# Type Like a Man! Inferring Gender from Keystroke Dynamics in Live-Chats

**Abeer A. N. Buker, Giorgio Roffo, and Alessandro Vinciarelli**
University of Glasgow

**This article analyzes** the interplay between gender and keystroke dynamics, i.e., the way people type as opposed to what they type. In particular, the experiments show that people of different gender tend to type differently when they interact through live-chat interfaces like those available in popular products such as Whatsapp or Skype. The analysis of the results shows that the differences concern mostly the expression of affective states, the way of conveying social presence, and the frequency of planning problems (the difficulties in deciding what to type next). In addition, the experiments show that the differences are sufficiently consistent to allow automatic gender recognition with accuracy higher than 95%. In other words, typing behavior appears to be a reliable gender marker, at least in the case of the interaction through text-based live-chats.

*Digital Object Identifier 10.1109/MIS.2019.2948514*
*Date of current version 22 January 2020.*

One of the main features of social interaction is the pervasive use of nonverbal behavioral cues (facial expressions, gestures, vocalizations, etc.) conveying socially and psychologically relevant information such as attitudes, emotions, personality, gender, etc. Social psychology has extensively investigated the phenomenon in the case of face-to-face interactions in which people have at disposition their natural means of expression (face, body, voice, etc.). However, it is still relatively unclear what happens when communication takes place through technologies (e.g., live-chats and messaging systems) that do not allow the use of nonverbal cues.[1] In other words, it is still relatively unclear whether communication mediated by such technologies involves a nonverbal component and, if yes, whether there are nonverbal cues that convey information about the people that communicate and their interaction.

This article tries to address the problems outlined above, at least to a partial extent, by analyzing dyadic live-chats and, in particular, the interplay between gender and *keystroke dynamics*. Overall, the results show that female and male participants tend to type in a different way, especially when it comes to affective and social aspects of an interaction. Furthermore, the observed differences are sufficiently consistent to allow automatic gender recognition with accuracy above 95%. These results suggest that, at least in the case of the data used in the experiments, communication through live-chats includes a measurable nonverbal component that can convey information about gender, one of the most important social dimensions of an individual.

Besides highlighting the role of nonverbal behavior in technology mediated communication, the results above are important from an application point of view. In fact, live-chats are nowadays one of the main channels through which companies interact with their customers.[2] In such a context, operators deal with individuals they cannot see or hear and, hence, any information available through typing behavior can be of help in maximizing the chances of a successful transaction. In addition, methodologies capable to understand social and psychological phenomena underlying live-chat interactions have been identified as key-technologies for the improvement of important application domains such as tutoring communication systems[3] or e-services.[4]

## PREVIOUS WORK

Until now, research efforts related to keystroke dynamics have focused on *identity verification*,[5] the task of automatically testing whether people actually are who they claim to be (see the work done by Banerjee and Woodard[6] for a survey). The key-assumption underlying the approaches in the area is that the way of typing is as specific of an individual as the many other sources of evidence commonly used for identity verification, including facial appearance, voice, fingerprints, and handwriting style.[7] For such a reason, the identity verification approaches based on keystroke dynamics are similar to those based on other forms of evidence. There is a *world model* $p(\vec{x}|\theta_w)$ that accounts for the distribution of a feature vector $\vec{x}$ across multiple individuals (the features are physical measurements extracted from the data that an individual provides as evidence of her identity) and there are *person models* $p(\vec{x}|\theta_i)$ that account for the distribution of the same feature vector for an individual $i$ in a predefined set of $N$ people ($\theta$ is the parameter set of a model and $i = 1, \dots, N$). When the ratio $p(\vec{x}|\theta_i)/p(\vec{x}|\theta_w)$ goes above a certain threshold, the claim of an individual to be person $i$ is accepted.

While the probabilistic approach is similar for all sources of identity evidence, the features change significantly from one case to the other. For keystroke dynamics, the most common features account for the timing in pressing the different keys that compose the most common $d$-graphs, i.e., the most common sequences of $d$ keys pressed consecutively. The motivation behind such a choice is that identity verification approaches typically involve the entry of a password or the capture of text typed for different reasons. The timing is represented in terms of distance in time between events such as, e.g., pressing and releasing of the same key, pressing of two consecutive keys, etc. (see the work done by Banerjee and Woodard[6] for full details).

