that does not appear in the document. To predict the class of a previously unseen author in the collection, they considered this as a query q and computed the consinus similarity between the $tfidf_d$ vector and the vector tf_q of term frequencies in q. This is represented in the equation below.

$$sim(d, q) = \frac{tfidf_d \cdot tf_q}{||tfidf_d|| \times ||tf_q||} \tag{21}$$

where $tf_{q,t} = \frac{n_{q,t}}{|d|}$.

Finally, the predicted class $\hat{C}_q$ was selected by choosing the query q with the largest similarity. For example, $\hat{C}_q$ for gender was defined as

$$\hat{C}_q = \max_{d \in \{male, female\}.} sim(d, q) \tag{22}$$

As stated above, a deep learning model based on CNN was also used. They defined a CNN as a kind of feed-forward artificial neural network (ANN), in which the patterns of connection between the neurons are inspired from the visual cortex. In their approach, they applied a CNN to a matrix representation of a 2-gram of letters for an author in a collection. The structure of the matrix representation of the 2-gram of letters for an author in a collection can be seen in Ref. [51]. This served as an input for the CNN which can be seen in Ref. [51].

The first layer which is the convolution layer consisted of 10 kernels of size 5×5. This layer served as the input for the second layer which was 20 kernels of size 5×5. After this layer was the drop out layer which served as input for two linear layers that were based on rectified linear unit (ReLU). The final outputs were obtained using a softmax function to predict the class of the author. The predicted class was defined as the class with the highest output from the softmax function.

For the training phase of the experiment, 90% of the dataset was used for training and the remaining 10% was used to evaluate the performances at each iteration. During the experiment they observed that English attained lower loss at 64 iterations, Spanish at 66 iterations, Portuguese at 87, and Arabic at 38 iterations.

They also noticed that the CNN model quickly overfitted and this posed a major challenge which was for them to effectively fight overfitting. After the training phase was done, they selected the CNN obtained at the iteration with the lowest loss.

The results show that using TF-IDF achieved the better results when used to predict language variety with the Portuguese collection yielding the best results at 99% while CNN achieved its best results when it was used to predict gender (Portuguese 85%). It is worth noting though that CNN also performed excellently well when it was used to predict language variety for Portuguese with an accuracy of 98%. The results obtained can be seen in Table 16 below.

6 Discussion

From this literature survey, we can deduce that

(1) There is insufficient study in the field of IQ estimation using written text due to the unavailability of a useful/large dataset,

(2) The use of deep learning methods has shown promising results when used for author profiling problems.

Stylometry however, has been effectively applied to many areas of research (author profiling, author identification, forensics, etc.) while yielding great results. Preliminary research has been carried out in the area of IQ estimation using stylometry and they have produced wonderful results with Hendrix and Yampolskiy[34] and Abramov and Yampolskiy[35] yielding a classification accuracy of 75% each and

Table 16 10-fold cross validation on the four training collections[48].

Corpus	Prediction result obtained from learning algorithms			
	TF-IDF	CNN	Final	Random
English variety	**0.8333**	0.6563	—	0.1666
English gender	0.6805	**0.7803**	—	0.5000
Both	0.4724	0.5228	**0.6502**	0.0833
Spanish variety	**0.9323**	0.7804	—	0.1428
Spanish gender	0.6491	**0.7238**	—	0.5000
Both	0.6051	0.5648	**0.6747**	0.0714
Portuguese variety	**0.9925**	0.9833	—	0.5000
Portuguese gender	0.7317	**0.8500**	—	0.5000
Both	0.5313	0.8358	**0.8436**	0.2500
Arabic variety	**0.8609**	0.6750	—	0.2500
Arabic gender	0.6888	**0.7500**	—	0.5000
Both	0.5929	0.5028	**0.6456**	0.1250
Average	—	—	0.7035	—

Note: "Both" means a combination of both variety and gender.

Abramov and Yampolskiy[35] identifying meanifulness (measured by rating words based on a meaningfulness rating corpus. Words that are highly associated with other words get a high meaningfulness rating compared to words that are weakly associated with other words) as the best feature for estimating intelligence from written text. The major challenge that is common in this research area has been the lack of a large dataset. The ideal corpus that is required for IQ estimation from written text is a corpus that comprises of a written text and a corresponding IQ result. This is a difficult corpus to gather.

