eng

MDPI

Article

Finding the Age and Education Level of Bulgarian-Speaking Internet Users Using Keystroke Dynamics

Denitsa Grunova and Ioannis Tsimperidis *

MLV Research Group, Department of Computer Science, International Hellenic University, 65404 Kavala, Greece; ntgkrou@teiemt.gr

* Correspondence: tsimperidis@cs.ihu.gr; Tel.: +30-251-046-2147

Abstract: The rapid development of information and communication technologies and the widespread use of the Internet has made it imperative to implement advanced user authentication methods based on the analysis of behavioural biometric data. In contrast to traditional authentication techniques, such as the simple use of passwords, these new methods face the challenge of authenticating users at more complex levels, even after the initial verification. This is particularly important as it helps to address risks such as the possibility of forgery and the disclosure of personal information to unauthorised individuals. In this study, the use of keystroke dynamics has been chosen as a biometric, which is the way a user uses the keyboard. Specifically, a number of Bulgarian-speaking users have been recorded during their daily keyboard use, and then a system has been implemented which, with the help of machine learning models, recognises certain acquired or intrinsic characteristics in order to reveal part of their identity. The results show that users can be categorised using keystroke dynamics, in terms of the age group they belong to and in terms of their educational level, with high accuracy rates, which is a strong indication for the creation of applications to enhance user security and facilitate their use of Internet services.

Keywords: keystroke dynamics; user classification; machine learning; Bulgarian language

Citation: Grunova, D.; Tsimperidis, I. Finding the Age and Education Level of Bulgarian-Speaking Internet Users Using Keystroke Dynamics. *Eng* **2023**, *4*, 2711–2721. https://doi.org/10.3390/eng4040154

Academic Editor: Goncalo Jesus

Received: 3 September 2023
Revised: 18 October 2023
Accepted: 23 October 2023
Published: 25 October 2023

1. Introduction

Communication technologies have brought about many different changes in the way the average person lives. As the Internet becomes an integral part of everyday life of more and more people, the need to accurately identify the demographic characteristics of Internet users has become paramount, for several reasons. The reasons for this are varied and related to user security and the best use of Internet services. Profiling unknown users by identifying certain inherent or acquired characteristics, such as their age and educational level, is essential for various applications, including personalised content delivery, targeted advertising, and customisation of the user experience. In this context, the use of keystroke dynamics as a means of extracting valuable demographic information has garnered considerable attention. Keystroke dynamics, a branch of behavioural biometrics, focuses on analysing the unique typing patterns exhibited by individuals [1]. These rhythms and patterns are idiosyncratic [2], in the same way as an individual's handwriting or signature, due to the similar underlying neurophysiological mechanisms. By studying different typing patterns and their correlations with demographic characteristics, keystroke dynamics provides a novel approach to demographic profiling.

Bulgarian is a Slavic language spoken by about 9 million people, mostly in Bulgaria and other neighbouring countries. It is the official language of Bulgaria and has historical importance in the Balkan region. Bulgarian has unique linguistic features, such as the Cyrillic alphabet used for writing [3]. However, the application of keystroke dynamics in the Bulgarian linguistic context remains relatively unexplored. Bulgarian, as a Slavic language, has distinct linguistic features that may influence typing behaviour. Investigating

the feasibility and effectiveness of keystroke dynamics within the Bulgarian-speaking population is crucial for the development of accurate and reliable demographic profiling techniques tailored specifically for this language group, which numbers approximately 400 million people mostly in Eastern Europe and Northern Asia.

Traditional methods of demographic profile creation, such as surveys and questionnaires, often suffer from limitations such as subjectivity, response bias, and reliance on user self-reporting. Also, some more modern methods, such as recognizing characteristics from facial photographs or studying the text the user has written, require the existence of specific multimedia files or access to personal data. In contrast, keystroke dynamics leverages data derived from how users type and not what they type. This means that no access to their personal information is required. Also, for data recording, it is not required sophisticated hardware, as a simple physical or virtual keyboard is enough. Furthermore, the method is non-intrusive, and it is possible to record data continuously without interfering with the ongoing work of the users.