In recent years, several works have pointed out that the approaches developed for identity verification can be used, with a few modifications, to classify people according to shared characteristics such as affective state,[5,8–11] age,[12,13] or gender.[13,14] In these cases, the most common approach is to train a different model $p(\vec{x}|\theta_c)$ for every class $c \in \mathcal{C} = \{c_1, \dots, c_L\}$, and then to assign a vector $\vec{x}$ that represents a test sample to the class $\hat{c} = \arg\max_{c \in \mathcal{C}} p(\vec{x}|\theta_c)$. The $d$-graph features originally developed for identity verification have been extensively used for classification. However, new measurements have been developed that better account for the phenomena targeted in the different classification problems.

Major attention has been paid to features that can capture the affective state of an individual.[8] Typing speed, typically measured as number of keys per second or average time between two keys pressed consecutively, has been shown to be particularly effective. The possible explanation is that emotions interplay with cognitive load and, hence, they can lead people to type faster or slower.[11] The main advantage of these features is that they are an implicit expression of affect, i.e., they are the result of processes that take place outside conscious awareness or are difficult to control. In this way, they are more likely to provide *honest* evidence of one's state. More recent works introduce features that measure how frequently explicit expressions of affect are (e.g., exclamation marks and emoticons). However, these correspond to the emotions that people claim to experience and not necessarily to the emotions actually being experienced. In addition, explicit affect expressions are not always frequent (e.g., exclamation marks account for only 0.63% of the total keys pressed in the data used in this article) and can lead to features that often take null value.

In a similar vein, significant efforts have been made to investigate the role of social presence, i.e., the ability to raise awareness of one's presence in others.[15] Overall, there is consensus that social presence is associated to more positive interaction outcomes, whether this means, e.g., tighter integration in a learning community[3] or higher customer loyalty in on-line services.[4] However, to the best of our knowledge, the efforts have focused on the design of chat interfaces capable to convey social presence rather than on the design of features capable to account for it. The last two aspects that have been extensively investigated are planning problems, possible difficulties in deciding what to type next, and style of the text being typed. The former is typically captured through the density of backspace keys (deleting texts means that there have been errors in typing individual keys or entire words)[10] and use of question marks or suspension points (expected to account for epistemic states and uncertainty, respectively). The style of the text is captured through the use of punctuation characters and capital letters (more frequent when the style is more similar to a written text).

## THE STUDY

The experiments have been performed over a corpus of 30 dyadic live-chats that have involved 60 fully unacquainted participants (35 female and 25 male). All of them are native English speakers and British nationals to limit as much as possible the effect of linguistic and cultural differences. The 30 chats revolve around the *Winter Survival Task* (WST), a scenario commonly used in human–human communication experiments. In the WST, the participants are asked to consider 12 items and to identify those that increase the chances of survival in a polar environment. Besides providing a subject of discussion to people that have never met before, the main advantage of the WST is that only a few people have experienced the problem of surviving in a hostile environment (nobody among the participants of the experiments presented in this article). As a consequence, outcomes and dynamics of the interaction do not depend on expertise or knowledge differences, but on social and psychological phenomena taking place during the chats.

The live-chat interface used for the experiments includes a key-logging platform that records and timestamps every key that the participants press. Therefore, the data corresponding to a particular participant can be thought of as a sequence of pairs $(k_i, t_i)$, with $i = 1, \ldots, N$, where $k_i$ is the $i\text{th}$ key that the participant presses ($k_i$ can be any of the keys available on the keyboard), $t_i$ is its timestamp (the time elapsed since January $1\text{st}$, 1970 at the moment the key has been pressed), $N$ is the total number of keys that the participant presses during the chat, and $t_j > t_i$ whenever $j > i$. During the live-chats, the participants type the texts they want to exchange and then they press the "*Enter*" key to make them accessible to their interlocutors (see Figure 1). Such an approach is common in widely diffused products such as, e.g., the chat interface available in *Skype*™ or *Whatsapp*™. For this reason, the sequence of pairs $(k_i, t_i)$ can be segmented into *chunks*, i.e., sequences of characters enclosed between two consecutive Enter keys (see Figure 1). A chunk can finally be further segmented into *tokens*, i.e., into sequences of characters enclosed between two consecutive blanks. Overall, the 30 chats of the corpus include 191,375 keys, including 119,618 pressed by females and 71,757 pressed by males. Correspondingly, the total number of

