frontiers in ICT

ORIGINAL RESEARCH
published: 31 May 2016
doi: 10.3389/fict.2016.00008

Understanding Online Behavior: Exploring the Probability of Online Personality Trait Using Supervised Machine-Learning Approach

Ikuesan Richard Adeyemi, Shukor Abd Razak and Mazleena Salleh*

Information Assurance and Security Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Malaysia

The notion of online anonymity is based on the assumption that on the Internet the means of identification are limited to network and system identifiers, which may not directly relate to the identity of the user. Personality traits as a form of identity have recently been explored. A myriad of relationships between the Internet and human personality traits have been examined based on correlation and regression of media usage specific to selected media platforms, such as social networking sites. In these studies, the link between humans and the Internet based on interests and disposition was studied. However, the paradigm of the existence of a platform-independent digital fingerprint of personality trait is yet to be explored. This paradigm considers the Internet an extension of human daily communication that is capable of exhibiting a digital behavioral signature. Therefore, in this study, using client–server interaction as the fundamental unit of online communication, the probability of a digital personality trait distinction was explored. A five-factor model of a personality trait measurement instrument and server-side network traffic data collected over 8 months from 43 respondents were analyzed using supervised machine-learning techniques. The results revealed a high probability that the signature of conscientiousness personality trait exists in online communication. This observation presents a novel platform for the exploration of online identity. Furthermore, it charts a new research focus on human digital signatures capable of characterizing online behavior.

Keywords: personality print, personality trait, digital behavioral signature, human–computer interaction, logistic model tree

OPEN ACCESS

Edited by:
Alessandro Vinciarelli,
University of Glasgow, UK

Reviewed by:
Marco Fyfe Pietro Gillies,
Goldsmiths – University of London, UK
Dinesh Babu Jayagopi,
International Institute of Information Technology Bangalore, India
Dilana Hazer,
University of Ulm, Germany

****Correspondence:***
Ikuesan Richard Adeyemi
raikuesan2@live.utm.my

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in ICT

Received: *01 June 2015*
Accepted: *21 April 2016*
Published: *31 May 2016*

Citation:
Adeyemi IR, Abd Razak S and Salleh M (2016) Understanding Online Behavior: Exploring the Probability of Online Personality Trait Using Supervised Machine-Learning Approach. Front. ICT 3:8. doi: 10.3389/fict.2016.00008

INTRODUCTION

Human–computer interaction (HCI) is a convolutional discipline that integrates diverse research disciplines. Research on HCI involves the study of broader societal implications and interactions that are based on computer system usage (Hooper and Dix, 2013). The integration of human psychological studies in HCI studies for understanding human dynamics on the Internet has received minimal (Hooper and Dix, 2013) attention, with online behavioral identity being one major aspect. Research on human behavioral identity on the Internet integrates HCI and Web science such that traditional identification mechanisms (examples include network domain identifiers, security, and authentication tokens) are supplemented to provide a more reliable identification and profile-building process. Initial exploratory studies on the Internet revealed behavioral tendencies, such as loneliness compensation tendencies (Amichai-Hamburger et al., 2002, 2004; Ross et al., 2009), separation from family and depression (Amichai-Hamburger, 2002), blogging and mass media usage patterns (Guadagno et al., 2008; Schrammel et al., 2009; de Oliveira et al., 2011; Quercia and Kosinski, 2011; Moore and McElroy, 2012), and even addictive tendencies (Samarein et al., 2013). The existence of these tendencies can be attributed to the nature of the Internet, which provides a suitable platform for the integration of domestic, professional, and family life and social desires, as well as for the exhibition of inherent desires. Such a platform presents a paradoxical agent capable of revealing the personality of online users. In other words, the Internet presents an integrated platform for the identification and simplification of the complex human identity.

Identity is an important factor in Web science and computer usage, which is conceptualized into domain identification (Joiner et al., 2007), including physiological biometrics, social identity, technical identity, and behavioral biometrics. Human personality traits constitute the most common behavioral biometric adopted for the Internet user identification process (Amiel and Sargent, 2004; Guadagno et al., 2008; Correa et al., 2010). Personality traits are variables that coordinate human action and experience through dynamic psychological organization and they constitute a major discriminant for determining online behavioral patterns (Amichai-Hamburger, 2002). Trait theory is characterized by two fundamental tenets: quantification and cross-situational consistency. Personality traits as measured by the five-factor model (FFM) or the Big Three models (Matthews et al., 2003) have been observed to satisfactorily meet these tenets. Additionally, it has been observed that they adequately capture human interaction on the Internet.

Research questions that involve the frequency of online media usage, the demographic composition of online users, the relationship between individual differences and online media usage and motives for online interaction, and the probable relationship between Internet users and their probable preference have been addressed (Joiner et al., 2007; Guadagno et al., 2008; Correa et al., 2010; Davis and Yi, 2012; Moore and McElroy, 2012). These studies assessed the influence of various personality traits on Internet usage, based on the assumption that the Internet cannot replace human communication and entertainment. While such an assumption holds true in a larger aspect, one key component for understanding Internet and human interaction, that of an online behavioral pattern among individuals who share a similar personality, remains largely ignored. It has been asserted that the Internet is directly or indirectly related to individual personality traits (Amichai-Hamburger et al., 2002; Amiel and Sargent, 2004; Ross et al., 2009) and is mostly under the control of the individual and is a moderating platform for expressing anonymized identity (Young and Rodgers, 1998; Tan and Yang, 2012; Samarein et al., 2013), as well as a salient predictor of cyber space usage (Guadagno et al., 2008; Golbeck et al., 2011; Davis and Yi, 2012). However, the question regarding the existence of personality trait signatures on the Internet remains unanswered.

In de Oliveira et al. (2011), the probability that the personality trait of mobile phone users can be inferred based on trait mean score and call pattern was explored. Similarly, in Murray and Durrell (2000), the probability that the demographic attributes of online users can be inferred was investigated. In Ross et al. (2009), it was suggested that research studies targeting social networking sites, such as Facebook, Twitter, and LinkedIn, are angled toward the investigation of personality presentation on the Internet. The verification that a personality signature on the Internet exists involves observing the dichotomization of personality factors (Ross et al., 2009; Amichai-Hamburger and Vinitzky, 2010; Moore and McElroy, 2012) to determine distinctive characteristics among various individuals on the trait continuum. This study attempts to answer a primal underlying question: "Can the personality traits of an individual be inferred from his/her network traffic?" This question aligns with the logic that human recurrent daily behavioral patterns are a subject of the inherent personality trait, which regulates the synergy between online and offline behavior. However, a reliable answer to this question requires a platform- or application-independent network data source. Existing studies in the literature are limited to the platform of interest, such as e-mail, blogs, or Facebook, which induces behavior peculiar to its features and application. Various assertions about personality trait based on platform-dependent Internet sources are presented in **Table 1**.

PERSONALITY TRAIT AND THE INTERNET

The use of the personality trait FFM in Web science research, which consists of openness to new experience, conscientiousness, extraversion, agreeableness, and neuroticism, allows for a common vocabulary and metrics for investigating and understanding individual dynamics. The study presented in Golbeck et al. (2011) showed that humans reveal their personality trait in online communication through self-description and online statistical updates on social networking sites through which the FFM can provide a well-rounded measure of the human–computer relation. The study observed that the personality trait of users can be estimated (in social media) to a degree of ≅11% accuracy for each factor based on the mean square error of observed online statistics. This implies that personality trait prediction can be achieved within 1/10 of its actual value. In Guadagno et al. (2008), a similar inference based on blogging behavior was observed. The study explored the correlation between blogging and openness to new experiences as well as neuroticism. Similarly, a study reported in Lim et al. (2006) revealed that the temporal variability in e-mail delay and response can be adopted to infer the personality trait of the individual involved. In Salleh et al. (2010a,b, 2014), it was observed that the personality trait of an individual can be inferred from his/her pair-programing tendency.

The growing tendency to use individual personality traits in online studies indicates that personality traits constitute an online biometric modality for understanding and identifying online users (Delgado-Gómez et al., 2010). This paradigm is widely applied toward the comprehension of individual personality and online social media. Social media in this context refers to online platforms where an individual's consumption of digital media is channeled for interaction and/or expansion of social influence through online media, intention notwithstanding. Research on

TABLE 1 | Summary of assertions on personality trait.