Creating the profile of an unknown user has several applications. First, unsuspecting users may be alerted to the possibility of falling victim to malicious users hiding their characteristics. Second, sites, forums, products, and services that match a user's profile can be recommended, saving them valuable time from searching the Internet. Third, authentication can be enhanced, as the identification will be enriched with additional information. Fourth, regarding the recording of data from the Bulgarian language, it facilitates a more comprehensive understanding of the unique typing behaviours exhibited by Bulgarian speakers, taking into account the specific linguistic characteristics of the Bulgarian language.

This paper endeavours to identify specific characteristics of unidentified Internet users by analysing their typing patterns. The primary objectives are twofold: firstly, to fortify security measures, safeguarding genuine users from potentially fraudulent activities; and secondly, to optimize the utilization of Internet services. This way, this work aims to enhance user authentication, to ensure the protection of unsuspecting individuals, and simultaneously to maximize the efficiency and user experience across various online services.

The rest of the paper follows the following structure. First, a summary of the existing literature relevant to the topic under consideration is provided. Then, the methodology followed is described and its details are discussed. In Section 4, the results for each of the four machine learning models used are presented. Finally, the paper concludes by presenting possible future directions for this research.

2. Background

The idea of keystroke dynamics dates back to the late 1800s. In fact, it came from a long-held belief that Morse code senders could identify each other by speed and rate of transmission. In addition, telegraphers identified each other through what they called the "sender's punch". The U.S. National Science Foundation, or NSF, conducted research in the 1980s that determined that each person has his or her own keyboard writing style. This is achieved through the NSF's keystroke recognition method, which analyses and processes the way a person writes on their keyboard [4].

As early as the mid-1970s, the examination of how the way one uses the keyboard can be a recognisable hallmark began. This was first highlighted in Spillane's research [5], where the idea of identifying users by the way they type was introduced. Also, an important contribution was made through the publication of the study by Forsen et al. [6], where keystroke dynamics was analysed as one of the biometric characteristics that can be used to verify the identity of a user requesting access to a system. One of the first studies on this topic was conducted by Gaines et al. [7]. They had a group of seven secretaries write the same three paragraphs twice over a period of four months. A total of 300 to 400 words were required both during the writing phase and for each comparison. Time delays between successive typing were measured, and the analysis was based on a limited number of

digraphs (two consecutive letters). Although the results were very encouraging (FAR 0% and FRR 4%), the sample size was too small and the volume of data required was too large.

Another study was conducted by Umphres and Williams in 1985 [8]. In this work, the time delay between consecutive key presses was also used to authenticate the user. It took approximately 1400 key presses to generate a profile for each user. Each time authentication was required, another 300 characters were required. The FAR achieved was 6%, but it is clear that the volume of data required was particularly large. Also, a similar study was conducted by Leggett and Williams [9] with data obtained from 17 computer programmers. The system developed showed an FAR of 5% and an FRR of 5.5%. However, a major drawback of this method is the need for a large amount of data. In total, each programmer had to write over 1000 words.

Canales et al. [10] attempted to create an authentication system for students who are examined online. For this reason, they used data coming from keystroke dynamics and stylometry. Specifically, known keystroke dynamics features and 82 stylometric features were extracted, which were character-based, word-based, and syntactic. Data derived from the recording of 40 students were collected, and a K-NN was used as a classifier. FAR and FRR were chosen as metrics, and experimental results showed that authentication was more successful when only keystroke dynamics features were used.

In their study, Zhong et al. [11] contributed a new distance metric in the research field of user authentication through keystroke dynamics, which was a combination of Manhattan distance and Mahalanobis distance, attempting to exploit the advantages of the two metrics and eliminate their disadvantages. To test the performance of their system they used the CMU keystroke dynamics benchmark dataset, and showed that the new metric they proposed outperformed other distance metrics.

In another study by Monrose and Rubin [12], the aim of the task was user recognition, and therefore, each volunteer was asked to enter a specific sentence as well as a sentence of their choice. The success rates were not so satisfactory when the texts were unfamiliar, but when the text copied by the user was specified, they reached a success rate of 90.7%.