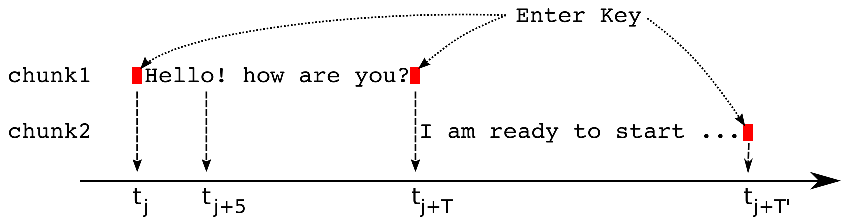

**Figure 1.** Red squares correspond to the Enter keys and delimit the chunks. As soon as the Enter key is pressed, the current chunk becomes visible to the interlocutor and a new chunk starts. Chunk 1 includes the keys of index $j+1$ to $j+T-1$ (included), whereas chunk 2 includes the keys of index $j+T+1$ to $j+T'-1$ (included).

chunks is 3,177, including 2,049 typed by females and 1,128 typed by males.

### Social and Affective Gender Differences

According to the definitions above, a chunk can be thought of as a set $C = \{(k_j, t_j), j = 1, \ldots, T\}$ that includes all the pairs enclosed between two consecutive Enter keys (see Figure 1). Every chunk can be represented as a vector of physical measurements, the *features*, that account for the keystroke dynamics. In the experiments of this article, there are 15 features and focus in particular on affective state and social presence.

Like any chat interface, the one used for the experiments does not allow one to convey information about emotions in terms of natural nonverbal cues (smiles, frowns, etc.). However, it is still possible to extract features that, according to the literature, account for the affective state of a person that is typing. These include *density of exclamation marks* (number exclamation mark divided by $T$), *density of emoticons* (number of emoticons divided by $T$), and *density of uppercase tokens* (number of tokens where all the letters are uppercase divided by the total number of tokens). Such features account for the emotional state an individual claims to experience in explicit terms and can be controlled consciously, i.e., people can decide to use them or not. However, the literature suggests that people convey their affective state implicitly and outside conscious awareness through how fast they type. For this reason, the features include the *typing speed* (average number of keys pressed per second calculated as number of pairs in $C$ divided by $t_T - t_1$) and *median of latency time* between consecutive keys (the median of the differences $t_{j+1} - t_j$ in the set $C$).

To the best of our knowledge, the literature does not suggest features that measure the social presence. However, a person involved in a live-chat makes her chunks visible by pressing "Enter" and this suggests that the frequency of such a key can be an effective measure. In fact, the more often a person presses the Enter key, the less time her interlocutor has to wait before getting an update with respect to the previous text received. As a consequence, the chunks are, on average, shorter and such a tendency can be represented with three features, namely *chunk length* (corresponding to the number $T$ of keys in $C$), *chunk duration* (corresponding to the difference $t_T - t_1$) and *number of tokens* (total number of tokens in $C$).

Hesitations and planning problems while typing can be captured through four features, namely *density of backspaces* (number of times the backspace key is pressed divided by $T$), *backspace time* (total amount of time spent in pressing the backspace key several times consecutively), *density of question marks* (number of question marks divided by $T$), and *density of suspension points* (measured as the average number of characters between consecutive points). Finally, three features account for the tendency to use a formal style, i.e., *density of points* (number of points divided by $T$), *density of capital letters* (number of capital letters divided by $T$), and *density of nonalphabetic characters* (number of nonalphabetic keys not considered in other features divided by $T$).

