I Feel it in Your Fingers: Inference of Self-Assessed Personality Traits from Keystroke Dynamics in Dyadic Interactive Chats

Abeer A.N. Buker[1,2] and Alessandro Vinciarelli[1]
[1]*School of Computing Science*, *University of Glasgow*, Glasgow, UK
[2]*Computer Science Department*, *Imam Abdulrahman Bin Faisal University*, Dammam, Saudi Arabia
a.buker.1@research.gla.ac.uk, Alessandro.Vinciarelli@glasgow.ac.uk

***Abstract*—The question at the core of this work is whether it is possible to infer self-assessed personality traits from keystroke dynamics (the way people type on a keyboard). The experiments were performed over a corpus of 30 dyadic chats, 60 participants in total, collected through a text-based chat interface similar to those available in popular products (e.g., Skype). The results show that keystroke dynamics (typing speed, frequency of deletions, etc.) allow one to infer whether someone is below median or not along the Big Five personality traits. In particular, it was possible to achieve F1 Scores up to 72% depending on the trait. To the best of our knowledge, this is the first work aimed at recognizing personality traits through analysis of keystroke dynamics.**

***Index Terms*—keystroke dynamics, personality traits, social signal processing, personality computing, Big Five**

I. Introduction

After the advent of Internet, the term "*chat*" refers not only to an informal conversation, but also to different forms of text-based online interaction. In the most general case, chats involve multiple users that either interact synchronously or post messages that can be read and answered asynchronously by others [1]. In dyadic chats, the focus of this work, there are only two participants and the interaction is synchronous. Such a case, far from being rare, applies to a wide range of settings and, in particular, to interactions taking place through popular platforms such as *WhatsApp* or *Messenger*. In this respect, dyadic chats appear "*to fall somewhere in between spoken and written language [...] it is a rare case when writing necessitates another person to co-construct the dialogue*" [2]. This means that dyadic chats can be thought of as conversations that take place through a text-based interface rather than face-to-face.

The goal of this work is to show that it is possible to infer self-assessed personality traits, at least to a certain extent, from keystroke dynamics, i.e., from the way people type on a keyboard while being involved in a dyadic chat. In particular, the article proposes experiments in which keystroke dynamics (typing speed, frequency of deletions, punctuation, etc.) allow one to infer whether a chat participant is below median or not along the Big Five traits [3]. The results show that such an inference can be performed with F1 Score up to 72% depending on the particular trait. The experiments were performed over a corpus of 30 dyadic chats each involving two unacquainted users (60 participants in total). To the best of our knowledge, this is the first work aimed at inferring personality-relevant information from keystroke dynamics (see Section II).

According to the terminology proposed in [4], the inference of self-assessed traits is referred to as *Automatic Personality Recognition* (APR). The term "*self-assessed*" means that the experiment participants fill questionnaires designed to capture information about their own personalities. Self-assessment questionnaires were shown to be sensitive to multiple biases, and "*[...] accuracy is not the only motive shaping self-perceptions [...] the other powerful motives are consistency seeking, self-enhancement, and self-presentation*" [5]. In other words, self-assessments might be affected by noise due to, e.g., attempts to convey good impressions. This makes APR challenging because the relationship between keystroke dynamics and self-assessed traits becomes less consistent.

However, self-assessment questionnaires are widely used in practice because, despite the noise mentioned above, they are effective in practical problems [6]. Furthermore, they are predictive of a wide range of important life aspects, including "*happiness, physical and psychological health, spirituality, and identity at an individual level; [...] quality of relationships with peers, family, and romantic others at an interpersonal level; [...]*" [7]). In this respect, the inference of self-assessed traits is an important problem because it can provide useful insights about individuals under observation.

In addition to the above, this work is important because dyadic chats attract increasingly more interest in everyday life. In 2010, marketing analyses showed that "*nearly one in five online US consumers has used chat for customer service in the past 12 months*" [8]. More recently, The Pew Research Center - probably the most important institution monitoring the use of digital technologies - showed that 36% of smartphone users in the USA communicate through messaging apps [9]. In other words, there is evidence that people interact increasingly more frequently through online chat systems like those considered in this work. Finally, a reason of interest is that it not fully clear how people exchange *social signals* - cues aimed at conveying social facts [10] - when technological platforms do not allow the use of nonverbal behavior (e.g., facial expressions, gestures and vocalizations) [11].

The rest of the article is organized as follows: Section II

surveys previous work, Section III shows the data, Section IV describes the APR approach, Section V reports on experiments and results, and the final Section VI draws some conclusions.

II. Survey of Previous Work

To the best of our knowledge, the psychological literature does not provide indications about the interplay between personality traits and keystroke dynamics. However, the literature shows that there are relationships between personality and two phenomena likely to be involved in typing (especially when being involved in an interactive chat aimed at addressing a task like in this work), namely cognitive load [12] and tendency to experience certain emotions rather than others [13]. In this respect, it is possible to expect that the traits can leave traces in terms of the way people type.

From a technological point of view, most of the works based on keystroke dynamics revolve around approaches for biometrics, i.e., around the attempt to recognize the identity of people through the way they type [14]. However, this survey focuses on emotion [15] and gender recognition, two problems that, while having been investigated less than biometrics, are more relevant to this work. Other potentially relevant problems (e.g., age recognition [16], stress detection [17]), or engagement assessment [18] were addressed only to a limited extent and, therefore, are not considered.