Ayotte et al. [13] attempted to address the problem of requiring a lot of data to achieve a high success rate for user authentication through keystroke dynamics. For this purpose, they introduced the snapshot-based tail area density metric (ITAD), a new graph comparison algorithm, to significantly reduce the number of keystrokes required for user authentication. The classifier they used was random forest, and in addition to the very good results they achieved, they showed that the most commonly used keystroke dynamics features, namely keystroke durations and digram latencies, are the most effective.

In a different field, Acien et al. [14] presented a comprehensive exploration of long short-term memory (LSTM) networks for keystroke biometric authentication on a large scale in free-text scenarios. Their research assessed the performance of LSTMs trained with a moderate number of keystrokes per user. They considered various machine learning models, training sample sizes, keystroke sequence lengths, and databases based on different device types, such as physical and virtual keyboards. Their methodology achieved an EER of 2.2% and 9.2% for physical and virtual keyboards, respectively. In fact, they showed that it can also be used in an authentication system involving many users, since the error rates increased only slightly even when there was data from 100,000 people.

Sahu et al. [15] dealt with the problem encountered in some systems where multiple users are involved and one user connects to another user's account. To solve this problem, they resorted to keystroke dynamics, and the algorithm they proposed involved techniques of data preprocessing, dimensionality reduction, data clustering, data embedding, and data localization and could be used directly on the typing data. To test the performance of their algorithm they used two available datasets. Ultimately, the rates of correct user identification they achieved were high.

In another study on user age search, Tsimperidis et al. [16] exploited a dataset containing 387 logs and extracted 700 keystroke dynamics features from them. The features extracted included keystroke durations and digram latencies. Using five different classifiers,

experiments were conducted with different feature sets. The results of the experiments led to the development of a system that could identify the age group of an unknown user with an accuracy of about 90%, among four different options.

Buriro et al. [17] tried to investigate the possibility of estimating, among other things, the age of a user who types a PIN/password between 4 and 16 digits in length, on mobile devices. Their data were collected from 150 volunteers on a specific device, and three classes were defined. They extracted temporal keystroke dynamics features and used several classifiers. Finally, the best results came from random forest, which had an accuracy of 87.9%.

In another study, Ulinskas et al. [18] utilised an existing keystroke dynamics dataset derived from a recording of 53 individuals typing the same password. The purpose of the study was to identify user fatigue. From these data, features relating to keystroke durations and digram latencies were extracted. Using six different classifiers, it was observed that the best results came from the latencies in the "up-up" graph, managing to identify fatigue with 91% accuracy.

In a different field, Pentel [19] collected keystroke data from various Web applications from 2011 to 2018. The log data were linked to the age and gender of the users and in some cases to other available information. A total of 2.3 million keystrokes from 7119 data logs were analysed, which came from approximately 1000 individuals and covered six different age groups. Binary and multi-class classification was applied using supervised machine learning methods. The results of the binary classification showed that performance was at the general baseline level, with the best F-score exceeding 0.92 and the lowest being 0.82. Through discriminative feature analysis, it was discovered that there was some overlap with features extracted from previous text mining studies.

In their effort, Yan and Yan [20] conducted a study in which a methodology was developed to categorize blog writers according to their gender. To achieve this, they utilised a collection of 75,000 blog entries and used various word features such as frequency of occurrence, blog background colour, font type and style, as well as other features such as punctuation marks and emoticons. Their methodology achieved an F-measure equal to 0.68.

In another study, Jones et al. [21] conducted a study where they collected data from user profiles and search keywords from Yahoo.com. They then created a model using a classifier based on SVM. This model was able to achieve very satisfactory accuracy in classifying the gender of users, with a rate of 83.8%. It was also able to predict the age of the users with an accuracy of 63.9%.

Tsimperidis et al. [22] proposed a method to identify some characteristics of an unknown user through keystroke dynamics. They collected data from 110 volunteers during their daily device usage, and then they trained five machine learning models using selected features and tried it to recognize the age group, handwriting, and education level of unknown users. The experimental results showed that this method can recognize the age group with 87.6% accuracy, handwriting with 97.0% accuracy, and education level with 84.3% accuracy for an unknown user.