Table 1 shows an upward pointing arrow whenever a feature tends to be higher, to a statistically significant extent, for a particular gender (according to a two-tailed $t$-test with false discovery rate[16] correction and confidence level 95%). Overall, the table shows that the difference is significant for 7 features out of the total 15, thus suggesting that female and male participants actually tend to type differently. Affect, social presence and planning problems appear to convey information about gender while the style of the text being typed does not. In the case of affect, the significant differences correspond only to the implicit expression of emotions. In particular, female participants tend to type faster (higher number of keys per second and shorter latency time between consecutive keys), a behavioral cue that typically accounts for the valence of the emotional states.[11]

Similarly to the above, female participants tend to project higher social presence by typing shorter chunks that, on average, require less time to be completed before they are made available to the interlocutors. Such an observation seems to confirm previous observations according to which female participants tend to manifest higher awareness of their interlocutors through, e.g., a more frequent use of back-channel[17] (short utterances like "*aha*" that signal attention to someone that is speaking). Finally, male participants seem to correct, more frequently, typos and misspellings (higher density of backspace keys and more time spent in pressing such the backspace key). One possible interpretation of such an observation is that there is an attempt to write like in an exchange of written messages rather than like in a conversation, where a misspelling can be considered acceptable as long as it does not interfere with the understandability of a message.

### Automatic Gender Recognition

So far, the analysis has addressed the features individually, but typing behavior, like any other form of observable behavior, should be represented as a pattern where multiple features jointly contribute to convey information of interest.[18] For this reason, this section presents experiments where vectors having the features described earlier as components are fed to a classifier trained to recognize automatically the gender of the person that has typed a chunk. The experiments make use of random forests (RF) because these, unlike most other classifiers, allow one to deal with unbalanced class distributions by manually setting misclassification costs associated to different classes. In particular, the costs associated to missing the least represented classes are set to be high to attenuate the tendency of the classifier to assign the test samples to the most frequent classes. Such a property is important because the distribution of the genders over the chunks is not uniform, with the chunks typed by female participants accounting for 64.5% of the available material.

**Table 1. Comparison between female and male participants.**

| Dimension | Feature | F | M |
|---|---|---|---|
| Affect | Density of exclamation marks | - | |
| Affect | Density of emoticons | - | - |
| Affect | Density of uppercase tokens | - | - |
| Affect | Typing speed | ↑ | ↓ |
| Affect | Median of latency time | ↓ | ↑ |
| Social presence | Chunk length | ↓ | ↑ |
| Social presence | Chunk duration | ↓ | ↑ |
| Social presence | Number of tokens | - | - |
| Planning problems | Density of backspaces | ↓ | ↑ |
| Planning problems | Backspace time | ↓ | ↑ |
| Planning problems | Density of question marks | - | - |
| Planning problems | Density of suspension points | ↓ | ↑ |
| Style | Density of points | - | - |
| Style | Density of capital letters | - | - |
| Style | Density of nonalphabetic characters | - | - |

According to a two-tailed $t$-test with false discovery rate correction, the difference between the feature values for female (F) and male (M) participants is statistically significant in 7 cases out of 15. The arrow points upwards for the gender showing the highest average value of the feature.

The training has been performed according to a *leave one participant out* approach, meaning that the training set includes the chunks typed by all participants except one and the test set includes only the chunks typed by the left out participant. The training process has been repeated as many times as there are participants and, at each iteration, a different participant has been left out. This makes the experiments *person-independent*, i.e., it ensures that the same individual is never represented in both training and test set so that the approach recognizes the gender and not a specific individual (like it happens, e.g., in identity verification). The RFs have been trained to recognize the gender of the person that has typed a particular chunk and the

**Table 2. Gender recognition performance.**

| Level | $\hat{\alpha}$ | Accuracy | $F1$ Score |
|---|---|---|---|
| Chunk | 54.5% | 84.6% | 88.4% |
| Participant | 51.4% | 98.3% | 97.5% |

The table includes accuracy $\hat{\alpha}$ (percentage of times the classifier makes the right decision) of a baseline approach randomly classifying the test samples according to the *a priori*distribution of the classes, accuracy of the approach, and $F1$ score (harmonic mean of precision and recall) at both chunk and participant level.

first row of Table 2 shows both accuracy and $F1$-score achieved for such a task. In addition, the table provides the accuracy $\hat{\alpha} = p_m^2 + p_f^2$ of a random classifier that assigns a test sample to the female or male class with probability $p_f$ and $p_m$, respectively ($p_f$ and $p_m$ are the *a priori* probabilities of female and male class). The accuracy of the proposed approach is higher, to a statistically significant extent, than $\hat{\alpha}$. Therefore, it is possible to say that the proposed approach performs better than chance.