Several works have shown that there is a significant interplay between emotions and typing speed. In particular, the experiments in [19], involving 52 people asked to type a predefined text, show that there is a significant correlation between speed and *arousal* (the "intensity" of an emotion). Similarly, in the case of unconstrained texts, experiments based on touch-screen keyboards have shown that error rate and speed tend to increase with both arousal and *valence* (the intrinsic attractiveness or averseness of an emotion) [20]. It is probably based on this type of observations that emotion recognition approaches often use features designed to measure typing speed [21]–[23].

For example, the approach proposed in [21] analyses typing patterns collected during the everyday activities of 25 people. The patterns were represented mainly in terms of how much time people tend to press a given key and how much time passes between two consecutive keys being pressed. The recognition task involved 7 emotion states (mostly overlapping with the six basic emotions) and the accuracies ranged roughly between 70% and 85% depending on the particular emotion. Similarly, the work presented in [22] analyses any text typed by 12 experiment participants during their everyday activities. The experiments targeted the recognition of 12 emotional states (e.g., *nervousness* and *relaxation*) and the recognition rates were up to 85%. The entire feature set was aimed at measuring how quickly people type different d-graphs, i.e., sequences of d consecutive keys (typically $d \leq 3$ [14], [24]). Finally, the work in [23] focuses on programmers working with or without time pressure, two conditions expected to elicit different emotions. The results show that the typing speed changes significantly across the two conditions, in line with the other works in the literature.

In the case of gender recognition, the task of automatically recognizing the gender of a person that types, no features appear to be more suitable or more effective than others [25]–[28]. For this reason, the experiments presented in [25] focused on feature selection methodologies. The results showed that a few hundreds features were still necessary to reach an accuracy of 95%. Earlier work [26] addressed the problem of recognizing gender through language-independent features. The experiments, performed over a collection of texts typed by 17 persons, showed that gender can be recognized with accuracy up to 75%. The proposed approach estimated the probability of every key being pressed by a person of a given gender and used a Naive Bayes classifier. The approach proposed in [27] was tested over data produced by 1,517 people during their daily activities (programming, e-mail writing, etc.). The feature set included around 2,000 features mostly related to the timing while typing key d-graphs. The results showed F-scores around 75%. In [28], the experiments showed an accuracy higher than 90% through the use of features expected to account for nonverbal aspects of typing.

III. The Data

The experiments of this work involved 60 participants randomly paired to form 30 dyads. The participants of each dyad were asked to interact through a text-chat interface similar to those available in popular products such as *Skype*, *Zoom*, *WhatsApp* or *Messenger*. The main element of the interface was a *textbox* where it was possible to type a message and then send it to the interlocutor by pushing the "*Enter*" key on the keyboard. The history of the messages sent during a particular chat was visible on the screen in the same way as it happens in the products mentioned above. The participants could see the messages typed by their interlocutors only after these latter pushed the Enter key. The motivation behind the use of dyadic chats is that one-to-one conversations are the "*primary site of human sociality*" [29]. Furthermore, all works in Personality Computing focus on one specific setting [4]. This makes behavior of different individuals comparable and, correspondingly, it makes it possible to associate variance in behavior and variance in traits through statistical methodologies.

The participants were instructed to use only laptop or desktop computers to limit keyboard size or response effects as much as possible. They were provided with a text describing the *Winter Survival Task*, a scenario in which the participants must identify items that increase the chances of survival after a plane crash in a polar area [30]. In particular, the participants were given a list of 12 items[1] and were asked to make a consensual decision about each of them ("*yes*" if it increases the chances of survival and "*no*" if it does not). The main advantage of the scenario is that users, on average, do not

[1] Steel wool, axe, pistol, butter can, newspaper, lighter without fuel, clothing, canvas, airmap, whisky, compass, chocolate.

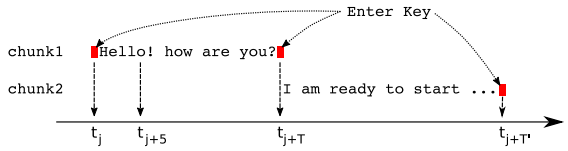

Fig. 1. The figure shows how the sequence of tokens is segmented into chunks.

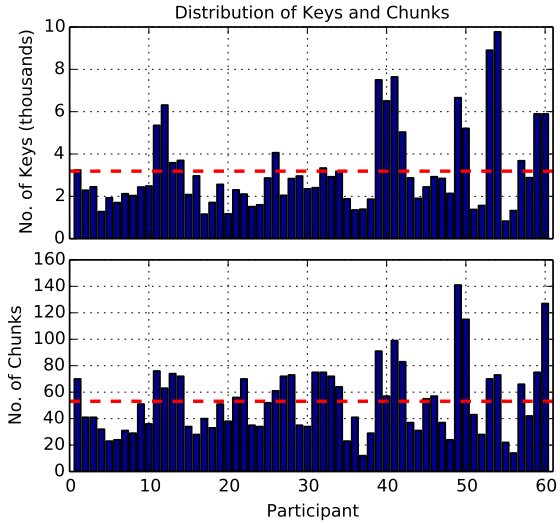

Fig. 2. The figure shows the distribution of keys (upper chart) and chunks (lower chart) across the experiment participants. The dashed red lines correspond to the averages.

hold any expertise relevant to the topic. Thus, the outcome of the conversations depends on social dynamics rather than on actual knowledge about the problem. The participants were asked to complete the task as quickly as possible and were not allowed to use any source of information (web, books, etc.). Furthermore, the participants of each chat were fully unacquainted and were not given any information about their interlocutors.