In a different field, Pentel [23] focused on the analysis of unintended user activities in human–computer interactions. While user interfaces are usually designed to react only to intentional commands, users often perform unintentional activities that produce many cues for the user and can be used to plan the appropriate response by the system. Specifically, the goal of this research was to predict the age and gender of users through the analysis of data generated from mouse and keyboard devices. These data were collected from six different systems from 2011 to 2017 and include information from 1519 individuals. The machine learning models were able to predict both the age and gender of the user with very high accuracy. In particular, the F-score and accuracy metrics were above 0.9.

In their study, Cascone et al. [24] used keystroke dynamics on touch devices to classify demographic information, such as the user's age and gender. The authors sought to investigate whether the process of touch typing, which includes information about the

pressure applied to the keys, can be used to detect user demographic information. To achieve this, the researchers analysed the data collected during the touch typing process using various machine learning algorithms. Among the findings of the research, it emerged that younger people tend to type faster but with more errors, while older people tend to type slower but with fewer errors. This may suggest possible correlations between keystroke dynamics and user age characteristics.

In the study conducted by Raul et al. [25] examined the use of keystroke dynamics as a biometric authentication method. The researchers analysed the data collected to evaluate the effectiveness of various authentication methods based on keystroke dynamics. They also studied various algorithms based on statistics and machine learning to analyse their positive and negative points. From the research, it was found that there is a need to extend the keystroke dynamics dataset to include all the key features.

One of the most prominent issues in classification studies is user classification based on their age, probably because age is one of the personal characteristics that people choose not to declare or often misrepresent in order to avoid being noticed in case they commit malicious actions.

The study of Schler et al. [26] was the search for the age of the author of a blog. The researchers collected their data from 71,493 blogs, which they classified according to the age of the author. For several of them, no age information was available, and for some of the classes they created, there were not enough data, resulting in three classes: the 10s (age group 13–17), the 20s (age group 23–27), and what they called the 30s (age group 33–46). As features for the classification, they used the frequency of occurrence of some words. The multi-class real Winnow algorithm was used for classification, in which for each class, a vector of as many dimensions as the set of parameters chosen was defined. The final results proved that the age group of blog creators could be correctly predicted with 73% accuracy.

The study by Rao et al. [27] aims to identify the characteristics of Twitter users, especially their age group, gender, region origin, and political orientation. They proposed an approach to automatically discover a number of user attributes by examining their status messages, the social network structure, and the communication behaviour of the users. SVM was chosen as the classifier, and users were divided into people over 30 and under 30. The researchers tested the system and attained a classification accuracy rate of about 74%.

Keystroke dynamics can be used to identify the under-18 age group, thus offering an effective way to create a model to protect children from online threats. By implementing a limited firewall, an environment that is more suitable for this particular user group will be created [28]. It can also be exploited in e-commerce problems by creating product recommendation services that are tailored to the age and gender of the users. Furthermore, the ability to identify the age and user through keystroke dynamics can allow the creation of a system where content or advertisements can be presented efficiently and targeted to the appropriate consumers, taking into account their preferences and characteristics [29].

Educational systems vary greatly between countries. International data on education should therefore be based on a classification that proposes, for all countries of the world, correct criteria for the distribution of educational programs at levels that can be considered comparable.

The educational level of an individual is an important characteristic in various surveys that have been carried out over the years.

While fixed-text keystroke dynamics biometrics are often used during the login process to provide an authentication, free-text biometric keystroke systems allow continuous authentication of a user during the entire session for increased security [30]. Furthermore, other studies [31,32] have exploited these additional user characteristics, such as age and gender, to improve the performance of the user authentication model.

Lin et al. [33] presented a proposal for an authentication system based on the analysis of the keystrokes dynamics features. This includes recording the duration required to press a key (known as keystroke duration) and the time between the release of a key and

the pressing of the next one (known as the up-down diagram). The purpose is to detect unauthorised users, even when they have knowledge of the genuine password for an account. After collecting the data and extracting the necessary features, a convolutional neural network is applied, which achieves 99% accuracy in detecting legitimate users.