Reference	Assertion from empirical/analytical observation
Amichai-Hamburger et al. (2002)	Neuroticism and extraversion: Internet usage is positively correlated with neuroticism and extraversion. Individuals high on neuroticism tend to reveal their true identity on the Internet as well as individuals low on the extraversion scale. Relationship in cyberspace can be transferred into real-life interaction
Salleh et al. (2010a)	Pair programing is not correlated with conscientiousness
Correa et al. (2013)	Emotional stability is negatively correlated with social media usage. Individual personality matters irrespective of demographics label. Individuals high on the extraversion and neuroticism scale tend to use social media more, while individuals high on openness use the social media more frequently
Salleh et al. (2010b)	Pair programing is not significantly correlated with the neuroticism personality trait
Moore and McElroy (2012)	Gender has significant effect on Facebook usage. Individuals high on extraversion have more Facebook friends, and report less regret, and the theory of social compensation holds true for individuals with high extraversion. Agreeableness and conscientiousness is positively related to social media use regret and individuals high on the openness scale report less frequency in social media usage. Neuroticism is negatively correlated to frequency of Facebook usages and regret. Individuals high on neuroticism scale spend more time on social media
Amichai-Hamburger and Vinitzky (2010)	Individuals high on the extraversion scale recorded higher number of friends. Individuals high on the neuroticism scale post more self-related pictures (assume such pictures imply a tendency to divulge personally identifying information)
Ross et al. (2009)	Preference for Facebook wall posting is positively correlated to neuroticism, while preference for photo posting on Facebook is negatively correlated neuroticism. Social media association is positively correlated to extraversion. Knowledge of computer-mediated communication is positively correlated to openness
Golbeck et al. (2011)	Individuals high on the extraversion scale tend to have more friends on Facebook. Such individuals tend to participate in more Facebook activity
Amiel and Sargent (2004)	Individuals high on extraversion prefer to use the Internet for research, information sharing, and opinionated tendency. Individuals high on neuroticism show particular interest in communal activities as opposed to one-to-one communication on the Internet (in respect to their desire to escape loneliness). Such individuals tend to show little interest in online discussion
Guadagno et al. (2008) and Yue et al. (2010)	Openness to new experience and neuroticism is a predictor of online blogging and such individuals mostly blog about themselves
de Oliveira et al. (2011)	Extraversion is significantly correlated to mobile phone usage. Extraversion and conscientiousness is significantly correlated to perceived usability of mobile phones
Swickert et al. (2002)	Neuroticism and conscientiousness personality is significantly correlated with computer usage for leisure
Landers and Lounsbury (2006)	Conscientiousness, extraversion, and agreeableness are inversely proportional to Internet usage. Conscientiousness is positively related to Internet usage for academic purpose
Schrammel et al. (2009)	Extraversion is directly proportional to number of friends on social media. Time spent on the Internet is not significantly related to conscientiousness. Agreeableness is not related to number of friends on Facebook. Openness scale is directly proportional to time spent on the Internet as well as the number of friends on social media
Tan and Yang (2014)	Extraversion is highly correlated with online application usage, followed by neuroticism

the relationship between an individual's personality and media consumption has explored correlation and regression, as shown in **Table 1**. In Correa et al. (2013), it was asserted that the synthesis of individual psychological make-up presents the capacity to reveal Internet usage. This assertion was further supported in Amichai-Hamburger (2005), where Internet usage was claimed to be dependent on the personality trait of the individual. The study suggested that the influence of personality traits can be observed in the duration of the individual's online browsing period and tendency to use the Internet. The duration of the online browsing period reflects the individual's choice, preference, and reflexes in cyberspace, which is largely controlled by his/her unique and stable psychological characteristics (Correa et al., 2013). Therefore, the browsing duration can reflect the tendency of loneliness, since highly neurotic individuals and introverts tend to spend more time on the Internet to compensate for a probable lack of physical interaction, while at the same time, it projects the interest of "real self" exploration in online interaction (Amichai-Hamburger, 2005; Schrammel et al., 2009). The tendency to use the Internet is defined in the context of the "rich get richer" and the "poor getting rich" theory (Amichai-Hamburger, 2002). For example, in Amichai-Hamburger and Vinitzky (2010), it was observed that individual high on the extraversion scale tend to use social media to enlarge their boundary of friends and influence, while individuals scoring high on the neuroticism scale tend to use anonymized online media for personal expression. In Hamburger and Ben-Artzi (2000), it was further suggested that the Internet can be described as a complex platform that presents a diverse paradoxical lexicon. The study, however, highlighted that Internet usage in itself does not explain the causation of individual usage (dis)similarity. As highlighted in **Table 1**, there appears to be a general consensus on the positive relationship between neuroticism and Internet usage, in particular on an anonymized channel.

Conversely, contrasting assertions seem to have been made about the relationship between personality trait factors and Internet service usage. For instance, in the study reported in Ross et al. (2009), which was grounded in a self-report measurement instrument, it was observed that conscientiousness is not a predictor of social networking. However, in the study reported in Amichai-Hamburger and Vinitzky (2010) and Moore and McElroy (2012), it was observed that conscientiousness is a predictor of online social networking. It is important to note that in Moore and McElroy (2012) and Amichai-Hamburger and Vinitzky (2010) the individual's profile was adopted as the measurement instrument, while in Ross et al. (2009), a self-report measurement instrument was used. Intuitively, the degree of observed correlation is dependent on the reliability of the measurement instrument. The features observed in the measurement instrument form the basis for the effective assertion that is predicated on the tempo-spatial properties of the data. **Table 2** presents a synopsis of the attributes considered in the studies in the literature on personality and the Internet. However, the features considered in these studies were platform- or application dependent. The observation reported in Schrammel et al. (2009) and Moore and McElroy (2012) substantiates the importance of the reliability of the data-centric measurement instrument. In order to study online patterns, tempo-spatial features that are independent of platform or application are required. In studies on online user identification (Herder, 2005; Padmanabhan and Yang, 2007; Kumar and Tomkins, 2010; Yang and Padmanabhan, 2010; Abramson, 2012; Herrmann et al., 2012; Abramson and Aha, 2013; Abramson and Gore, 2013), such platform-independent features, which include but are not limited to Web page visit characteristics, Web request characteristics, Web session characteristics, and Web genre characteristics, were adopted.

The integration of these applications and platform-independent features results in a robust mechanism for exploring individual behavior on the Internet. This study observed the probability of the existence of digital personality traits based on platform-independent features. It, thus, differs from existing studies as follows.

- The features considered are based solely on human action and are platform independent. Semantic structures in the observed features are adapted for pattern classification.
- The personality trait signature is considered based on the classification of trait dichotomy, in contrast to the correlation and regression of trait mean score. Dichotomy is defined in this context to mean a categorization of continuous variable as stipulated in Ören and Ghasem-Aghaee (2003).
- The measure of experimental repeatability and validation is based on a standard perspective of measurement in addition to the generic method of dichotomization of traits (Oren and Ghasem-Aghaee, 2003). This differs from the n-sigma thumb rule and equal thirds method applied in Ross et al. (2009), Amichai-Hamburger and Vinitzky (2010), and Moore and McElroy (2012) or the 40:30:30 dichotomy applied in Salleh et al. (2010a,b, 2014). The intuition behind the generic dichotomy is based on the limitation inherent in a data-centric dichotomy. A data-centric dichotomy yields varying borderlines for every dataset as revealed in the dichotomy observed in Ross et al. (2009) and Amichai-Hamburger and Vinitzky (2010).

TABLE 2 | Summary of features used in personality trait studies.

Reference	Features considered	Dichotomy of personality trait
Tan and Yang (2012, 2014)	Games and online friend, entertainment, basic application, finance, social networking, and online transaction	N/A
de Oliveira et al. (2011, 2013)	Duration of received calls, number of received call, number of placed call, number of SMS, MMS sent/received, and call detail records	Trait mean score
Guadagno et al. (2008)	Estimate of time spent using Internet for recreational purposes, time spent using instant message, number of e-mails written daily, hours spent maintaining blog, hours spent reading blogs, number of blogs read, frequency of updating blog, the use of real name, and content of blog	N/A
Landers and Lounsbury (2006)	Internet usage based on duration of hours spent, frequency of Internet service usage	N/A
Swickert et al. (2002)	Average time spent per week, search, visits to bulletin board, and visits to chat rooms	N/A
Golbeck et al. (2011)	Structural features: edge of friendship, personal information, activity and preference, language feature, and Internal Facebook statistics	N/A
Moore and McElroy (2012)	Time spent online, frequency of use, actual number of friends, number of photos, self-posting, and self-report on what the user does online	N/A
Ross et al. (2009)	Self-report computer-mediated communication competence instrument, self-report of basic use of Facebook, attitude associated with Facebook, and posting of personally identifying information	Equal third: only upper and lower cut-offs for each factor
Amichai-Hamburger and Vinitzky (2010)	Basic information, personal information, contact information, education, and work information on Facebook	Equal third: only upper and lower cut-offs for each factor
Correa et al. (2010)	Scale-self-report social media usage and socio-demographics	N/A
Salleh et al. (2010a,b, 2011)	Feedback (assignment, test, and examination) on pair-programing tutorial class	Low–Average–High: 40–30–30 30:40:30
Quercia and Kosinski (2011)	Logarithm of number of followed users, number of followers, number of listings, number of likes and retweets, the arithmetic sum of doubled number of followers, and Facebook friends, divided by two	N/A

According to these observed distinctions, current study focuses on answering the question: given a dichotomous continuum of personality trait, do individuals in a dichotomy exhibit a consistent signature distinguishable from individuals in another dichotomy? To answer this research question, two key assumptions were considered:

1. An FFM of a self-report personality trait instrument is sufficient to describe an individual on the personality trait continuum.
2. The self-report instrument of an individual is independent of other individuals.