The chat interface used in the experiments was equipped with a key-logging platform that recorded and timestamped every key the users pressed. Therefore, the collected data can be thought of as a sequence of pairs (k_i, t_i), with $i = 1, \ldots, M$, where k_i is the i^{th} key that was pressed, t_i is the time at which key k_i had been pressed and M is the total number of keys that were pressed. Overall, the corpus includes $191,375$ keys and the sequence of keys can be segmented into *tokens*, i.e., strings of consecutive non-blank characters delimited by blank spaces. During the chats, the participants had to press the *Enter* key to send a sequence of tokens to their interlocutor (see above). Some people pressed the Enter key every few words, while others did it only after having written long and articulated messages. In both cases, the chat can be segmented into *chunks*, i.e., sequences of tokens delimited by Enter keys (see Figure 1). The chunks are, in total, $3,177$ and they are the analysis unit of the experiments (see Section IV for more

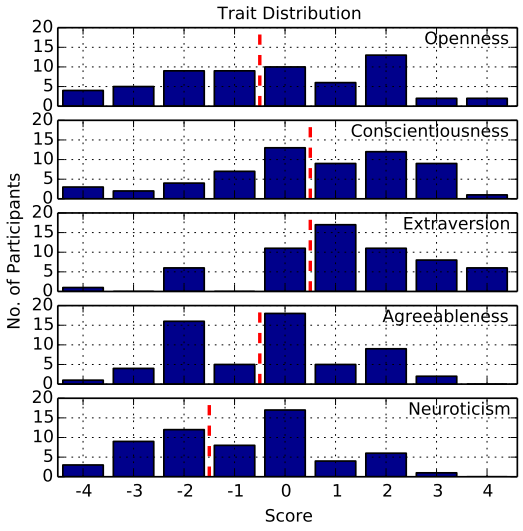

Fig. 3. The charts show the distribution of the different trait scores across the participants of the experiments. The red vertical line corresponds to the separation between classes A (equal to the median or above) and B (below median).

details). The charts of Figure 2 show the distribution of keys and chunks across the 60 experiment participants.

In the days before participating in the experiments, all participants filled the Big-Five Inventory 10 questionnaire (BFI-10) [31], an instrument allowing personality self-assessment in terms of the Big-Five traits [3]:

- Extraversion: Active, Assertive, Energetic, etc.
- Agreeableness: Appreciative, Kind, Generous, etc.
- Conscientiousness: Effcient, Organized, Planful, etc.
- Neuroticism: Anxious, Self-pitying, Tense, etc.
- Openness: Artistic, Curious, Imaginative, etc.

The motivation behind the choice of the Big-Five personality model is that it is the most widely applied in research and practice [32]. In the particular case of computing applications, the model has the major advantage of representing personality as a five-dimensional vector, a format particularly suitable for computer processing [4]. Each of the five vector components is a score that measures how well the adjectives in the list above describe the behavior of an individual (the higher the score for a trait, the better the adjectives associated to the trait describe well the behavior of a person). In this respect, similar vectors correspond to similar personalities.

Figure 3 shows the distribution of the Big-Five scores across the participants. Every score is a discrete variable in the range $[-4, 4]$ and the charts show that most of the participants tend to cluster around the median. Such a phenomenon is common and probably reflects the tendency of social cognition to categorize (e.g., *extravert* vs *introvert*), i.e., to represent self and others in terms of discrete categories rather than distribution along a continuum [33]. For this reason, the scores have been binarized and the participants have been assigned to one of two possible classes for each trait, namely *equal to the median or above* (A) and *below median* (B). The medians

Name	Description	Type	Refs
Backspace (ρ)	Density of "*Backspace*" keys	Density	[35]
Exclamation Marks (ρ)	Density of "*!*" key	Density	[36]
Emoticons (ρ)	Density of key sequences corresponding to emoticons	Density	[36]
Uppercase Tokens (ρ)	Density of tokens written in uppercase letters	Density	[37]–[39]
Uppercase Letters (ρ)	Density of keys corresponding to uppercase letters	Density	[36]
Points (ρ)	Density of keys corresponding to "."	Density	[36]
Question Marks (ρ)	Density of keys corresponding to "*?*"	Density	[36]
Non-Alphabetic Keys (ρ)	Density of keys that do not correspond to letters	Density	[22]
Suspension Points	Average distance between consecutive "." keys	Density	[40]
Chunk Length	Total number N of keys in the chunk	Global	[28], [36], [41]
Number of Tokens	Total number of tokens in the chunk	Global	[28], [41]
Chunk Duration	$\Delta t = t_N - t_1$, with t_k time at which key k is pressed	Temporal	[28], [36], [41]
Backspace Time	Total time spent in pressing the "*Backspace*" key	Temporal	[35]
Typing Speed	Number of keys pressed per unit of time	Temporal	[42]
Median Latency Time	Median of time between two keys pressed consecutively	Temporal	[35]

TABLE II

THE TABLE SHOWS THE 15 FEATURES EXTRACTED FROM EVERY CHUNK. FOR EVERY FEATURE, THE TABLE PROVIDES NAME, DESCRIPTION, TYPE AND REFERENCES WHERE THE FEATURES HAVE BEEN USED. THE SYMBOL ρ STANDS FOR DENSITY.