This seems to also be a problem with IQ estimation problems in general as the earlier researches into IQ estimation also showed how difficult it is to gather any corpus or dataset that can be used to estimate IQ because they are not readily available as seen in Wang et al.[12] which yielded root mean square error (RMSE) of 8.695 and 9.166 using a multi-kernel and single-kernel support vector regression, respectively. They can be also expensive and time consuming[22,23]. Firooz and Setarehdan[23] which showed a minimum relative error of 3.093% and 3.690% using a linear regression and support vector regression, respectively. The IQ estimation algorithm used was based on fNIRS and can be improved by increasing the number and variety of subjects, this is due to the fact that recording fNIRS signal is much simpler in the areas limited to the significant channels compared to the EEG-fNIRS or EEG recording and processing across the entire head. However, we reviewed other research into other author profiling problems (gender, age, native language, and personality type) and identified the success using stylometry and machine learning from written text. Koppel et al. (2005)[37] exploited common errors with native language speakers of a specific language and used multi-class linear support vector machine. This method achieved a classification accuracy of 80.2% which when compared with the method employed by Argamon et al.[36], achieved a good result. Argamon et al.[36] employed the use of Bayesian multinomial Regression to estimate gender, age, native language, and personality type from written text. The best result that was achieved with this methodology was with content-based features which achieved a classification accuracy of 82.3%. In this review, the methodology proposed by Argamon et al.[36] achieved the best result with estimating native language (when compared with the methods that did not use deep learning). However, it is

worth noting that the method proposed by Kheng et al.[48] achieved a classification accuracy of 91.98% and 97.5% in estimating authors of only Spanish and Portuguese, respectively, but performed badly with predicting authors from the other native language considered (like Arabic).

Koppel et al. (2002)[43] employed a method that automatically selected features that contributed the most to achieving an accurate result and used a winnow like algorithm for classifying the gender of formal text. This achieved a classification accuracy of 80% which when compared to other methods that did not use deep learning achieved the best results in predicting the gender of an author. The distance-based method used by Khan[44] showed to have a flaw which caused it to yield bad results while the instance-based method employed by Adame-Arcia et al.[47] performed well with gender identification but performed badly with language variety identification. Also, we see the use of support vectors in most of the author profiling problems[12, 23, 37]. We were also able to identify the success of deep learning methods with Franco-Salvador et al.[49] (deep averaging networks) achieving a classification accuracy of 86.8% (gender) and 77.4% (language). Kodiyan et al.[39] also showed the success of using recurring neural networks (RNNs) and achieved a classification accuracy of 75.67% (gender) and 87.11% (language). Schaetti[51] showed the success of using convolutional neural networks (CNN) to predict gender and also the success of using TF-IDF to predict language variety. The success of deep learning methods is mostly because of the ability of deep networks to exploit simple and complex compositional features of data representations[13]. Finally, we also see in this review the general success of author profiling using stylometry[37] and machine learning[39] using formal texts[43] and informal texts[38].

7 Conclusion

A lot of research has been done by psychologist in IQ testing and determining a person's intelligence quotient. The most common types of IQ test developed over the years include:

- Stanford-Binet Intelligence Scale,
- Universal Nonverbal Intelligence,
- Differential Ability Scales,
- Peabody Individual Achievement Test,
- Wechsler Individual Achievement Test,
- Wechsler Adult Intelligence Scale,
- Woodcock Johnson III Tests of Cognitive Disabilities.

Very little research has been done in estimating IQ from written text and this is because it is relatively a new area of research. But the researches done in this area have yielded promising results. Very few researchers are involved in IQ estimation using various datasets/corpora and different machine learning methods to estimate the IQ of an individual. This is because there is very little dataset available that could be applied in IQ estimation. So, there is very little material to review in this area. IQ estimation from written text is a new and emerging field with the first study done in 2017[34]. Although there has been success in estimating an author's traits or characteristics from written text, estimating intelligence from written text is a field that is still being researched. Based on this review, we can extrapolate from the success of stylometry and machine learning in the area of author profiling coupled with the relative success seen in the area of IQ estimation from written text that an individual's IQ can be estimated from a body of written text. For future research, we suggest that acquiring a larger text-IQ dataset would be very useful and machine learning model can be trained using stylometry features. Also, reinforcement learning can be applied on this dataset to continually improve the accuracy of the model.