These assumptions provide a singular composition of the individual for which a dichotomization and subsequent classification process can be performed as detailed in subsequent sections.

METHOD

In order to examine the probability that a personality trait distinction based on online interaction exists, server-side network data (from the fundamental building block of the Internet, client–server communication) were adopted. Additionally, the FFM personality trait measurement instrument was administered. Server-side network data was collected from the functioning servers in the Research Management Centre (RMC) in Universiti Teknologi, Malaysia for a period of 8 months. The network data inclusion in this study was conditioned on two criteria.

- The observed client (computer in this case) is used by only one individual throughout the duration of the data collection.
- Each client communicated frequently with the server for the duration of the data collection.

The RMC server is a research and development information system server that host research and daily operational activity for academic and non-academic staffs of the university. Network data collected from the server is in the form of log activity of each sampled user as discussed in the subsequent subsection.

Sample and Procedure

Server data were captured at the RMC server using a URL-request dump script that records the activity of each client in the organization. In order to enroll users for this study, proposal for the study was initially sent to the Director of the Research Centre where the Ethics committee recommended its approval. Furthermore, consent forms were distributed to RMC staff members. A total of 64 staff members volunteered for this study. Daily monitoring of the physical presence of these 64 staff members was conducted to satisfy the network data collection criteria. The 50-item International Personality Item Pool (IPIP) measure of personality trait was administered to the 64 respondents. However, only 43 respondents satisfied the network data collection criteria. Hence, this experimental study used 43 respondents, which accounts for 67% of the users who volunteered. An exploratory analysis of the 43 responses yielded a Cronbach's alpha reliability, presented in the comparative analysis in **Table 3**.

The Cronbach's alpha reliability of the conscientiousness personality trait (0.734) was observed to be closer to the reference IPIP-value (0.790). In addition, the conscientiousness personality trait have higher distribution of respondents as reflected by the value of its mean and SD (2.61 ± 1.02). Thus, for the analysis of network data, in this study, respondents were considered based on their conscientiousness trait (as highlighted in bold in **Table 3**). Conscientiousness personality trait is a continuous dimension of personality trait that describes individual tendency to demonstrate thorough and careful thought process, efficient and organized method of handling task, and systematic behavioral tendencies. The choice of conscientiousness personality trait does is primarily constrained to the reliability of the measured instrument as reflected in the Cronbach's alpha. Coincidentally, the choice of conscientiousness personality trait will also provide a better alternative to the contrasting correlational studies between Ross et al. (2009) and Amichai-Hamburger and Vinitzky (2010) and Moore and McElroy (2012). The network traffic of each respondent in the Conscientiousness personality trait was collected from 26th April 2014 to 31st December 2014.

Network Feature

A heuristic methodology was developed to clean the raw log file of the requested URL and to extract relevant human-centric features. The heuristics consider Web requests that originate as a result of human action, as opposed to requests initiated by a system or network facility on behalf of the individual. The heuristics were applied to individual requests and the following human-centric features were extracted based on a 30-min session boundary, which is the generally accepted session duration (Kumar and Tomkins, 2010; Yang and Padmanabhan, 2010). Network features considered in this study is based on the human-centric characteristic features defined in Adeyemi et al. (2014). The features that represent behavioral characteristics are intrinsic to human routine. Such behavior have been collectively applied in human studies (Adeyemi et al., 2014). These features are elucidated in the following subsections.

Web Request Characteristics

The individual Web request pattern was observed through the inter-request characteristics extracted from each session. Inter-request time (also referred to as interval) is the time difference between two consecutive requests within a session. The statistical properties of Web request characteristics as defined in Adeyemi et al. (2014), which include mean, SD, variance, kurtosis, and skewness of individual Web requests, were extracted from each session. These standardized features were considered with respect to interval and flight time. A total of 10 human-centric features were extracted from the Web request characteristics.

Visitation Pattern

The University Centre operates a two-server load-balancing client–server communication architecture. This implies that the possible number of probable Web pages is bounded by the total Web pages in the two servers as represented by

$$\mathrm{URL}_{\mathrm{Total}} = \sum_{i=1}^{s}\left[\int_{j=1}^{N}\mathrm{URL}_{j}\right] \tag{1}$$

TABLE 3 | Descriptive analysis of measured items.

Factor	Mean	SD	Variance	Skewness	Observed model	Cronbach's alpha reference model
Extraversion	3.26	0.96	0.95	−0.13	0.64	0.87
Neuroticism	2.96	0.9	0.82	−0.07	0.519	0.86
Agreeableness	3.13	0.9	0.85	−0.2	0.57	0.82
Conscientiousness	2.61	1.02	1.06	0.11	**0.734**	**0.79**
Openness	3.09	0.87	0.77	−0.15	0.721	0.84

s = total number of servers and N = number of unique URLs in each server.

In this study, it was assumed that individual Web request patterns obey a power law distribution as asserted in Barabasi (2005) and Zhou et al. (2008) from empirical experimentation. The visit characteristics considered in this study include aggregation of visit within session, rate of revisit per session, and session length with respect to visit aggregation, presented in Eqs 2–4, respectively.

$$V_{agg} = \sum\nolimits_{i=1}^{n}\left(\text{URL per session}\right)_i \Big/ \sum\nolimits_{j}^{N}\left(\text{URL under observation}\right)_j \tag{2}$$

$$\begin{aligned} R_{vs} &= \sum\nolimits_{i=1}^{n}\left(\text{URL per session}\right)_i \Big/ S_d \ , \ S_{d\Rightarrow}\text{session duration} \\ &= \int_{j=1}^{n} t_j \, dt, \leq 30 \text{ min} \end{aligned} \tag{3}$$

$$S_{agg} = \sum\nolimits_{j}^{N}\left(\text{URL under observation}\right)_j \Big/ S_d \tag{4}$$

The logic of rate of visit is in conformity with Eq. 1, on the premise that the probable URLs that an individual can visit are limited to the observable URLs in the server. In addition, this presupposes that the interest-driven model and priority-queue model of probable request patterns (Zhou et al., 2008) are captured by the bounded URL distribution such that all the observed users share similar working conditions and the major observable distinction can be revealed through observing the human behavioral composition. Three features were derived from the visitation pattern. In addition, session duration and total number of requests per session were also derived. A total of 15 features were extracted from the network traffic as summarized in **Table 4**.

The duration of the server-side data collection was divided into a pattern observation (training) phase and pattern validation phase. For the model training and validation phases, 21 and 15 weeks were adopted, respectively. In order to explore the distinction among the observed dichotomies, six supervised machine-learning algorithms were examined. The selection of the six classifiers was based on the initial exploration of applicable classifiers using the extracted features. Twenty two supervised classifiers were initially explored. This include Baseline classifier, BF tree, Decision stump, Hoeffding tree, J48, Logistic Model Tree (LMT), NB tree, Random forest, Random tree, Naïve Bayes, Bayes Net, Simple logistics, Hidden Markov Model, SMO, SVM, multilayer perceptron, Decision table, JRip, Partial Decision Tree (PART), k-NN, Decision Table Naïve Bayes (DTNB), and logistics regression model. Six classifiers performed significantly better than the baseline classifier. The six classifiers consisted of a logistic regression model, LMT, J48 decision tree, Reduced Error Pruning Decision Tree (REPTree), DTNB, and PART. Discussions of these classification algorithms can be found in Kotsiantis et al. (2007), Othman et al. (2007), and Nguyen and Armitage (2008).

TABLE 4 | Summary of features used in classification process.

Features used	Brief description
Visit session aggregation (V_{agg})	It is the ratio of the sum of the total URL visited in a session to the sum of URL-count (URL under observation) in the session. URL-count refers to the sum of the number of times any URL is revisited within the duration of a session. It also shows the sequential/parallel characteristics of the individual
Rate of visit per session (R_{vs})	It is the ratio of the sum of the total number of URL visited in Session to the duration of session. This feature shows the visit behavior of an individual within a session
Rate of visit-count per session (S_{agg})	It is the ratio of the sum of URL-count to the duration of the session. This feature shows the re-visitation behavior of an individual within the duration of a session
Total number of request per session	It is the total number of requests made within the duration of a session. This feature shows the behavior of an individual with respect to the amount of request capacity. It also indicates the nature of task being handled by the individual
Session duration	It is the absolute difference between the end-time of session and the start-time of session. This feature reflects the behavior of user within the delimited session of 30-min
Interval and flight mean	The mean of a distribution reveals the standard shape parameter of individual request pattern over the observed duration
Interval and flight SD	This feature reveals the degree of spread-out of individual request pattern within the period of observation
Interval and flight variance	This feature is similar to the SD. It measures the degree of proximity of individual request pattern over the period of observation
Interval and flight skewness	The skewness of interval and flight measures the degree of asymmetry of individual request pattern within the period of observation
Interval and flight kurtosis	These features show the behavior of individual request pattern based on its peak-width and tail-weight. They also measure the degree of outlier in request pattern

Flight = : $t_{i+1} - t_i$ where t_i is the ith time, corresponding to the converted time, while interval = : $(t_{i+1} - t_i) + 1$ s.