The results obtained at the chunk-level confirm that the typing behavior conveys information about gender. However, it is unclear whether the performance of the approach is similar for all participants or whether it tends to be higher for some of them. For this reason, Figure 2 shows the accuracy achieved over the chunks of each person involved in the experiments and, in particular, it shows that the accuracy is above 50% in all cases except one. Therefore, the application of a *majority vote*, meaning that the participants are assigned to the class their chunks are most frequently assigned to, leads to an accuracy of 98.3%, better than the corresponding $\hat{\alpha}$ to a statistically significant extent (see Table 2). Such a result suggests that the features described earlier can act as reliable gender markers.

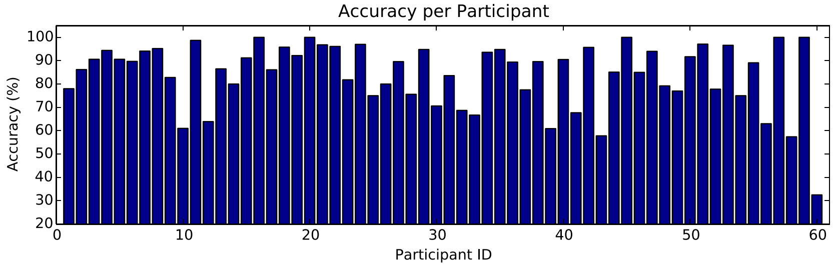

**Figure 2.** Chart shows the accuracy achieved over the chunks of every participant.

## DISCUSSION AND CONCLUSIONS

Nonverbal communication is often referred to as *body language*, an expression that accounts for the major role that the body plays in interaction, especially when it comes to conveying socially and psychologically relevant information. Such a role is the result of a long evolutionary process that has shaped the brain to be sensitive to the signals sent by co-located others more than to any other signal in the environment (e.g., the human voice is one of the sounds that requires the lowest energy to be heard). Still, despite such an evolutionary history, people communicate increasingly more frequently through technologies that prevent, partially or totally, the use of nonverbal behavior. For example, phones allow one to use nonverbal vocal behavior (laughter, sobbing, intonation, pauses, etc.), but not facial expressions or gestures.

In the context outlined above, it is important to investigate whether body language is still possible when the body cannot play its role. For this reason, this article has shown that there is a significant interplay between gender and keystroke dynamics at least in the case of interactions taking place through live-chat interfaces. In particular, the experiments have shown that it is possible to infer the gender of a person from her typing behavior with an accuracy higher than 95%. In addition, the experiments have shown that such a performance relies mostly on features (physical and machine detectable measures extracted through a key-logging platform) that account for implicit and explicit expression of affect, social presence, and planning problems.

According to a recent survey, 36% of adults owning a smartphone use messaging systems (https://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/). In addition, The market for technologies supporting live-chats is expected to grow with an average rate of 7.3% until 2023 when it is expected to reach a total volume close to one billion dollars per year (https://www.alliedmarketresearch.com/press-release/live-chat-software-market.html). In this respect, the extension of domains like affective computing, originally limited to observable behavior in face-to-face interactions, to keystroke dynamics promises not only to provide new insights about the way people manifest their inner states through observable behavior, but also to have societal and economic impact.

## ACKNOWLEDGMENTS

This work was supported by the UK EPSRC under Grant EP/M025055/1 and Grant EP/N035305/1.