Appendix

This section contains Tables A1 and A2. Table A1 shows a summary the papers reviewed with the dataset or corpora used in the respective study. It also shows the evaluation metrics used and the results obtained in each of the studies. Table A2 shows the publications reviewed with the conferences/journals, authors, and a link to access the publications.

References

[1] P. Hallinan, Book review: Psychological testing (5th edn), *Aust. Educ. Dev. Psychol.*, vol. 2, no. 2, p. 18, 1985,

[2] M. Lewis and P. Scale, RATIO IQ, https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470373699.speced1103,1983.

[3] K. Cherry, Alfred Binet and the Simon-Binet intelligence scale, https://www.verywellmind.com/alfred-binet-biography-2795503, 2020.

[4] E. Byington and W. Felps, Why do IQ scores predict job performance? An alternative, sociological explanation, *Res. Organ. Behav.*, vol. 30, pp.175–202, 2010.

[5] Historical importance of ASVAB testing, https://asvabmilitarytest.com/history-of-asvab-test, 2022.

[6] F. J. Tweedie, S. Singh, and D. I. Holmes, Neural network applications in stylometry: The federalist papers, *Comput. Hum.*, vol. 30, no. 1, pp. 1–10, 1996.

[7] N. Ali, M. Hindi, and R. V. Yampolskiy, Evaluation of authorship attribution software on a chat bot corpus, in

Table A1 Summary of papers reviewed, dataset/corpora used, evaluation metrics, and results.

(to be continued)

N/A	Publication	Machine learning method	Dataset	Source	Size	Evaluation metrics	Result
1	MRI-based IQ estimation with sparse learning[12]	SVR — Multi-kernel SVR, Single-kernel SVR	MRI samples of developing children between 6 and 15 years scanned at 5 different sites (NYU, KKI, SU, OHSU, and UCLA)	Autism Brain Imaging Data Exchange	164 samples (male/female: 130/34)	Correlation coefficient, root mean square error	• Multi Kernel SVR yielded a CC of 0.718 and an RMSE of 8.695. • Single kernel SVR yielded a CC of 0.684 and an RMSE of 9.166.
2	IQ estimation by means of EEG-fNIRS recordings during a logical–mathematical intelligence test[23]	Linear regression, support vector regression	fNIRS and EEG signals readings	fNIRS and EEG signals readings gotten from graduate students while they solved the RPM intelligence test	11 samples (male/female: 6/5)	Relative error between the real IQ (Cattle test) and estimated IQ	• A combination of fNIRS and EEG features selected using PCA yielded the best results.
3	Automated IQ estimation from writing samples[34]	Stylometry	Common crawl corpus and SAT vocabulary	https://aws.amazon.com/public-datasets/common-crawl/	Samples from common crawl with more than 100 words	Percentage error between Real IQ (from social media contacts) and estimated IQ	• The results show an accuracy of about 75%.
4	Automatic IQ estimation using stylometric methods[35]	Stylometry	Written text samples of American English published since 1990	Open American National Corpus (OANC) and SAT Vocabulary	6516 written samples and 5000 words from the SAT Vocabulary	Error between expected IQ range (gotten from sample GRE cores mapped to a range of IQs) and calculated IQ	• There was a high correlation between estimated IQ and calculated IQ with WRDMEAc feature providing the best estimation with a 75% accuracy.
5	Automatically profiling the author of an anonymous text[36]	Bayesian Multinomial Regression (BMR)	Three separate corpora. One to detect age and gender, one to detect native language, and the last one to detect personality type.	**Age and gender**: Full sets of postings from blog authors written in English **Native language**: International Corpus of Learner English **Personality**: Essays written by psychology undergraduates at the University of Texas, Austin, as part of their course requirements	**Age and gender**: 19 320 authors with a mean length of 7250 words/author **Native language**: 1290 authors. Between 279 and 846 words/author **Personality**: 198 authors. Between 251–1951 words/author	Classification accuracy	• Content-based and style-based features yielded the best results for age (76.1%) and gender (77.7%). • Content-based features only yielded the best results for language (82.3%). • Style-based features yielded the best results for neuroticism (65.7%).

Table A1 Summary of papers reviewed, dataset/corpora used, evaluation metrics, and results.