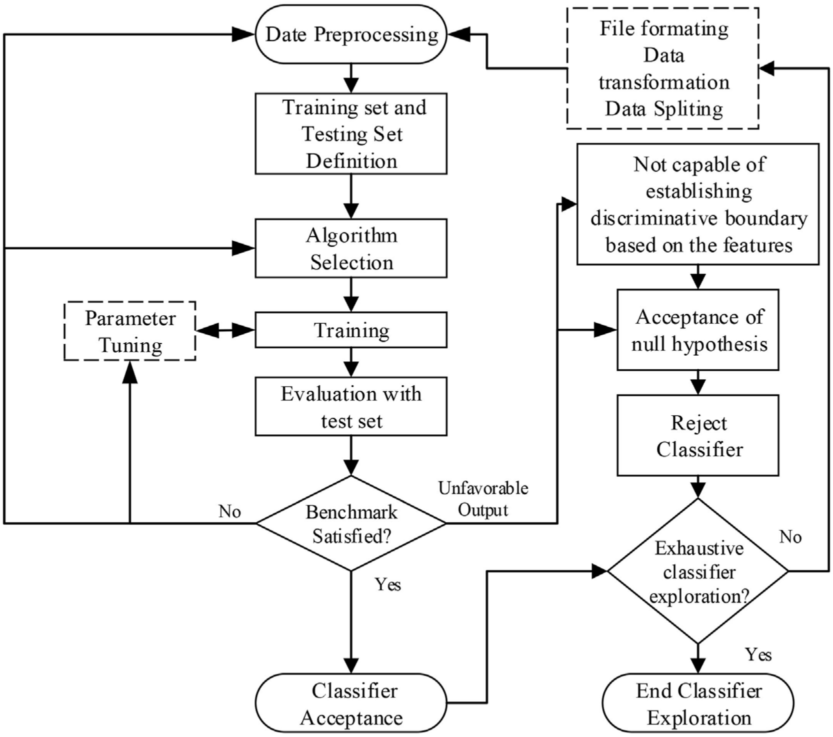
Date Preprocessing
Training set and Testing Set Definition
Algorithm Selection
Parameter Tuning
Training
Evaluation with test set
Benchmark Satisfied?
No
Yes
Unfavorable Output
Classifier Acceptance
File formating Data transformation Data Spliting
Not capable of establishing discriminative boundary based on the features
Acceptance of null hypothesis
Reject Classifier
Exhaustive classifier exploration?
No
Yes
End Classifier Exploration

FIGURE 1 | **Procedure for classifier exploration**.

The process adopted in this study to explore classification is similar to that defined in Kotsiantis et al. (2007), as presented in **Figure 1**. However, the exploration process applied in this study included an exhaustive search method for finding an applicable classifier. This involves a search for all the applicable classifiers capable of establishing a discriminative boundary among classes in the dataset based on the informative structure of the feature space. The process starts with the arrangement and sorting of the data to obtain uniformity. The result of this process is then input into the pre-processing section. Pre-processing involves data cleaning, extraction of the sequence of request, and sessionization of the request based on the adopted session threshold. The next stage involves splitting the dataset into training and testing samples. This is followed by the classifier exploration process. This process involves selecting a classification algorithm, splitting the dataset into training and testing, and then comparing the accuracy of the algorithm with the baseline accuracy. The default baseline for the exploration process is based on the highest class probability that can be measured by the ZeroR algorithm in the WEKA® tool kit. A classifier is considered applicable if the attained accuracy is significantly better than the baseline accuracy. Some classifiers, such as the logistic regression model, require parameter optimization through parameter tuning. For such a classifier, tuning is conducted to verify the applicability of the classifier to the feature space. A classifier that meets these criteria is then accepted. However, a classifier is considered inapplicable when is it not capable of establishing a discriminative boundary among the classes in the dataset as presented in the feature space, as depicted in **Figure 1**.

WEKA software was adopted for the classifier exploration process in this study. This is because it is a Java-based open source software that has gained widespread adoption for pattern classification and machine-learning processes because of its robustness to feature size, ease of integration (Othman et al., 2007) and within-script automation (Yang, 2010). The experimental process was based on the accuracy obtained using 10-fold cross validation and a 10-iteration process to prevent overfitting. The default setting in the WEKA tool kit was adopted for all parameters in the selected classifiers.

To evaluate the performance of each classifier, seven evaluation metrics were considered: accuracy, Kappa Statistics, root mean square error (RMSE), precision, recall, F-measure, and area under the receiver operating characteristics curve (AUC). The accuracy of each classifier is described by the degree of difference between the correctly classified [true-positive (TP) and true-negative] instance and the actual instance. The RMSE measures the magnified difference between the correctly classified instances and actual instances. RMSE (range from 0 → 1) is biased toward larger errors, a characteristic that makes it suitable for prediction performance evaluation. Precision (0 → 1) computes the ratio of correctness over the classified instances. It describes the consistency of the classifier. Recall (0 → 1) evaluates the performance based on the probability of the correctly classified instance. AUC (0 → 1) is the cumulative distribution function (CDF) of the TP to the CDF of the false-positive (FP). F-measure (0 → 1) measures the average rate of precision and recall of a classifier. It balances precision/recall trade-offs. Kappa (Cohen's kappa coefficient) statistics, however, measures the accuracy with respect to the *p*-value; thus, Kappa statistics measures the coincidental concordance between the output of a classifier and the label generation process. It compensates for random accuracy in a multi-class phenomenon. Its values range from −1 (total disagreement) through 0 (random agreement) to 1 (complete agreement), which implies that the computed accuracy depends on the efficiency and effectiveness of the classifier for the given observation.

Measurement Item Dichotomy

In order to define class membership based on the FFM domain score of each respondent, the dichotomy defined in **Figure 2** was adopted. High-, moderate-, and low-class dichotomy was obtained for 21, 12, and 10 respondents, respectively. A statistical *t*-test revealed a statistically significant difference between the mean of the observed dichotomies, which suggests a high inter-class boundary and low intra-class boundary. In order to prevent data redundancy and enhance computational efficiency, only individuals who exhibited unique request patterns were considered for inclusion in the pattern observation process. A unique request pattern was observed based on the discretization and symbolic translation of individual inter-request patterns. It was observed that the overall session instance and size of respondents for each dichotomy were reduced by 27.3, 8.33, and 5% for the low, moderate, and high classes, respectively, for the training dataset.

RESULTS

In order to explore the probability that a digital personality exists on the Internet, the conscientiousness trait was dichotomized to form clusters of the low, moderate, and high classes as depicted in **Figure 2** based on the assertion in Oren and Ghasem-Aghaee (2003) that identified a five-class dichotomy on the personality trait continuum. The adopted cut-off is independent of the distribution of the data. This is to prevent a data-centric dichotomy, while only three dichotomies were extracted from the trait score distribution. Six machine-learning classification schemes were applied to the extracted features to observe the structural relationship capable of revealing a dissimilar inter-dichotomy pattern. The schemes included J48 decision tree, a logistic regression model, LMT, DTNB, REPTree, and PART. Decision trees (DTs) are capable of presenting a high-level abstractive relationship between observable variables and lend themselves to ease of computation. Initial observations revealed that the observed schemes exhibit a higher classification accuracy than other types of classification scheme.

Signature Observation Based on Training Data

In order to examine the probability of distinguishing individuals based on the conscientiousness dichotomy and consequently to answer the research question, eight standardized machine-learning evaluation criteria were considered. The results of the experimental process are presented in **Table 5**. The experimental process was based on 10-runs of 10-fold cross validation. The selection of the observed classifier was based on a preliminary investigation to find a suitable classifier and on the preference that the results obtained should be logically interpretable. In addition, tree-based classifiers have been widely applied in online user identification (Yang and Padmanabhan, 2010). The results revealed that five classifiers, DTNB, PART, J48, LMT, and REPTree, achieved statistically significant classification accuracy relative to the baseline accuracy. The baseline classifier achieved on average an accuracy of 48.81%. However, DTNB, PART, J48, LMT, and REPTree achieved an average accuracy of 79.21, 76.66, 82.36, 84.96, and 80.74%, respectively. The LMT classifier attained a higher accuracy in distinguishing individuals along the conscientiousness personality trait continuum with a noteworthy low error rate (Type-I and Type-II), which indicates that the achieved results are reliable. The research question posed in Section "Personality Trait and the Internet" of this paper is answered using a statistical significance level of $p > 0.001$. The obtained accuracy for each classifier was measured relative to ZeroR, the baseline classifier adopted in this study. **Table 5** shows a baseline accuracy (ZeroR classifier) of 48.81%, which forms the null hypothesis (the probability of obtaining an accuracy level lower or equal to the baseline could be explained by random variation) for the study. The accuracy achieved by the logistic regression model is statistically insignificant. This implies that

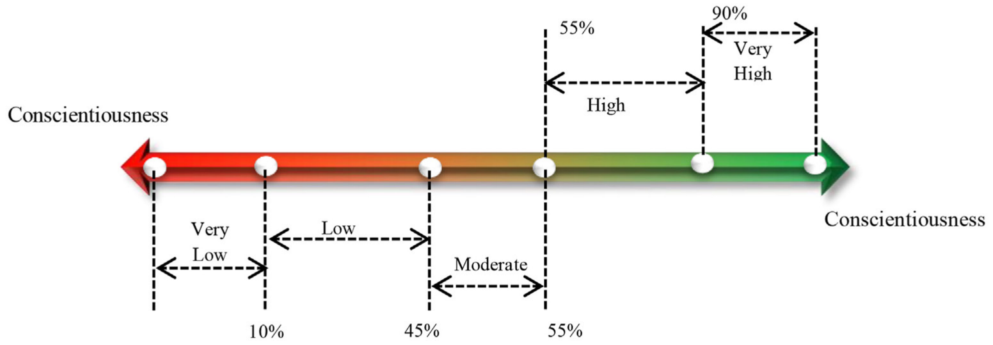

FIGURE 2 | Adapted cut-off for trait score dichotomy/class formation.