Trait	Male	Female	Avg. Age	Total
Ope. (A)	14	19	25.3± 2.6	33
Ope. (B)	11	16	25.7 ± 3.9	27
Con. (A)	10	21	25.9 ± 3.4	31
Con. (B)	15	14	25.0 ± 2.8	29
Ext. (A)	21	21	25.4 ± 3.4	42
Ext. (B)	4	14	25.5 ± 2.7	18
Agr. (A)	16	18	25.1 ± 3.0	34
Agr. (B)	9	17	26.0 ± 3.3	26
Neu. (A)	17	19	25.3 ± 2.6	36
Neu. (B)	8	16	25.7 ± 3.9	24

TABLE I

THE TABLE SHOWS THE NUMBER OF PARTICIPANTS BELONGING TO EACH CLASS AND TRAIT, A STANDS FOR *equal to the median or above* AND B STANDS FOR *below median*. ACCORDING TO A χ^2 TEST WITH BONFERRONI CORRECTION, THERE IS NO STATISTICALLY SIGNIFICANT DIFFERENCE BETWEEN A AND B CLASSES IN TERMS OF GENDER DISTRIBUTION. SIMILARLY, THERE IS NO STATISTICALLY SIGNIFICANT DIFFERENCE IN TERMS OF AGE (ACCORDING TO A t-TEST).

observed in the data appear to be the same, within statistical fluctuations as those observed in the population of the UK, the country where the data was collected [34]. Such a practice is in line with the literature showing that most personality recognition approaches perform the inference of the traits as a binary classification [4]. Table I shows the distribution of the participants across classes and genders for all traits.

The data collection was performed according to the ethical regulations of the British Psychological Association. In particular, chat users had the right to leave the experiment at any moment and without giving any explanation. Furthermore, participants were given 15 days to delete, partially or totally, the data collected during their chat. All participants signed an informed consent letter explaining the use of the data. These are stored in password-protected repositories and the format is fully anonymized (there is no possibility to establish a connection between the data and the identity of the participants).

IV. THE APPROACH

The approach used in the experiments includes two main steps, namely *feature extraction* and *recognition*. The goal of the first step is to convert every chunk (see Section III) into a vector of physical measurements that account for keystroke dynamics. The goal of the second step is to assign every vector to a class that can correspond to being below median or not along each of the Big-Five traits.

A. Feature Extraction

The feature extraction step converts every chunk into a vector in which the components, the features, are physical measurements inspired by either *biometrics* (the domain aimed at automatically verifying the identity of an individual) or *authorship attribution* (the field aimed at automatically identifying the author of a text). The assumption behind the choice is that these features have been designed to capture effectively the identity of a person and, therefore, they are likely to account for salient individual characteristics such as personality traits (see Section II). The features are 15 (see Table II) and can be grouped into three main categories, namely *densities* (how frequently certain keys are pressed), *temporal* (time-related aspects of typing) and *global* (characteristics of each chunk as a whole).

The goal of the density features is to measure the tendency of a chat user to display a given typing pattern, whether it corresponds to an individual key (e.g., question and exclamation marks), a sequence of keys (e.g., emoticons or uppercase tokens) or a class of keys (e.g., upper case letters or non-alphabetic characters). The density ρ_p of a pattern p can be calculated as follows: $\rho_p = \frac{n_p}{N_p}$, where n_p is the number of times the pattern p appears in the chunk and N_p is the total number of relevant patterns in the chunk, i.e., the total number of keys when p is a key or a class of keys, the total number of key sequences when p is a sequence of keys, etc.

The temporal features take into account that the chat interface used in the experiments is equipped with a key-logging platform and, therefore, the data is not just a written text, but a sequence of pairs (k_i, t_i), $i = 1, \ldots, M$, where t_i is the time at which the i^{th} key has been pressed and M is the total number of keys in the chunk (see Section III). This makes it

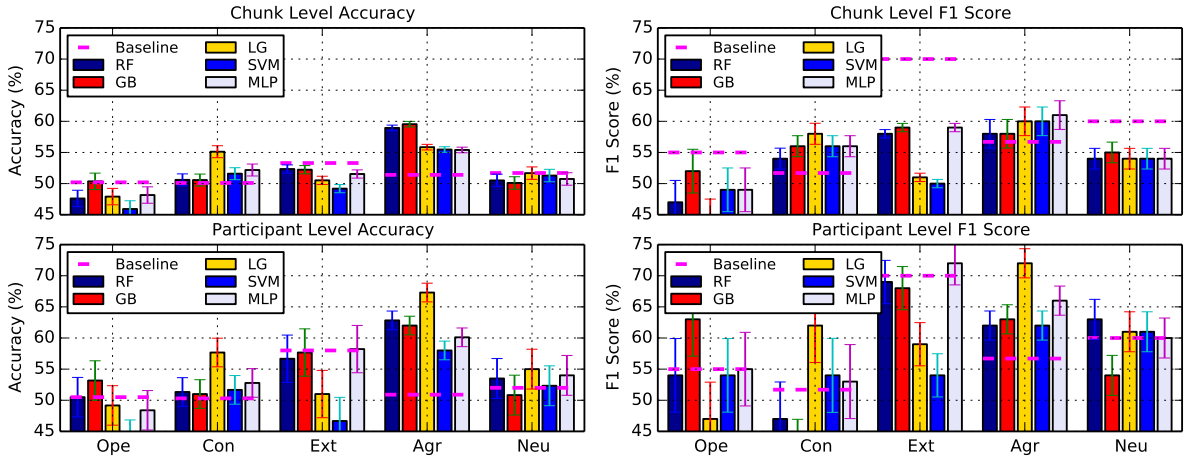

Fig. 4. The figure shows Accuracy and F1 Score for the five traits at chunk (upper charts) and participant (lower charts) level. The error bars correspond to the standard deviation over the R repetitions of the experiments ($R = 10$ in all cases except MLP, when $R = 30$). The horizontal dashed lines correspond to the performance of the random system used as a baseline (see Section IV). The lower right chart (F1 Score at Participant Level) is the one that matters most from an application point of view because the ultimate goal of the approach is to classify persons.

possible to measure how typing patterns distribute over time (e.g., the median latency time between consecutive keys) or how much time has been spent in pressing a given sequence of keys (e.g., the length of the time interval between the first and the last of several pressings of the backspace key). Finally global characteristics measure aspects of the chunk that result from the whole set of keys that have been pressed (e.g., the chunk length or the number of tokens).