(continued)

N/A	Publication	Machine learning method	Dataset	Source	Size	Evaluation metrics	Result
6	Author profiling, instance-based similarity classification[47]	Instance-based similarity classification	XML-based tweets from twitter in four different languages (Arabic, English, Portuguese, and Spanish)	www.twitter.com	Arabic: 2400 documents English: 3600 documents Portuguese: 1200 documents Spanish: 4200 documents 100 tweets/documents	Classification accuracy	• Performed well in gender classification but poorly in language classification.
7	Arabic tweeps gender and dialect prediction[38]	Support vector machines (SVMs), sequential minimal optimization (SMO)	XML-based tweets from twitter in Arabic language	www.twitter.com	240 000 tweets written in Arabic by 2400 authors	Classification accuracy	SMO yielded the best results with • Language variety = 75.5%, • Gender = 72.25%.
8	Subword-based deep averaging networks for author profiling in social media[49]	Deep averaging networks (DANs)	XML-based tweets from twitter in four different languages (Arabic, English, Portuguese, and Spanish)	www.twitter.com	Arabic: 2400 documents English: 3600 documents Portuguese: 1200 documents Spanish: 4200 documents 100 tweets/documents	Classification accuracy between an instance-based and prototype-based classification	• DAN with subword embeddings yielded the best results. • DAN performs well in author profiling to magnify the most discriminant values contained in an embedding average. • It is a competitive alternative.
9	Author profile prediction using trend and word frequency based analysis in text[44]	Distance-based method	XML-based tweets from twitter in four different languages (Arabic, English, Portuguese, and Spanish)	www.twitter.com	Arabic: 2400 documents English: 3600 documents Portuguese: 1200 documents Spanish: 4200 documents 100 tweets/documents	Classification accuracy	•There is a flaw in the system which is a decrease in the prediction of variety when there is an increase in the number of language varieties. • The method yields bad results.
10	INSA LYON and UNI PASSAU's participation at PAN@CLEF'17: Author profiling task: Notebook for PAN at CLEF 2017[48]	SVMs, multinomial Naïve Bayes classifier (MNBC), random forest	XML-based tweets from twitter in four different languages (Arabic, English, Portuguese, and Spanish)	www.twitter.com	Arabic: 235 781 tweets English: 358 445 tweets Spanish: 418 090 tweets Portuguese: 118 105 tweets	Classification accuracy	• Combining TF-IDF features on unigram and bigrams using Naïve Bayes classifier yielded the best results. • Predicting Portuguese (97.5%) and Spanish (91.98%) yielded the best results.

(to be continued)

Table A1 Summary of papers reviewed, dataset/corpora used, evaluation metrics, and results.

(continued)

N/A	Publication	Machine learning method	Dataset	Source	Size	Evaluation metrics	Result
11	Author profiling with bidirectional RNNs using attention with GRUs[39]	Recurrent neural networks (RNNs)	XML-based tweets from twitter in four different languages (Arabic, English, Portuguese, and Spanish)	www.twitter.com	500 authors, 100 tweets per author	Classification accuracy between the RNN and a CNN based model as the baseline	RNN yielded better results than CNN with an average classification accuracy of • 75.67% for gender, • 87.11% for language variety.
12	TF-IDF and deep learning for author profiling[51]	TF-IDF based method, convolutional neural networks (CNNs)	XML-based tweets from twitter in four different languages (Arabic, English, Portuguese, and Spanish)	www.twitter.com	English: 360 000 tweets Spanish: 420 000 tweets Portuguese: 120 000 tweets Arabic: 240 000 tweets	Classification accuracy between TF-IDF and CNN	• TF-IDF performed better for predicting language variety. • CNN performed better when used to classify gender.
13	Automatically categorizing written texts by author gender[43]	Winnow-like Algorithm, Naïve Bayes, decision trees	Documents in British English that are labeled both for author gender and for genre: fiction and several non-fiction genres and sub-genres	http://www.ir.iit.edu/∼argamon/gender.html	Between 554 and 61 199 words with an average of about 34 320 words each (female = 34 795; male = 33 845).	Classification accuracy	• Function words combined with parts-of-speech yielded the best results across all genres with about 80% accuracy.
14	Determining an author's native language by mining a text for errors[37]	Multi-class linear SVM	Written text from non-native English-speaking students	International Corpus of Learner English	258 authors each from Russia, Czech Republic, Bulgaria, Spanish, and French sub-corpus	Confusion matrix	• Classification accuracy of 80.2% when all features are used in tandem with one another.
15	Intelligence quotient classification from human MRI brain images using convolutional neural networks[21]	CNN based IQ classification	ABIDE (autism brain image data exchange) provided by NITRC (neuroimaging informatics tools and resources)	Autism Brain Imaging Data Exchange	5000 bi-dimensional slices from each of the three brain views (15 000)	Classification accuracy	• ResNet-50 yielded a maximum accuracy of 85.9%. • Using the images from the sagittal view proved to yield the best results.
16	Predicting individualized intelligence quotient scores using brainnetome-atlas based functional connectivity[22]	Regression algorithms	MRI brain scans	MRI brain scans obtained using a Tesla magnetic resonance scanner	360 subjects between the ages of 17 and 24. 174 females and 186 males	Comparison of regression coefficients	• ReliefF + LASSO produced the best results with a regression coefficient of 0.5122 for all subjects and 0.7212 for all female subjects.