TABLE 5 | Result of signature pattern exploration.

Classifier	Accuracy	Kappa stats	RMSE	Precision	Recall	F-measure	AUC
ZeroR	48.81	0.00	0.45	0.49	1.00	0.66	0.50
Logistic	48.88	0.01	0.45	0.49	0.97	0.65	0.58
DTNB	79.21***	0.66	0.29	0.79	0.88	0.83	0.93
PART	76.66***	0.61	0.31	0.82	0.86	0.83	0.93
J48	82.36***	0.71	0.30	0.87	0.88	0.88	0.93
LMT	**84.96*****	**0.76**	**0.28**	**0.89**	**0.89**	**0.89**	**0.94**
REPTree	80.74***	0.69	0.30	0.84	0.87	0.85	0.94

**** indicates statistical significance at p > 0.001 with reference to the baseline model (ZeroR in this case).*

TABLE 6 | Result of signature pattern validation.

Classifier	Accuracy	Kappa stats	RMSE	Precision	Recall	F-measure	AUC
ZeroR	39.25	0.00	0.47	0.00	0.00	0.00	0.50
Logistic	43.8	0.10	0.46	0.44	0.26	0.33	0.58
DTNB	61.99***	0.40	0.40	0.67	0.60	0.63	0.80
PART	72.46***	0.58	0.35	0.78	0.70	0.74	0.88
J48	78.34***	0.67	0.33	0.79	0.79	0.79	0.90
LMT	**80.50*****	**0.70**	**0.33**	**0.83**	**0.82**	**0.82**	**0.92**
REPTree	75.54***	0.62	0.34	0.78	0.77	0.78	0.91

**** indicates statistical significance at p > 0.001 with reference to the baseline model (ZeroR in this case).*

the obtained accuracy using the logistic regression model can be explained by random variation. However, the other classifiers, DTNB, PART, J48, LMT, and REPTree, achieved a statistically significant accuracy level. The statistical significance of the results of these classifiers implies that the achieved accuracy is a function of the structural composition and efficiency of the classifiers.

The performance of LMT is significantly superior (as highlighted in bold in **Tables 5** and **6**) to that of the baseline classifier based on the value of the AUC (plot of the false-positive rate versus the true-positive rate). The AUC value ranges from 0 to 1, where a value <0.5 indicates that the result of the classifier is not better than a random guess. Values closer to 1 indicate that the classifier is robust and reliably accurate. AUC is robust to imbalanced data, thus providing a reliable metric that indicates how well a classifier separates the classes in the dataset. The AUC value (averaged at 0.94) shows that the LMT classifier can correctly separate instances (conscientiousness trait) of respondents into their appropriate dichotomies. Similarly, the F-measure, averaged at 0.89, indicates that LMT provides a reliable discriminatory boundary detector for the various classes in the dataset. The F-measure is the harmonic mean of precision and recall that indicates the precision recall property of a classifier. Furthermore, an assessment of LMT using RMSE and Kappa statistics indicated a fairly consistent performance. The RMSE value, averaged at 0.28, indicates that the LMT classifier can reliably estimate the posterior probabilities of each class. The RMSE value ranges from 0 to 1, where a value closer to 0 reflects the capability of the classifier to reliably estimate the posterior probability of each class in the experimental data.

Signature Consistency Based on Validation Dataset

In order to ascertain and verify the observed reliability of the results presented in **Table 5**, a separate validation dataset was evaluated. The results, shown in **Table 6**, were based on the same experimental condition as the training dataset, revealing consistencies in the probability of distinguishing individuals on the conscientiousness personality trait continuum. The performance of J48 and LMT is consistently higher than that of the other classifiers. These results indicate a very high probability of the existence of a digital fingerprint along the conscientiousness continuum.

ANALYSIS

The results in **Table 6** further substantiate the existence of a digital fingerprint as asserted by the results shown in **Table 5**. LMT performed consistently higher than the other classifiers. The average accuracy in the exploration phase and the validation phase shows a relative similarity at a value >80%, which suggests a statistically significant probability of the existence of a personality print. The results achieved by the five classifiers, as indicated in **Table 6**, show a very high statistically significant classification of the data space of the validation dataset relative to the baseline accuracy classifier (ZeroR). The results presented in **Table 6** are based on a class prior probability of 35.6, 39.3, and 25.1% for high-, moderate-, and low-class dichotomy, respectively. The observed average accuracy of LMT at 80.50% for the conscientiousness dichotomy, as demonstrated in **Figure 3**, indicates that individuals can be distinguished on the conscientiousness continuum to accuracy strength of 7.2, 7.0, and 6.9 out of every 10 instances for the high, moderate, and low classes, respectively. **Figure 3** further presents a comparative analysis of the accuracy for each class in the conscientiousness dichotomy. The parameters considered include the class prior probability of each class, the achieved accuracy of each class, the difference between the class prior probability, and the achieved accuracy. The results also indicate reliable internal consistency with the F-measure and AUC value at approximately

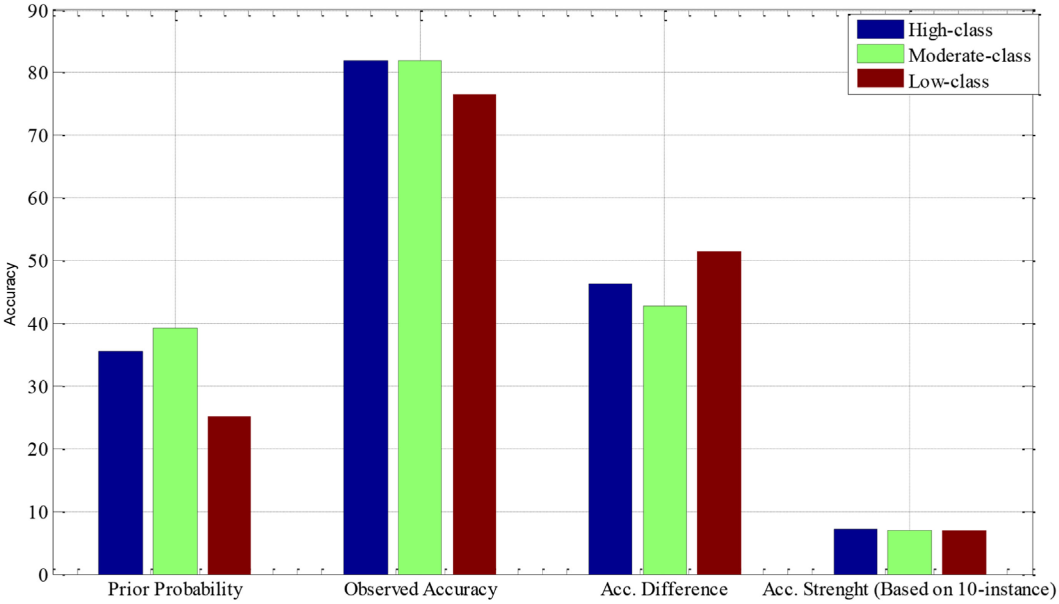

FIGURE 3 | **Analysis of validation result**.

TABLE 7 | Analysis of significance test.

Data	**≈Classifier accuracy (%)**					
	J48	**DTNB**	**PART**	**REPTree**	**Logistic**	**LMT**
Training dataset	82	79	77	81	49	85***
Testing dataset	78	62	72	76	44	81***

**** indicates statistical significance at p > 0.001.*

0.82 and 0.92, respectively. Furthermore, the sensitivity (recall) shows individual identification is reliable at approximately 0.82. This implies that for every given instance, the LMT model can correctly distinguish individuals on the conscientiousness continuum with 82% reliability. In order to verify the performance of LMT over the other classifiers (using paired *t*-test), the experiment was repeated in reference to LMT as the baseline classifier. The result of the test is presented in **Table 7**.