None of the features take into account the actual content of the chunks, i.e., what people type. From a technological point of view, the main advantage of such a choice is that the approach is language-independent. From the scientific point of view, the reason of the choice is that nonverbal behavior has been widely shown to convey socially and psychologically-relevant information, both in face-to-face [43] and technology-mediated settings [11]. In other words, non-verbal aspects of typing are expected to act as personality markers that, being displayed outside conscious control, are more likely to be *honest* [44], i.e., to leak genuine and reliable information about the person that types.

B. Recognition

The recognition problems addressed in this work are binary classifications, i.e., problems in which a feature vector, automatically extracted from a chunk, must be assigned to one of two possible classes (c_1 and c_2 hereafter). In particular, there are five binary classification problems in which the two classes correspond to being below median or not along each of the Big Five traits (see Section III). The classification approaches selected for the experiments were *Random Forests* (RF) [45], *Gradient Boosting* (GB) [46], *Logistic Regression* (LG) [47], *Support Vector Machines* (SVM) [48] and *Multi Layer Perceptrons* [49]. The main reason behind the choice of these classifiers is that they are the most commonly used in the Personality Computing literature [4] and, therefore, they can provide reliable baseline results. In addition, unlike other methodologies, the classifiers above do not require large amounts of material to be trained.

Given that the classification was performed at the level of individual chunks, there were multiple classification outcomes for every participant. The aggregation of these chunk-level outcomes was performed through a *majority vote*, i.e., by assigning a participant to the class her or his chunks were most frequently assigned to: $\hat{c} = \arg\max_{c \in \{c_1, c_2\}} n(c)$, where $n(c)$ is the number of chunks assigned to class c for a given participant. All classifiers were compared to a baseline system that assigns an unseen chunk to class c_k with probability corresponding to its prior $p(c_k)$. The accuracy of such a system is $\hat{\alpha} = p(c_1)^2 + p(c_2)^2$, while its Precision, Recall and F1 score all correspond to the prior of the class corresponding to the positive case.

V. Experiments and Results

The experiment participants were randomly split into four disjoint subsets each including 15 persons. All data generated by the people belonging to one of these subsets were inserted into a fold so that, at the end of the process, the data were distributed across 4 disjoint folds. This made it possible to perform the experiments according to a k-fold protocol with $k = 4$. Given that all the data of a participant were in the same fold, the experiments were *person independent*, meaning that the same person was never represented in both training and test set. This ensures that the approach actually recognizes whether a person is below median along the Big-Five traits and not just the identity of the participants.

Table I shows that the classes are imbalanced, i.e., that the samples do not distribute uniformly across the classes of the five binary classification tasks (one per trait). For this reason, in the case of GB, the experiments involved the application of the *Synthetic Over-Sampling Minority Technique*

(SMOTE) [50]. SMOTE considers the T nearest neighbors $\vec{x}_i$ of every sample $\vec{x}$ in the minority class and, for each of them, it creates a synthetic sample $\vec{x}' = \vec{x} + r \cdot (\vec{x}_i - \vec{x})$, where r is a random number between 0 and 1. The parameter T was set to have, for each binary classification task, the same number of samples in both classes. The use of such an algorithm ensures that GB does not always assigns test samples to the most represented class.

Figure 4 shows the results at both chunk and participant level, i.e., both before and after the majority vote aimed at aggregating the decisions made at the level of individual chunks (see Section IV). Given that training the classifiers requires a random step (for initialization and for distributing the data across folds), the experiments were repeated $R = 10$ times and, correspondingly, all performance metrics are reported in terms of average and standard deviation over the R repetitions. Overall, the charts suggest that both accuracy (17 times out of 25) and F1 Scores (21 times out of 25) tend to be greater at person level than at chunk level. This suggests that the majority vote is an effective strategy to combine the decisions made at the level of individual chunks. The rest of this section focuses on the F1 Score at the participant level (lower right chart in Figure 4) because this is the metric that matters most from an application point of view. In fact, the goal of the work is not to classify individual chunks, but to predict whether a person is below median or not along the traits.

According to a two-tailed one sample t-test, there are three traits for which there is at least one classifier that improves over the baseline random approach ($p < 0.01$ in all cases): Openness, Conscientiousness and Agreeableness. In two of these three cases, the best performing classifier is LG (Conscientiousness and Agreeableness), while in the remaining case it is GB (Openness). In the case of Agreeableness, all classifiers perform better than the baseline, but LG is better than the others to a statistically significant extent ($p < 0.0001$ according to a two-tailed t-test). Overall, the performances of Figure 4 appear to be similar to those observed in other works aimed at Automatic Personality Recognition (the task of predicting self-assessed personality traits) [4]. However, the comparison should not be considered fully rigorous because, to the best of our knowledge, this is the first experiment that performs the task using keystroke dynamics.

The probable reason why the approach performs over chance only for certain traits is that, according to Personality Psychology, the possibility of achieving satisfactory performances depends on "*Relevance (i.e., the environment must allow the person to express the trait) and Availability (i.e., the trait must be perceptible to others)*" [51]. This means that every situation allows one to manifest only some of the traits and, therefore, only these can be recognized effectively. In this respect, Figure 4 suggests that, in the experiments of this work, Agreeableness is the most relevant and available trait (all classifiers recognize it beyond chance), followed by Conscientiousness and Openness.