Table A2 Summary of papers reviewed with conference, authors, and link to access them.

N/A	Publication	Year published	Journal/ Conference name	Authors	Link
1	MRI-Based intelligent quotient (IQ) estimation with sparse learning[12]	2015	*Plos One* (Journal)	Liye Wang, Chong-Yaw Wee, Heung-Il Suk, Xiaoying Tang, and Dinggang Shen	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117295
2	IQ estimation by means of EEG-fNIRS recordings during a logical-mathematical intelligence test[23]	2019	Elsevier (Journal)	Shabnam Firooz and Seyed Kamaledin Setarehdan	https://www.sciencedirect.com/science/article/abs/pii/S0010482519301738
3	Automated IQ estimation from writing samples[34]	2017	MAICS 2017 (Conference)	Austin Hendrix and Roman Yampolskiy	http://ceur-ws.org/Vol-1964/CS1.pdf
4	Automatic IQ estimation using stylometric methods[35]	2019	ThinkIR (Electronic Thesis & Dissertation)	Polina Shafran Abramov and Roman Yampolskiy	https://ir.library.louisville.edu/cgi/viewcontent.cgi?article=4132&context=etd
5	Automatically profiling the author of an anonymous text[36]	2009	*Communications of the ACM* (Journal)	Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler	https://www.researchgate.net/publication/220427266_Automatically_Profiling_the_Author_of_an_Anonymous_Text
6	Author profiling, instance-based similarity classification[47]	2017	PAN at CLEF 2017 (Conference)	Yaritza Adame-Arcia, Daniel Castro-Castro, Reynier Ortega Bueno, and Rafael Mu-ñoz	https://pan.webis.de/downloads/publications/papers/adamearcia_2017.pdf
7	Arabic tweeps gender and dialect prediction[38]	2017	PAN at CLEF 2017 (Conference)	Khaled Alrifai, Ghaida Rebdawi, and Nada Ghneim	https://www.semanticscholar.org/paper/Arabic-Tweeps-Gender-and-Dialect-Prediction-Alrifai-Rebdawi/21b3341024ec0df3a73f7d30cf067686f0103464?p2df
8	Subword-based deep averaging networks for author profiling in social media[49]	2017	PAN at CLEF 2017 (Conference)	Marc Franco-Salvador, Nataliia Plotnikova, Neha Pawar, and Yassine Benajiba	https://www.semanticscholar.org/paper/Subword-based-Deep-Averaging-Networks-for-Author-in-Franco-Salvador-Plotnikova/a9d7350eb6c3381b43454ede11ad07789dccbf20
9	Author profile prediction using trend and word frequency based analysis in text[44]	2017	PAN at CLEF 2017 (Conference)	Jamal Ahmad Khan	https://www.semanticscholar.org/paper/Author-Profile-Prediction-Using-Trend-and-Word-in-Khan/4da5e57d2d07cf5336b2d6623bae090ea1c38e58
10	INSA LYON and UNI PASSAU's participation at PAN@CLEF'17: Author profiling task: Notebook for PAN at CLEF 2017[48]	2017	PAN at CLEF 2017 (Conference)	Guillaume Kheng, Léa Laporte, and Michael Granitzer	https://www.semanticscholar.org/paper/INSA-LYON-and-UNI-PASSAU's-Participation-at-Author-Kheng-Laporte/c61506eab622da887cbe6c8202435197735b3b67
11	Author profiling with bidirectional RNNs using attention with GRUs[39]	2017	PAN at CLEF 2017 (Conference)	Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak	https://www.semanticscholar.org/paper/Author-Profiling-with-Bidirectional-RNNs-using-with-Kodiyan-Hardegger/915654a4a29fd86621caeb7022fa484092f5e33b

(to be continued)

Table A2 Summary of papers reviewed with conference, authors, and link to access them.