The statistical significance test was measured based on 10 repetitions of 10-fold cross validation result of each classifier (which generated 100 instances of accuracy for each classifier). Each classifier was paired with LMT and the sample mean of the pairs were tested on the assumption that there are no statistically significant difference between the achieved classification accuracy of LMT and each paired classifier. The result showed that the accuracy of LMT was statistically significant at 95 and 99% confidence interval, in reference to the other classifiers. This test, thus, lends further credence to the performance of LMT on the training and validation datasets.

Initial observation of the results of the validation process suggests higher accuracy for classification of the low class. However, a granular dissection of the accuracy strength reveals that the accuracy of the high class is slightly better than that of the other classes.

Discussion

The research question of this study stated in the form of a hypothesis asserts that individuals can be distinguished in online interaction based on their position on the personality trait continuum. The results presented in **Tables 5** and **6** support this assertion, showing a high probability of the existence of a personality trait signature. The dichotomy explored in this study is different from that adopted in Ross et al. (2009), Amichai-Hamburger and Vinitzky (2010), and Moore and McElroy (2012). It also differs from the dichotomy considered in Salleh et al. (2010a) and Salleh et al. (2014) as well as from n-sigma rules. The defined dichotomy adopted in this study is considered generic and unbiased toward data skewness. Such generic scaling provides a basis for experimental repeatability and a corresponding comparison with subsequent studies. Earlier studies in the literature, defined dichotomies based on trait mean score distribution (equal thirds, for instance). Such a cut-off boundary is biased toward the distribution of the data, which amounts to different boundary cut-offs for every dataset [the dichotomies observed in Ross et al. (2009) and Amichai-Hamburger and Vinitzky (2010), for example] and, hence, context dependent. Individuals classified as high class in one context can be classified as moderate in another. To frame research findings on such a basis contradicts the universality of the personality trait measurement instrument. Furthermore, an initial experimental evaluation based on the five-sigma rule and equal third dichotomy produced an accuracy below 10% for the prior probability of each class. The F-measure and sensitivity of each classifier based on the five-sigma rule were below 0.60. This suggests that a context-dependent dichotomy [such as presented in Ross et al. (2009)] cannot provide a reliable measure for the exploration of the digital conscientiousness trait. The results shown in **Tables 5** and **6** consistently show the capability of the explored dichotomy.

The results further demonstrate the superiority of the LMT classifier to other classifiers in terms of overall accuracy. An LMT classifier is a hybrid classifier that integrates a linear logistic regression model in a DT classification mechanism. Classification is achieved by generating decisions with logistic models at its leaves and the prediction estimate is obtained by using posterior class probability. The integration of DT into LMT enhances its superiority to linear regression models when applied to a highly multidimensional dataset that requires ease of human interpretability. The performance of the DTNB classifier was inferior to that of LMT in this study. DTNB constitutes the integration of a naïve Bayes algorithm in a decision table mechanism. An initial experiment based on naïve Bayes showed a very poor classification performance. The naïve Bayes classifier assumes that all the attributes in the dataset are independent. The capability of LMT to infer larger structural knowledge from a high dimension dataset can be attributed to its superiority to DTNB. PART is a rule-based induction algorithm that builds a DT by avoiding global optimization in order to reduce time and processing complexities. PART uses the separate-and-conquer approach of RIPPER and combines it with the DT mechanism of C4.5 by removing all instances from the training dataset that are covered by this rule and proceeds recursively until no instance in the dataset remains. PART builds a partial DT for the current set of instances by choosing leafs with the largest coverage as the new rule. However, LMT demonstrated a higher classification capability than PART in this study.

A REPTree applies regression tree logic and generates multiple trees in altered iterations by sorting the values of numeric attributes once. This is achieved through the information gain principle (which measures the expected reduction in entropy), tree pruning based on reduced-error pruning with a back fitting method, and integration of the C4.5 mechanism for missing values by splitting each corresponding instance into fractional instances. However, LMT demonstrated higher classification accuracy than REPTree in both the training and testing process. The J48 DT classifier is a Java coded version of the C4.5 DT implemented in the WEKA workbench. C4.5 is an induction-based learning algorithm that uses the information gain ratio, as opposed to the ordinary information gain, which is biased toward large value attributes, as the splitting criterion for recursively partitioning instances of attributes into attribute space. Instances are classified by constructing nodes that form the root of the tree using singular incoming edges to link nodes, while supporting multiple outgoing edges through a predefined discrete function of the input attribute value. The performance of LMT showed a higher classification accuracy than did that of J48 in terms of the measured parameters in **Table 7**. This is further shown in the average number of correctly and incorrectly classified instances in **Table 8**. A granular observation of the resultant model reveals that LMT generated a smaller size of rules for classification than did J48. However, the time taken to build the LMT model is significantly longer than that taken to build the J48 DT model. A comparative analysis of the performance of each classifier is presented in **Table 8**. Parameters considered include the time taken to build each model (time elapsed), average number of correctly classified instances in 10 iterations (average number of correct), average number of incorrectly classified instances in 10 iterations (average number of incorrect), and number of rules or trees generated by the algorithm to learn the instances of an attribute for prediction (number of trees/rules).

As shown in **Table 8**, the number of rules generated for classification is not a direct indicator of the effectiveness of the classifier. This is evident in the number of rules generated by LMT (903), J48 (1469), REPTree (779), and PART (181) in decreasing order of performance. In terms of time taken to build the model, REPtree outperformed all the other classifiers with an average number of 9019.3 correctly classified instances. Next in the performance rank based on the time taken to build the model is the J48 DT. The performance of this classifier is of particular interesting because in terms of its average number of correctly classified instances it is closer to that of LMT. Therefore, if time is considered a factor in the selection of a classifier, the J48 DT presents a better classification performance than LMT. Network forensic analysis processes (such as catch-it-while-you-can), which consider time an important factor, are an example of applications that favor the time taken to build the model. However, in a data analysis process, where accuracy is a critical factor irrespective of the time taken, the LMT classifier presents a better performance. Furthermore, the time elapsed (also referred to as the running time of a classifier) is dependent on factors, such as

i. the processor core; single versus multiple core system configuration,
ii. the generation of the processor core with respect to the read/write speed of memory,
iii. architecture: 32- or 64-bit, and
iv. input data to the classification algorithm.

The data input to a classifier determines its intrinsic classification complexity. WEKA measures the complexity gain of a classifier based on a log-loss function of base 2 of entropy gain. The complexity improvement column in **Table 8** shows the level of complexity gain on the class prior entropy for each observed classifier. The LMT and DTNB classifiers demonstrated high complexity gain as compared to other classifiers. This further suggests the effectiveness of LMT in terms of the accuracy, overall complexity, and reliability of the developed model.

An example of such an application is the stop-look-and-listen type of network forensic process. A depiction of the DT generated by the J48 classifier is presented in **Figure 4**. The DT from J48 is presented as against the DT from LMT model. This is because, LMT model generate rules that combines logistic model and decision tree that are relatively complex to interpret, as against rules generated from J48, and other rule-based classifiers. The figure illustrates a body of rules generated using the DT process of the J48 classifier. The results for the training dataset, as depicted in the partial DT presented in **Figure 4**, show that individuals in each dichotomy exhibit structural patterns that integrate the following network feature combinations on the conscientiousness scale.

TABLE 8 | Comparative analysis of performance of classifiers.

Classifier	Time elapsed	Complexity improvement	Average number of correct	Average number of incorrect	Number of trees/rules
ZeroR	0.01	0	5453.0	5718.0	Not applicable
Logistic	3.29	0.0321	5460.7	5710.3	Not applicable
DTNB	47.04	**0.8571**	8849.0	2322.0	Not applicable
PART	10.29	−19.0316	8564.0	2607.0	181
J48	2.16	−66.8363	9200.5	1970.5	1469 (735 leafs)
LMT	33.31	**0.5274**	9490.7	1680.3	903 (452 leafs)
REPTree	0.54	−18.0266	9019.3	2151.7	779

15-attribute, 11,171-instance, 10-fold cross validation, 10-iteration. 1469 (735) implies the classifier generated a total of 1469 trees with 735 leafs.

Low Class

In **Figure 4**, it can be seen that when the rate of visit was ≤0.106557, the rate of visit-count per session was ≤10.8684, the rate of visit was <0.009975 and >0.00, the session duration was >320 s, the rate of visit-count per session was >7.655462, the session duration was >1227 s, and the rate of visit per session was ≤8.488285, a total of 142 structural patterns were extracted and all the patterns were observed to belong to individuals who scored low on the conscientiousness trait continuum.

Moderate Class

In **Figure 4**, it can be seen that when the rate of visit was ≤0.106557, the rate of visit-count per session was ≤10.8684, the rate of visit was ≤0.009975, and the rate of visit was <0.00, a total of 47 structural patterns were extracted and all the patterns belonged to individuals in the moderate class of the conscientiousness trait continuum.