One possible explanation behind relevance and availability of Conscientiousness is that, according to the literature, "*[...] there are two dimensions that underlie most judgments of traits, people, groups, and cultures [...] the first makes reference to attributes such as competence [...] and the second to warmth [...]*" [52], where "*competence*" is one of the terms that the literature uses to define Conscientiousness. In addition, the dyadic chats revolve around a scenario that requires the participants to complete a task as quickly as possible and, therefore, skills and characteristics underlying Conscientiousness (being efficient, organized, planful, etc.) might emerge with particular evidence. In a similar vein, the need to complete the Winter Survival Task, a problem people are typically not competent or knowledgeable about, might explain relevance and availability of Openness. In fact, such a trait corresponds to creativity and imaginativeness skills that people might tap when trying to solve problems they are not familiar with.

Name	Ope.	Con.	Ext.	Agr.	Neu.
Chunk Length	-	↓	-	↓	-
Chunk Duration	-	↓	-	↓	-
Number of Tokens	-	↓	↓	↓	-
Backspace (ρ)	-	-	-	-	-
Backspace Time	-	↓	↑	↓	-
Typing Speed	↓	↑	↓	↑	↓
Median Latency Time	↑	↓	↑	-	-
Exclamation Marks (ρ)	-	-	-	-	↓
Emoticons (ρ)	-	-	↑	-	-
Uppercase Tokens (ρ)	-	-	-	-	-
Uppercase Letters (ρ)	-	-	-	↓	-
Points (ρ)	-	↓	-	-	-
Question Marks (ρ)	-	↑	-	-	-
Non-Alphabetic Keys (ρ)	-	-	-	↓	-
Suspension Points	-	↑	-	-	-

TABLE III

FOR EVERY FEATURE AND EVERY TRAIT, A t-TEST HAS BEEN APPLIED TO COMPARE THE MEANS OF THE FEATURE VALUES BETWEEN PEOPLE BELOW MEDIAN AND THE OTHERS (THE SIGNIFICANCE LEVEL IS 5% AND THE FALSE DISCOVERY RATE CORRECTION [53] HAS BEEN USED TO DEAL WITH THE MULTIPLE COMPARISONS PROBLEM). THE SYMBOL "↑" MEANS THAT THE AVERAGE VALUE OF THE FEATURE IS HIGHER, TO A STATISTICALLY SIGNIFICANT EXTENT, FOR PEOPLE ABOVE OR EQUAL TO MEDIAN (VICE VERSA FOR SYMBOL "↓").

To the best of our knowledge, the literature does not provide indications that can explain the good performances observed for Agreeableness. However, one possible explanation is that several features account for how frequently participants push the "Enter" key, thus making the text they type visible to their interlocutors. People with high Agreeableness scores might tend to push the "Enter" key more frequently because this reduces the time others have to wait before they get a response. In fact, such a tendency to limit the chances of an unpleasant *doorbell effect* (the need to wait long time for a response without indications about what is happening) appears to be in line with the characteristics underlying the Agreeableness trait. In other words, it can be expected that highly Agreeable people worry more about their interlocutors and, therefore, they tend to make their texts visible more frequently.

Table III seems to confirm the hypothesis above by showing that people above or equal to median along Agreeableness actually tend to type shorter chunks (the Chunk Length is

lower to a statistically significant extent) that take less time to be completed (the Chunk Duration is lower to a statistically significant extent) and include less tokens, on average (the Number of Tokens is lower to a statistically significant extent). In addition, high Agreeableness participants tend to type faster (the Typing Speed is higher to a statistically significant extent), thus further confirming the attempt to limit as much as possible the time interlocutors need to wait before they can see the text being typed.

One interesting aspect of Table III is that the number of individual features for which there is a statistically significant effect depends on the trait. Such a number is large for two of the three recognized traits (Conscientiousness and Agreeableness) and this might further explain the results of Figure 4. However, the lowest number of effects corresponds to Openness, one of the traits recognized beyond chance. One possible explanation is that t-tests are performed individually for every feature, while the recognition is performed using a feature vector as a whole. In other words, even if there are multiple features for which there is a statistically significant difference, what really matters is that such differences distribute according to a pattern, i.e., that they are associated in such a way that feature vectors belonging to different classes actually tend to distribute in different regions of the feature space.

VI. Conclusions

To the best of our knowledge, this article presents the first APR work based on keystroke dynamics, i.e., on the way people type on a keyboard. The experiments involved 60 participants that were asked to interact with an unacquainted interlocutor through a text-based chat interface equipped with a key-logging platform. The results show that it is possible to recognize whether someone is below median or not along the Big-Five personality traits with F1 Score up to 72% depending on the particular trait. Overall, the performances appear to be similar to those obtained using other types of data (see [4] for an extensive survey). However, besides the effectiveness in predicting the traits, the key-result of the paper is the very possibility to infer personality-relevant information from keystroke dynamics.

The analysis of keystroke dynamics can be of help in a wider spectrum of application domains, including the detection of mental health issues (often addressed through the analysis of nonverbal behavior in the literature) or the analysis of social and psychological phenomena in interaction. Not surprisingly, the state-of-the-art presented in Section II reports on efforts aimed at emotion recognition. In all cases, the use of live-chats promises to be effective from an application point of view, especially when considering that, according to KBV Research, "*The Global Live Chat Software Market size is expected to reach $987.3 million by 2023, rising at a market growth of 7.3% [Compound Annual Growth Rate] during the forecast period*" (https://www.kbvresearch.com/live-chat-software-market/).