(continued)

N/A	Publication	Year published	Journal/ Conference Name	Authors	Link
12	TF-IDF and deep-learning for author profiling[51]	2017	PAN at CLEF 2017 (Conference)	Nils Schaetti	https://www.researchgate.net/publication/320287536_UniNE_at_CLEF_2017_TF-IDF_and_Deep-Learning_for_Author_Profiling_Notebook_for_PAN_at_CLEF_2017
13	Automatically categorizing written texts by author gender[43]	2002	*Literary and Linguistic Computing* (Journal)	Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni	https://academic.oup.com/dsh/article-abstract/17/4/401/1019830
14	Determining an author's native language by mining a text for errors[37]	2005	KDD'05: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining	Moshe Koppel, Jonathan Schler, and Kfir Zigdon	https://dl.acm.org/doi/10.1145/1081870.1081947
15	Intelligence quotient classification from human MRI brain images using convolutional neural networks[21]	2020	12th International Conference on Computational Intelligence and Communication Networks	A Arya and Manju Manuel	https://ieeexplore.ieee.org/document/9242552
16	Predicting individualized intelligence quotient scores using brainnetome-atlas based functional connectivity[22]	2017	2017 IEEE International Workshop on Machine Learning for Signal Processing	Rongtao Jiang, Shile Qi, Yuhui Du, Weizheng Yan, Vince D. Calhoun, Tianzi Jiang, and Jing Sui	https://ieeexplore.ieee.org/document/8168150

Proc. 2011 23rd Int. Symp. Information, Commun. Autom. Technol., Sarajevo, Bosnia and Herzegovina, 2011, pp. 1–6.

[8] N. Ali, D. Schaeffer, and R. V. Yampolskiy, Linguistic profiling and behavioral drift in chat bots, *CEUR Workshop Proceedings*, vol. 841, pp. 27–30, 2012.

[9] R. V. Yampolskiy, N. Ali, D. D'Souza, and A. A. Mohamed, Behavioral biometrics, *Int. J. Nat. Comput. Res.*, vol. 4, no. 3, pp. 85–118, 2014.

[10] R. J. Sternberg, Intelligence, *Dialogues Clin. Neurosci.*, vol. 14, no. 1, pp. 19–27, 2012.

[11] V. Čavojová and E. B. Mikušková, Does intelligence predict academic achievement? Two case studies, *Procedia - Soc. Behav. Sci.*, vol. 174, pp. 3462–3469, 2015.

[12] L. Wang, C. -Y. Wee, H. -I. Suk, X. Tang, and D. Shen, MRI-based intelligence quotient (IQ) estimation with sparse learning, *PLoS One*, vol. 10, no. 3, p. e0117295, 2015.

[13] M. Badar, M. Haris, and A. Fatima, Application of deep learning for retinal image analysis: A review, *Comput. Sci. Rev.*, vol. 35, p. 100203, 2020.

[14] M. Brammer, The role of neuroimaging in diagnosis and personalized medicine-current position and likely future directions, *Dialogues Clin. Neurosci.*, vol. 11, no. 4, pp. 389–396, 2009.

[15] C. J. Price, S. Ramsden, T. M. H. Hope, K. J. Friston, and M. L. Seghier, Predicting IQ change from brain structure: A cross-validation study, *Dev. Cogn. Neurosci.*, vol. 5, pp. 172–184, 2013.

[16] T. Ohtani, P. G. Nestor, S. Bouix, Y. Saito, T. Hosokawa, and M. Kubicki, Medial frontal white and gray matter contributions to general intelligence, *PLoS One*, vol. 9, no. 12, p. e112691, 2014.