High Class

Top-down navigation of **Figure 4**, right navigation to node N185, reveals that when node N185 was >4.352697 and the session duration was >1615, 100 structural patterns were extracted. Ninety-six of these were observed to belong to individuals in the high class on the conscientiousness continuum. Only four patterns belong to a different class. A further navigation through the DT space reveals that 18 structural patterns were also extracted and 17 of the 18 structural patterns belong to individuals in the high class of the conscientiousness continuum. The structures reveal that individuals in this class share a similar session duration, which is relatively higher than individuals in other classes.

These rules demonstrate the manner in which DTs can be used to classify an individual on the Internet into the low, moderate, or high classes on the conscientiousness continuum. The rules further show the relationship between the personality trait of online users and the observed usage pattern of the Internet. In essence, it shows that online users who are on the Low-class, Moderate-class, or High-class on the Conscientiousness personality trait continuum can be distinguished on the Internet within a 30-min session span. Since this study could not extract classes for very low and very high dichotomies, it is logical to assume that the novel signatures of individuals who belong to such dichotomies on the conscientiousness trait continuum are not captured in the present study. However, this result is particularly insightful because it uses a DT. A DT offers several advantages over other types of classifiers. These include ease of interpretation by humans and ease of presentation and application. Furthermore, a DT does not require a prior assumption about the structure of the data, but builds knowledge based on the structure of the data.

Implications

Given the advances in modern day technology, the Internet continues to present challenging requirements for effective service delivery systems, as well as reliable information security and assurance. The idea of the existence of a personality print on the Internet presents a paradigm that is capable of explicating and subsequently understanding the complexities of Internet technology. The notion of a personality print stands up to the challenge and ultimately debunks the general dictum of "on the Internet, no one knows you are a dog." Personality print from a service delivery perspective such as e-commerce and e-learning, as well as recommender systems, presents a platform for the development of an intelligent Internet service system that is capable of detecting the characteristics of an individual and consequently predicting or suggesting effective personalized-service processes. An appropriate dissection of the personality identity of the end-user opens for researchers a more fundamental discourse on the underlying causation of network burstiness and the probability that sophisticated artificial networks capable of replicating or mimicking human dynamics can be built. Furthermore, it opens the path for researchers to better comprehend the evolution of Internet consumption. In terms of online security, the personality print introduces a complementary platform for online identification, specifically in one-to-many authentication processes. The integration of the personality print in the identification and authentication process further suggests a robust and reliable security mechanism laced with inherent human characteristics. The observed high probability of the existence of personality prints substantiates the assertions in studies in the literature related to personality and the Internet (Ross et al., 2009; Correa et al., 2010; Moore and McElroy, 2012), such that a generic description of the relationship between an online user and an observed behavioral pattern of network features can be explained.

The observation in this study is limited to several factors. First, only one personality factor was observed in this study. This is

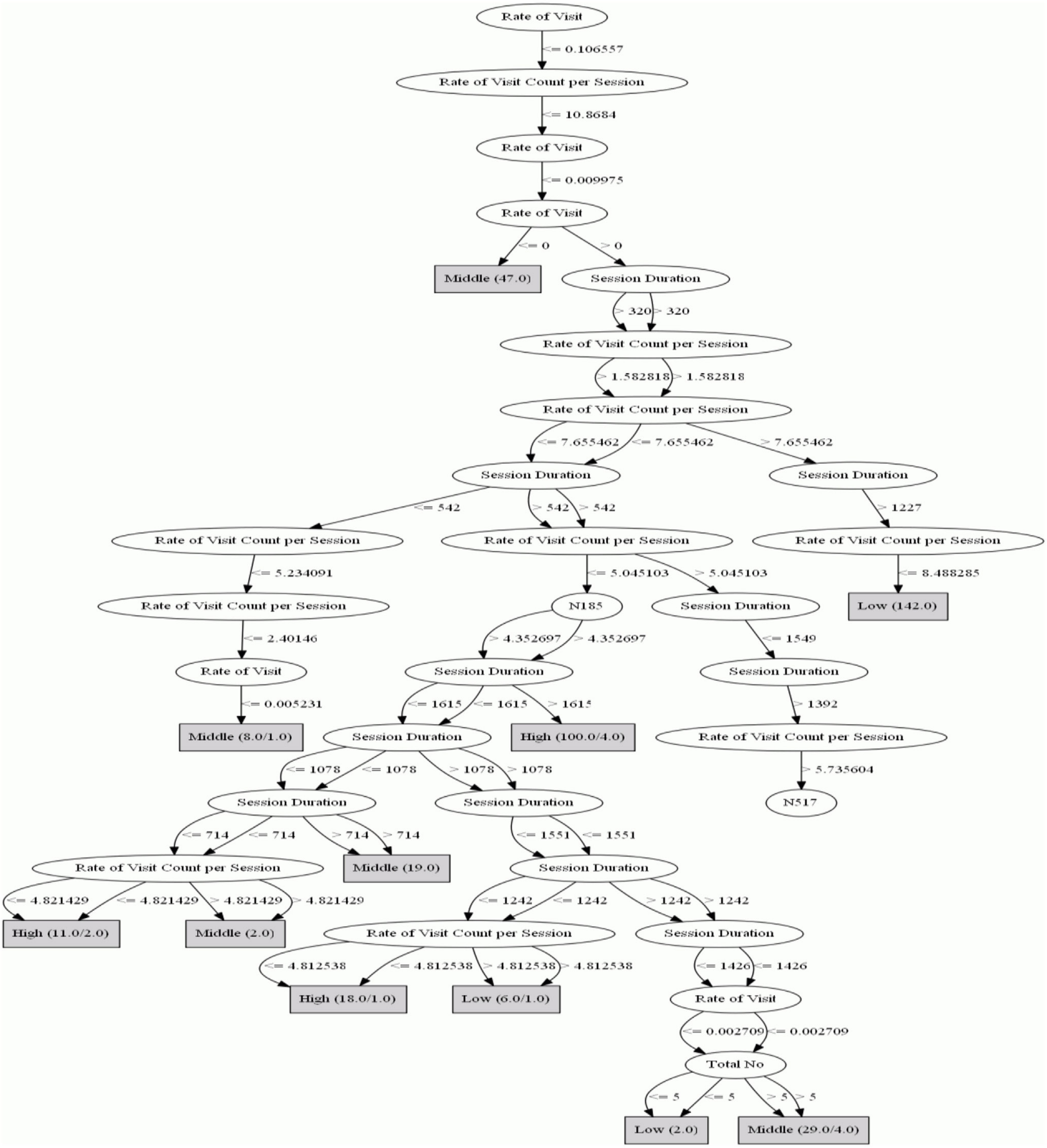

FIGURE 4 | **Partial decision tree of personality print**.

attributed to the limited sample size of respondents in the study. The observed classification accuracy of the LMT is below 100%. This implies that the results obtained do not represent a perfect classification but do provide a probabilistic analysis of the existence of a personality print. Furthermore, only three classes out of five dichotomous distinctions defined in **Figure 2** were explored in this study. This can also be attributed to the sample size adopted in this study. The integration of client-side network data into the pattern observation process presents a more comprehensive dataset for personality print exploration. An on-going process is being performed to expand this preliminary investigation, such that a larger sample size can be studied. In addition, other factors of the Big Five personality trait models are being explored.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

This research was supported/partially supported by Universiti Teknologi Malaysia and Ministry of Higher Education Malaysia under the vote number: R.J130000.7813.4F804.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fict.2016.00008

REFERENCES

Abramson, M. (2012). "Toward the attribution of Web behavior," in *IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA)*, (Ottawa: IEEE), 1–5.

Abramson, M., and Aha, D. (2013). "User authentication from Web browsing behavior," in *FLAIRS Conference*, (Palo Alto, CA:AAAI Digital Library), 268–273.

Abramson, M., and Gore, S. (2013). *Associative Patterns of Web Browsing Behavior* (Palo Alto, CA: AAAI Digital Library).

Adeyemi, I. R., Razak, A. S., and Salleh, M. (2014). "A psychographic framework for online user identification," in *International Symposium on Biometrics and Security Technologies (ISBAST)*, (Kuala Lumpur: IEEE), 198–203.

Amichai-Hamburger, Y. (2002). Internet and personality. *Comput. Human Behav.* 18, 1–10. doi:10.1016/S0747-5632(01)00034-6

Amichai-Hamburger, Y., Fine, A., and Goldstein, A. (2004). The impact of Internet interactivity and need for closure on consumer preference. *Comput. Human Behav.* 20, 103–117. doi:10.1016/S0747-5632(03)00041-4

Amichai-Hamburger, Y., (ed.). (2005). "Personality and the Internet," in *The Social Net: Human Behavior in Cyberspace* (New York, NY: Oxford University Press), 27–55.