A direction for future work is to combine the classifiers used in the experiments in an ensemble. The main rationale behind such a choice is that none of the classifiers clearly outperforms the others and the best performing classifier tends to be different for every trait. Furthermore, the difference between the best performing classifiers for a given trait is not always statistically significant and, therefore, it can be expected that they mutually compensate for their errors (provided they are sufficiently *diverse* [54]). In a similar vein, it is possible to investigate multimodal approaches that take into account not only the way people type, but also the text they type. In fact, the literature shows that there is a major interplay between personality and lexical choices [3]. Therefore, the use of text-based approaches in combination with those proposed in this work can lead to performance improvements.

Besides the technological research avenues mentioned above, one important problem is how to deal with noise and biases inherent to self-assessment questionnaires (see Section I). According to the literature, these problems reduce the validity of questionnaires, but to an acceptable extent in most practical cases [6]. In other words, while being affected by noise, self-assessment scores might still be sufficiently reliable to address practical problems. However, one common approach is to collect alternative evidence that can help to corroborate or reject the results of the tests [55]. In the case of this work, Table III shows that there are tendencies associated to the traits (e.g., highly extravert people tend to include less tokens in their chunks). Based on the assumption that most people tend to provide honest self-assessments, non-reliable self-assessments might be detected through the statistical analysis of the misalignments between between observed behavior (e.g., a large number of tokens in the chunks) and self-assessed traits (e.g., a high Extraversion score).

Approaches like those presented in this work involve ethical issues that need to be addressed before any application outside the laboratory. One of the main implications of the experiments is that, while typing, people leak information about their personality traits without realizing it. This is in line with previous observations [56] showing that, whenever sharing data through a digital platform, people allow the inference of information they do not necessarily wish to disclose. This means that the notion of privacy should cover not only the content people produce (the focus of current privacy regulations), but also any information that can be inferred from the content (currently not protected or regulated).

References

[1] D. Uthus and D. Aha, "Multiparticipant chat analysis: A survey," *Artificial Intelligence*, vol. 199, pp. 106–121, 2013.

[2] M. Freiermuth, "Online chat," *The International Encyclopedia of Language and Social Interaction*, 2015.

[3] G. Saucier and L. Goldberg, "The language of personality: Lexical perspectives on the five-factor model," in *The Five-Factor Model of Personality*, J. Wiggins, Ed., 1996.

[4] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[5] D. Paulhus and S. Vazire, "The self-report method," in *Handbook of Research Methods in Personality Psychology*, R. Robins, R. Fraley, and R. Krueger, Eds. Gilford, 2007, pp. 224–239.

[6] G. Matthews, I. Deary, and M. Whiteman, *Personality Traits*. Cambridge University Press, 2009.

[7] D. Ozer and V. Benet-Martinez, "Personality and the prediction of consequential outcomes," *Annual Reviews of Psychology*, vol. 57, pp. 401–421, 2006.

[8] D. Clarkson, C. Johnson, E. Stark, and B. McGowan, "Making proactive chat work: Maximizing sales and service requires ongoing refinement," Forrester, Tech. Rep., 2010.

[9] M. Duggan, "Mobile mesaging and social media 2015," Pew Research Center, Tech. Rep., 2015.

[10] I. Poggi and F. D'Errico, "Social signals: a framework in terms of goals and beliefs," *Cognitive Processing*, vol. 13, no. 2, pp. 427–445, 2012.

[11] A. Vinciarelli and A. Pentland, "New social signals in a new interaction world: The next frontier for social signal processing," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 1, no. 2, pp. 10–17, 2015.

[12] V. Kumari, D. Ffytche, S. Williams, and J. Gray, "Personality predicts brain responses to cognitive demands," *Journal of Neuroscience*, vol. 24, no. 47, pp. 10 636–10 641, 2004.

[13] R. Reisenzein and H. Weber, "Personality and emotion," in *The Cambridge Handbook of Personality Psychology*, P. Corr and G. Matthews, Eds. Cambridge University Press, 2009, pp. 54–71.

[14] S. Banerjee and D. Woodard, "Biometric authentication and identification using keystroke dynamics: A survey," *Journal of Pattern Recognition Research*, vol. 7, no. 1, pp. 116–139, 2012.

[15] A. Kołakowska, "A review of emotion recognition methods based on keystroke dynamics and mouse movements," in *Proceedings of the IEEE International Conference on Human System Interactions*, 2013, pp. 548–555.

[16] I. Tsimperidis, S. Rostami, and V. Katos, "Age detection through keystroke dynamics from user authentication failures," *International Journal of Digital Crime and Forensics*, vol. 9, no. 1, pp. 1–16, 2017.

[17] S. Gunawardhane, P. De Silva, D. Kulathunga, and S. Arunatileka, "Non invasive human stress detection using key stroke dynamics and pattern variations," in *Proceedings of the International Conference on Advances in ICT for Emerging Regions*, 2013, pp. 240–247.

[18] S. Bixler, R.and D'Mello, "Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits," in *Proceedings of the International Conference on Intelligent User Interfaces*, 2013, pp. 225–234.

[19] P.-M. Lee, W.-H. Tsui, and T.-C. Hsiao, "The influence of emotion on keyboard typing: An experimental study using auditory stimuli," *PLOS ONE*, vol. 10, no. 6, pp. 1–16, 2015.