[17] F. J. Navas-Sánchez, Y. Alemán-Gómez, J. Sánchez-Gonzalez, J. A. Guzmán-De-Villoria, C. Franco, O. Robles, C. Arango, and M. Desco, White matter microstructure correlates of mathematical giftedness and intelligence quotient., *Hum. Brain Mapp.*, vol. 35, no. 6, pp. 2619–2631, 2014.

[18] K. L. Narr, R. P. Woods, P. M. Thompson, P. Szeszko, D. Robinson, T. Dimtcheva, M. Gurbani, A. W. Toga, and R. M. Bilder, Relationships between IQ and regional cortical gray matter thickness in healthy adults, *Cereb. Cortex*, vol. 17, no. 9, pp. 2163–2171, 2007.

[19] G. S. P. Pamplona, G. S. Santos Neto, S. R. E. Rosset, B. P. Rogers, and C. E. G. Salmon, Analyzing the association between functional connectivity of the brain and intellectual performance, *Front. Hum. Neurosci.*, vol. 9, p. 61, 2015.

[20] H. E. H. Pol, H. G. Schnack, D. Posthuma, R. C. W. Mandl, W. F. Baaré, C. V. Oel, N. E. V. Haren, D. L. Collins, A. C. Evans, K. Amunts, et al., Genetic contributions to human brain morphology and intelligence, *J. Neurosci.*, vol. 26, no. 40, pp. 10235–10242, 2006.

[21] A. Arya and M. Manuel, Intelligence quotient classification from human MRI brain images using convolutional neural network, in *Proc. 2020 12th Int. Conf. Comput. Intell. Commun. Networks (CICN)*, Bhimtal, India, 2020, pp. 75–80.

[22] R. Jiang, S. Qi, Y. Du, W. Yan, V. D. Calhoun, T. Jiang, and J. Sui, Predicting individualized intelligence quotient scores using brainnetome-atlas based functional connectivity, in *Proc. 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Tokyo, Japan, 2017, pp. 1–6.

[23] S. Firooz and S. K. Setarehdan, IQ estimation by means of EEG-fNIRS recordings during a logical-mathematical intelligence test, *Comput. Biol. Med.*, vol. 110, pp. 218–226, 2019.

[24] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556, 2015.

[25] Brainnetome Atlas, https://atlas.brainnetome.org/, 2021.

[26] V. J. Schmithorst and S. K. Holland, Sex differences in the development of neuroanatomical functional connectivity underlying intelligence found using Bayesian connectivity analysis, *NeuroImage*, vol. 35, no. 1, pp. 406–419, 2007.

[27] N. Jaušovec, Differences in cognitive processes between gifted, intelligent, creative, and average individuals while solving complex problems: An EEG study, *Intelligence*, vol. 28, no. 3, pp. 213–237, 2000.

[28] R. J. Haier, B. Siegel, C. Tang, L. Abel, and M. S. Buchsbaum, Intelligence and changes in regional cerebral glucose metabolic rate following learning, *Intelligence*, vol. 16, nos. 3&4, pp. 415–426, 1992.

[29] U. Basten, C. Stelzel, and C. J. Fiebach, Intelligence is differentially related to neural effort in the task-positive and the task-negative brain network, *Intelligence*, vol. 41, no. 5, pp. 517–528, 2013.

[30] U. Basten, K. Hilger, and C. J. Fiebach, Where smart brains are different: A quantitative meta-analysis of functional and structural brain imaging studies on intelligence, *Intelligence*, vol. 51, pp. 10–27, 2015.

[31] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, Automatic text categorization in terms of genre and author, *Comput. Linguist.*, vol. 26, no. 4, pp. 471–495, 2000.

[32] A. Abbasi and H. Chen, Applying authorship analysis to extremist-group web forum messages, *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 67–75, 2005.

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

[34] A. Hendrix and R. Yampolskiy, Automated IQ estimation from writing samples, https://aws.amazon.com/publicdatasets/common-crawl/, 2017.

[35] P. S. Abramov and R. V. Yampolskiy, Automatic IQ estimation using stylometric methods, in *Handbook of Research on Learning in the Age of Transhumanism*, S. Sisman-Ugur and G. Kurubacak, eds. Hershey, PA, USA: IGI Global, 2019, pp. 32–45.

[36] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, Automatically profiling the author of an anonymous text, *Commun. ACM*, vol. 52, no. 2, pp. 119–123, 2009.