## REFERENCES

1. A. Vinciarelli and A. S. Pentland, “New social signals in a new interaction world: The next frontier for social signal processing,” *IEEE Syst., Man, Cybern. Mag.*, vol. 1, no. 2, pp. 10–17, Apr. 2015.
2. D. Clarkson, C. Johnson, E. Stark, and B. McGowan, “Making proactive chat work: Maximizing sales and service requires ongoing refinement,” Forrester Research, Cambridge, MA, USA, Tech. Rep., 2010.
3. T. Traphagan *et al.*, “Cognitive, social and teaching presence in a virtual world and a text chat,” *Comput. Educ.*, vol. 55, no. 3, pp. 923–936, 2010.
4. D. Cyr, K. Hassanein, M. Head, and A. Ivanov, “The role of social presence in establishing loyalty in e-service environments,” *Interacting Comput.*, vol. 19, no. 1, pp. 43–56, 2007.
5. C. Epp, M. Lippold, and R. Mandryk, “Identifying emotional states using keystroke dynamics,” in *Proc. Assoc. Comput. Mach. Human-Comput. Interact.*, 2011, pp. 715–724.
6. S. Banerjee and D. Woodard, “Biometric authentication and identification using keystroke dynamics: A survey,” *J. Pattern Recognit. Res.*, vol. 7, no. 1, pp. 116–139, 2012.
7. K. Delac and M. Grgic, “A survey of biometric recognition methods,” in *Proc. IEEE Int. Symp. Electron. Mar.*, 2004, pp. 184–193.
8. A. Kołakowska, “A review of emotion recognition methods based on keystroke dynamics and mouse movements,” in *Proc. IEEE Int. Conf. Human Syst. Interact.*, 2013, pp. 548–555.
9. A. Kołakowska, “Recognizing emotions on the basis of keystroke dynamics,” in *Proc. IEEE Int. Conf. Human Syst. Interact.*, 2015, pp. 291–297.
10. P. Shukla and R. Solanki, “Web based keystroke dynamics application for identifying emotional state,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 11, pp. 4489–4493, 2013.
11. M. Trojahn, F. Arndt, M. Weinmann, and F. Ortmeier, “Emotion recognition through keystroke dynamics on touchscreen keyboards.” in *Proc. Int. Conf. Enterprise Inf. Syst.*, 2013, pp. 31–37.
12. D. Brizan, A. Goodkind, P. Koch, K. Balagani, V. Phoha, and A. Rosenberg, “Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics,” *Int. J. Human-Comput. Studies*, vol. 82, no. 1, pp. 57–68, 2015.
13. A. Pentel, “Predicting age and gender by keystroke dynamics and mouse patterns,” in *Proc. ACM Conf. User Model., Adaptation Personalization*, 2017, pp. 381–385.
14. I. Tsimperidis, A. Arampatzis, and A. Karakos, “Keystroke dynamics features for gender recognition,” *Digit. Investigation*, vol. 24, no. 1, pp. 4–10, 2018.
15. 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,” *Comput. Human Behav.*, vol. 27, no. 1, pp. 513–521, 2011.
16. Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *J. Roy. Statist. Soc. Series B*, vol. 57, pp. 289–300, 1995.
17. A. Vinciarelli, P. Chatziioannou, and A. Esposito, “When the words are not everything: The use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls,” *Frontiers ICT*, vol. 2, p. 4, 2015.
18. J. Kagan, *Five Constraints on Predicting Behavior.* Cambridge, MA, USA: MIT Press, 2017.

**Abeer A. N. Buker** is currently a Research Student with the University of Glasgow, Glasgow, U.K. Her research interest is the analysis of keystroke dynamics. Contact her at 2171949B@student.gla.ac.uk.

**Giorgio Roffo** is currently a Research Associate with the University of Glasgow, Glasgow, U.K. His research interests are machine learning and computer vision. He received the Ph.D. degree in computer science with the University of Verona, Verona, Italy. Contact him at Giorgio.Roffo@glasgow.ac.uk.

**Alessandro Vinciarelli** is currently a Full Professor with the University of Glasgow, Glasgow, U.K. His research interest is the analysis of nonverbal behavior in social interactions. He received the Ph.D. degree in applied mathematics from the Idiap Research Institute, Switzerland. Contact him at Alessandro.Vinciarelli@glasgow.ac.uk.