[20] M. Trojahn, F. Arndt, M. Weinmann, and F. Ortmeier, "Emotion recognition through keystroke dynamics on touchscreen keyboards." in *Proceedings of the International Conference on Enterprise Information Systems*, 2013, pp. 31–37.

[21] A. Nahin Nazm Haque, J. Alam, H. Mahmud, and K. Hasan, "Identifying emotion by keystroke dynamics and text pattern analysis," *Behaviour & Information Technology*, vol. 33, no. 9, pp. 987–996, 2014.

[22] C. Epp, M. Lippold, and R. Mandryk, "Identifying emotional states using keystroke dynamics," in *Proceedings of ACM CHI*, 2011, pp. 715–724.

[23] A. Kołakowska, "Towards detecting programmers' stress on the basis of keystroke dynamics," in *Proceedings of the Federated Conference on Computer Science and Information Systems*, 2016, pp. 1621–1626.

[24] F. Bergadano, D. Gunetti, and C. Picardi, "User authentication through keystroke dynamics," *ACM Transactions on Information and System Security*, vol. 5, no. 4, pp. 367–397, 2002.

[25] I. Tsimperidis, A. Arampatzis, and A. Karakos, "Keystroke dynamics features for gender recognition," *Digital Investigation*, vol. 24, pp. 4–10, 2018.

[26] I. Tsimperidis, V. Katos, and N. Clarke, "Language-independent gender identification through keystroke analysis," *Information & Computer Security*, vol. 23, no. 3, pp. 286–301, 2015.

[27] A. Pentel, "Predicting age and gender by keystroke dynamics and mouse patterns," in *Proceedings of the International Conference on User Modeling, Adaptation and Personalization*, 2017, pp. 381–385.

[28] A. Buker, G. Roffo, and A. Vinciarelli, "Type like a man! Inferring gender from keystroke dynamics in live-chats," *IEEE Intelligent Systems*, vol. 34, no. 6, 2019.

[29] E. Schegloff, "Analyzing single episodes of interaction: An exercise in conversation analysis," *Social Psychology Quarterly*, pp. 101–114, 1987.

[30] M. Joshi, E. Davis, R. Kathuria, and C. Weidner, "Experiential learning process: Exploring teaching and learning of strategic management framework through the winter survival exercise," *Journal of Management Education*, vol. 29, no. 5, pp. 672–695, 2005.

[31] B. Rammstedt and O. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[32] G. Matthews, I. Deary, and M. Whiteman, *Personality traits*. Cambridge University Press, 2003.

[33] C. Macrae and G. Bodenhausen, "Social cognition: Thinking categorically about others," *Annual Review of Psychology*, vol. 51, no. 1, pp. 93–120, 2000.

[34] P. Rentfrow, M. Jokela, and M. Lamb, "Regional personality differences in Great Britain," *PloS one*, vol. 10, no. 3, p. e0122245, 2015.

[35] L. Vizer, "Detecting cognitive and physical stress through typing behavior," in *Proceedings of ACM CHI*, 2009, pp. 3113–3116.

[36] M. Brooks, K. Kuksenok, M. Torkildson, D. Perry, J. Robinson, T. Scott, O. Anicello, A. Zukowski, P. Harris, and C. Aragon, "Statistical affect detection in collaborative chat," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, 2013, pp. 317–328.

[37] C. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 579–586.

[38] A. Yadollahi, A. Shahraki, and O. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Computing Surveys*, vol. 50, no. 2, pp. 25:1–25:33, 2017.

[39] M. Riordan and R. Kreuz, "Cues in computer-mediated communication: A corpus analysis," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1806–1817, 2010.

[40] M. Cristani, G. Roffo, C. Segalin, L. Bazzani, A. Vinciarelli, and V. Murino, "Conversationally-inspired stylometric features for authorship attribution in instant messaging," in *Proceedings of the ACM International Conference on Multimedia*, 2012, pp. 1121–1124.

[41] M. Weinel, M. Bannert, J. Zumbach, H. Hoppe, and N. Malzahn, "A closer look on social presence as a causing factor in computer-mediated collaboration," *Computers in Human Behavior*, vol. 27, no. 1, pp. 513–521, 2011.

[42] Y. Lim, A. Ayesh, and M. Stacey, "Using mouse and keyboard dynamics to detect cognitive stress during mental arithmetic," in *Intelligent Systems in Science and Information*, K. Arai, S. Kapoor, and R. Bhatia, Eds. Springer International Publishing, 2015, pp. 335–350.

[43] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[44] A. Pentland, *Honest Signals*. MIT Press, 2007.

[45] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] J. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[47] C. Bishop, *Pattern recognition and machine learning*. Springer Verlag, 2006.

[48] L. Wang, *Support Vector Machines: theory and applications*. Springer Science & Business Media, 2005.

[49] E. Charniak, *Inreoduction to Deep Learning*. MIT Press, 2018.

[50] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[51] A. Wright, "Current directions in personality science and the potential for advances through computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 292–296, 2014.

[52] C. Judd, L. James-Hawkins, V. Yzerbyt, and Y. Kashima, "Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth." *Journal of Personality and Social Psychology*, vol. 89, no. 6, pp. 899–913, 2005.

[53] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, pp. 289–300, 1995.

[54] R. Ranawana and V. Palade, "Multi-classifier systems: Review and a roadmap for developers," *International Journal of Hybrid Intelligent Systems*, vol. 3, no. 1, pp. 35–61, 2006.

[55] W. Arthur, D. Woehr, and W. Graziano, "Personality testing in employment settings: Problems and issues in the application of typical selection practices," *Personnel Review*, 2001.

[56] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.