[37] M. Koppel, J. Schler, and K. Zigdon, Determining an author's native language by mining a text for errors, in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Chicago, IL, USA, 2005, pp. 624–628.

[38] K. Alrifai, G. Rebdawi, and N. Ghneim, Arabic tweeps gender and dialect prediction: Notebook for PAN at CLEF 2017, *CEUR Workshop Proc.*, vol. 1866, pp. 1–9, 2017.

[39] D. Kodiyan, F. Hardegger, S. Neuhaus, and M. Cieliebak, Author profiling with bidirectional RNNs using attention with GRUs: Notebook for PAN at CLEF 2017, *CEUR Workshop Proc.*, vol. 1866, pp. 1–10, 2017.
[40] PAN at CLEF, https://pan.webis.de/, 2022.
[41] F. Rangel, P. Rosso, M. Potthast, and B. Stein, Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter, *CEUR Workshop Proc.*, vol. 1866, pp. 1–26, 2017.
[42] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein, Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations, *CEUR Workshop Proc.*, vol. 1609, pp. 750–784, 2016.
[43] M. Koppel, S. Argamon, and A. R. Shimoni, Automatically categorizing written texts by author gender, *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.
[44] J. A. Khan, Author profile prediction using trend and word frequency based analysis in text: Notebook for PAN at CLEF 2017, *CEUR Workshop Proc.*, vol. 1866, pp. 1–7, 2017.
[45] K. Nishiyama, G. O. Adebayo, and R. Yampolskiy, Authorship identification of translational algorithms, in *Proc. 2021 IEEE 15th International Conference on Semantic Computing* (*ICSC*), Laguna Hills, CA, USA, 2021, pp. 90–91
[46] Ramyaa, C. He, and K. Rasheed, Using machine learning techniques for stylometry, in *Proc. Int. Conf. Artif. Intell. IC-AI'04*, Las Vegas, NV, USA, 2004, pp. 897–903.
[47] Y. Adame-Arcia, D. Castro-Castro, R. O. Bueno, and R. Mu-ñoz, Author profiling, instance-based similarity classification: Notebook for PAN at CLEF 2017, *CEUR Workshop Proc.*, vol. 1866, pp. 1–7, 2017.
[48] G. Kheng, L. Laporte, and M. Granitzer, INSA LYON and UNI PASSAU's participation at PAN@CLEF'17: Author profiling task: Notebook for PAN at CLEF 2017, *CEUR Workshop Proc.*, vol. 1866, pp. 1–11, 2017.
[49] M. Franco-Salvador, N. Plotnikova, N. Pawar, and Y. Benajiba, Subword-based deep averaging networks for author profiling in social media: Notebook for PAN at CLEF 2017, *CEUR Workshop Proc.*, vol. 1866, pp. 1–10, 2017.
[50] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information, *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.
[51] N. Schaetti, UniNE at CLEF 2017: TF-IDF and deep-learning for author profiling: Notebook for PAN at CLEF 2017, *CEUR Workshop Proc.*, vol. 1866, pp. 1–11, 2017.

Roman V. Yampolskiy received the BS/MS degree in computer science from Rochester Institute of Technology in 2004 and the PhD degree in engineering and computer science from University of NY-Buffalo in 2008. He has been a tenured associate professor in the Department of Computer Science and Engineering at the Speed School of Engineering, University of Louisville since 2008. He is the founding and current director of the Cyber Security Lab and an author of many books including *Artificial Superintelligence: A Futuristic Approach*. During his tenure at UofL, he has been recognized as: Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and Outstanding Early Career in Education award winner among many other honors and distinctions. He is a senior member of IEEE and AGI, and member of Kentucky Academy of Science. His main areas of interest are AI safety and cybersecurity. He is an author of over 200 publications including multiple journal articles and books.

Glory O. Adebayo received the BS degree in computer science from Covenant University, Ota, Nigeria in 2012 and the MSc degree in computing: information engineering from Robert Gordon University, Aberdeen in 2015. He is currently pursuing the PhD degree in the Department of Computer Science and Engineering at the Speed School of Engineering, University of Louisville. He is also a research assistant in the Cyber Security Lab in University of Louisville. His main research areas are stylometry, NLP, and artificial intelligence. During his study at UofL, he has held leadership positions on the Student Governing Association and also been recognized for a student leadership award.