Amichai-Hamburger, Y., and Vinitzky, G. (2010). Social network use and personality. *Comput. Human Behav.* 26, 1289–1295. doi:10.1016/j.chb.2010.03.018

Amichai-Hamburger, Y., Wainapel, G., and Fox, S. (2002). "On the Internet no one knows I'm an introvert": extroversion, neuroticism, and Internet interaction. *Cyberpsychol. Behav.* 5, 125–128. doi:10.1089/109493102753770507

Amiel, T., and Sargent, S. (2004). Individual differences in Internet usage motives. *Comput. Human Behav.* 20, 711–726. doi:10.1016/j.chb.2004.09.002

Barabasi, A. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* 435, 207–211. doi:10.1038/nature03459

Correa, T., Bachmann, I., Hinsley, A. W., and Gil de Zúñiga, H. (2013). "Personality and social media use," in *Organizations and Social Networking: Utilizing Social Media to Engage Consumers*, Vol. 2013, eds E. Y. Li, S. Loh, E. Cain, and L. Fabiana (Hershey, PA: Business Science Reference), 41–61.

Correa, T., Hinsley, A., and De Zuniga, H. (2010). Who interacts on the Web? The intersection of users' personality and social media use. *Comput. Human Behav.* 26, 247–253. doi:10.1016/j.chb.2009.09.003

Davis, J. M., and Yi, M. Y. (2012). User disposition and extent of Web utilization: a trait hierarchy approach. *Int. J. Hum. Comput. Stud.* 70, 346–363. doi:10.1016/j.ijhcs.2011.12.003

de Oliveira, R., Cherubini, M., and Oliver, N. (2013). Influence of personality on satisfaction with mobile phone services. *ACM Trans. Comput. Interact.* 20, 10. doi:10.1145/2463579.2463581

de Oliveira, R., Karatzoglou, A., Cerezo, P. C., Armenta Lopez de Vicua, A., and Oliver, N. (2011). "Towards a psychographic user model from mobile phone usage," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2191–2196.

Delgado-Gómez, D., Sukno, F., Aguado, D., Santacruz, C., and Artés-Rodriguez, A. (2010). Individual identification using personality traits. *J. Netw. Comput. Appl* 33, 293–299. doi:10.1016/j.jnca.2009.12.009

Golbeck, J., Robles, C., and Turner, K. (2011). "Predicting personality with social media," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems* (New York: ACM), 253–262.

Guadagno, R., Okdie, B., and Eno, C. (2008). Who blogs? Personality predictors of blogging. *Comput. Human Behav.* 24, 1993–2004. doi:10.1016/j.chb.2007.09.001

Hamburger, Y. A., and Ben-Artzi, E. (2000). The relationship between extraversion and neuroticism and the different uses of the Internet.pdf. *Comput. Human Behav.* 16, 441–449. doi:10.1016/S0747-5632(00)00017-0

Herder, E. (2005). "Characterizations of user Web revisit behavior," in *Proceedings Workshop on Adaptivity and User Modeling in Interactive Systems* (Hershey: Idea Group Publishing), 1–6.

Herrmann, D., Gerber, C., Banse, C., and Federrath, H. (2012). "Analyzing characteristic host access patterns for re-identification of web user sessions," in *Information Security Technology for Applications*, eds A. Tuomas, J. Kimmo, and N. Kaisa (Espoo, Finland: Springer), 136–154.

Hooper, C., and Dix, A. (2013). Web science and human-computer interaction: forming a mutually supportive relationship. *Interactions* 20, 52–57. doi:10.1145/2451856.2451868

Joiner, R., Brosnan, M., and Duffield, J. (2007). The relationship between Internet identification, Internet anxiety and Internet use. *Comput. Human Behav.* 23, 1408–1420. doi:10.1016/j.chb.2005.03.002

Kotsiantis, S., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: a review of classification techniques. *Informatical* 31, 249–268.

Kumar, R., and Tomkins, A. (2010). "A characterization of online browsing behavior," in *Proceedings of the 19th International Conference World Wide Web* (North Carolina: ACM), 561–570.

Landers, R., and Lounsbury, J. (2006). An investigation of Big Five and narrow personality traits in relation to Internet usage. *Comput. Human Behav.* 22, 283–293. doi:10.1016/j.chb.2004.06.001

Lim, M. J.-H., Negnevitsky, M., and Hartnett, J. (2006). Personality trait based simulation model of the email system. *Int. J. Netw. Secur.* 3, 164–182.

Matthews, G., Deary, I. J., and Whiteman, M. C. (2003). *Personality Traits*. Cincinnati: University of Cincinnati.

Moore, K., and McElroy, J. (2012). The influence of personality on Facebook usage, wall postings, and regret. *Comput. Human Behav.* 28, 267–274. doi:10.1016/j.chb.2011.09.009

Murray, D., and Durrell, K. (2000). "Inferring demographic attributes of anonymous Internet users," in *International WEBKDD'99 Workshop San Diego; CA; USA; August 15; 1999 Revised Papers*, ed. M. S. Brij Masand (Berlin, Heidelberg: Springer), 7–20.

Nguyen, T. T. T., and Armitage, G. (2008). "A survey of techniques for Internet traffic classification using machine learning", in *IEEE Communications Surveys & Tutorials,* IEEE, 56–76.

Oren, T., and Ghasem-Aghaee, N. (2003). "Personality representation processable in fuzzy logic for human behavior simulation," in *Society for Computer Simulation International: In Summer Computer Simulation Conference*, Society for Computer Simulation International, 11–18.

Ören, T. I., and Ghasem-Aghaee, N. (2003). "Towards fuzzy agents with dynamic personality for human behavior simulation," in *Proceedings of the 2003 Summer Computer Simulation Conference* (San Diego: SCS), 11–18.

Othman, F. M., Moh, T., and Yau, S. (2007). "Comparison of different classification techniques using WEKA for breast cancer," in *3rd Kuala Lumpur International Conference on Biomedical Engineering*, Vol. 15, 520–523.

Padmanabhan, B., and Yang, Y. (2007). *Clickprints on the Web: Are There Signatures in Web Browsing Data?* SSRN. Available at: http//ssrn.com/abstract=931057

Quercia, D., and Kosinski, M. (2011). "Our Twitter profiles, our selves: predicting personality with Twitter," in *Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, (Boston, MA: IEEE), 180–185.

Ross, C., Orr, E. E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., and Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Comput. Human Behav.* 25, 578–586. doi:10.1016/j.chb.2008.12.024

Salleh, N., Mendes, E., and Grundy, J. (2011). "The effects of openness to experience on pair programming in a higher education context," in *Proceeding of the 24th IEEE-CS Conference on Software Engineering Education and Training (CSEE&T)*, (Honolulu, HI: IEEE), 149–158.

Salleh, N., Mendes, E., and Grundy, J. (2014). Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments. *Empir. Softw. Eng.* 19, 714–752. doi:10.1007/s10664-012-9238-4

Salleh, N., Mendes, E., Grundy, J., and Burch, G. (2010a). "An empirical study of the effects of conscientiousness in pair programming using the five-factor personality model," in *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering*, Vol. 1 (Cape Town: ACM), 577–586.

Salleh, N., Mendes, E., Grundy, J., and Burch, G. (2010b). "The effects of neuroticism on pair programming: an empirical study in the higher education context," in *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (Bolzano: ACM), 22.

Samarein, Z. A., Far, N. S., Yekleh, M., Tahmasebi, S., Ramezani, Y. V., and Sandi, L. (2013). Relationship between personality traits and Internet addiction of students at Kharazmi University. *Int. J. Psychol. Behav. Res.* 2, 10–17.

Schrammel, J., Köffel, C., and Tscheligi, M. (2009). "Personality traits, usage patterns and information disclosure in online communities," in *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, (Cambridge: ACM), 169–174.

Swickert, R., Hittner, J., Harris, J., and Herring, J. (2002). Relationships among Internet use, personality, and social support. *Comput. Human Behav.* 18, 437–451. doi:10.1016/S0747-5632(01)00054-1

Tan, W., and Yang, C. (2012). "Personality trait predictors of usage of Internet services," in *International Conference on Economics, Business Innovation (IIPEDR)*, Vol. 38 (Singapore: ACSIT Press), 185–190.

Tan, W., and Yang, C. (2014). Internet applications use and personality. *Telematics Inform.* 31, 27–38. doi:10.1016/j.tele.2013.02.006

Yang, Y., and Padmanabhan, B. (2010). Toward user patterns for online security: observation time and online user identification. *Decis. Support Syst.* 48, 548–558. doi:10.1016/j.dss.2009.11.005

Yang, Y. C. (2010). Web user behavioral profiling for user identification. *Decis. Support Syst.* 49, 261–271. doi:10.1016/j.dss.2010.03.001

Young, K. S., and Rodgers, R. C. (1998). "Internet addiction: personality traits associated with its development," in *69th Annual Meeting of the Eastern Psychological Association*, Boston, MA, 1–6.

Yue, C., Kan, S., Xiaomeng, H., and Zhen, L. (2010). "Psychological influences of blogging: blog use, personality trait and self-concept," in *2010 IEEE 2nd Symposium Web Society*, IEEE, 71–76.

Zhou, T., Han, X., and Wang, B. (2008). Towards the understanding of human dynamics. *Sci. Matters Humanit. Complex Syst.* 1, 207–233. doi:10.1142/9789812835949_0012

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.