3. Methodology

Datasets on keystroke dynamics are difficult to find on the Internet. In fact, most of them come from recording users typing fixed text. In contrary, the publication of free-text logging data carries the risk of leaking personal data and is therefore rarely found on the Internet by studies or surveys. The methodology of this research consists of three consecutive phases. In the first phase, free-text data were collected from Bulgarian-speaking volunteers who agreed to participate in the process. In the second phase, appropriate keystroke dynamics features were extracted. Finally, in the third phase, machine learning algorithms, namely naïve Bayes, SVM, multilayer perceptron, and random forest, were used to classify users according to their age and educational level.

3.1. Data Collection

For the needs of the research, keylogger software was developed and installed on the volunteers' computers. To ensure that sensitive and personal data of the volunteers, such as passwords or credit card numbers, would not be leaked, the researchers committed themselves by signing a consent form not to disclose the data to third parties and to the exclusive use of these data in this research. Furthermore, volunteers were given the option to activate the keylogger only when they wished to do so, in order to choose which data would be recorded.

For the needs of data collection, hundreds of individuals were approached, and ultimately, several dozen participated, generating a number of logfiles, each of which contained data from the use of approximately 3500 keystrokes. Each participant could type at any moment of the day and at any application. This was done to capture as much of the participants' daily typing as possible. That is, no specific time frame was imposed, adding versatility and reliability to the dataset, and therefore, no specific keylogging sessions were defined.

Each keystroke action performed by the volunteers was recorded in the logs, which were in the following format:

82,#2022-04-14#,1741707,"dn"
82,#2022-04-14#,1741879,"up"
86,#2022-04-14#,1742754,"dn"
86,#2022-04-14#,1742817,"up"
69,#2022-04-14#,1742840,"dn"
69,#2022-04-14#,1742950,"up"

Each line represents a record of the volunteer's action. The first field represents the virtual key code of the key that the volunteer pressed or released. The second field, enclosed by the symbol "#", indicates the date on which the action took place. The third field corresponds to the exact time when the action took place, expressed as an integer number representing milliseconds from the beginning of the day. The fourth field describes the type of action, with the word "dn" indicating the pressing of a key and "up" referring to the release of a key.

The recording of the volunteers whose native language is Bulgarian was carried out between 29 March 2022 and 16 May 2022. During this period 46 logfiles were collected. Table 1 shows the demographic data of the Bulgarian-speaking volunteers studied in this research.

Table 1. Logfiles per age and educational level.

Characteristic	Class	Logfiles
Age	18–25	9
	26–35	6
	36–45	6
	46+	25
Educational Level	ISCED 2	0
	ISCED 3	17
	ISCED 5	1
	ISCED 6	12
	ISCED 7–8	16

Admittedly, the dataset created is relatively small. But since there is no dataset on the Internet with data from Bulgarian-speaking users, at least according to what is known, which is the target group of this research, combined with the difficulty of creating such a dataset, due to the distrust of individuals to participate in a process where their typing is recorded, the particular dataset is considered the best that can be used.

Looking at the logfiles, which were obtained from the recording of Bulgarian-speaking volunteers, it seemed that an attempt could be made to study the data and classify the users based on their age. It was important to test this characteristic to find a way to group the recorded ages of the users. An option was made to create two classes, those up to 45 years old and those over 45 years old. The first class consists of 21 logfiles, while the second class consists of 25.

Analysing the data of Bulgarian-speaking users, the choice of classify users by their educational level is an interesting and feasible classification. In their personal data, when registering for the project, users indicated their educational level according to the ISCED scale. The ISCED (International Standard Classification of Education) classification was developed by UNESCO in the mid-1970s and was first revised in 1997. A further revision of ISCED took place between 2009 and 2011 with extensive global consultations with countries, regional experts, and international organisations. Finally, ISCED 2011 was adopted by the UNESCO General Conference in November 2011 [34]. Users who have declared ISCED2, ISCED3, and ISCED5 are considered non-university level, while users with ISCED6 and ISCED7–8 have university education and above.

3.2. Feature Extraction

Keystroke dynamics comes with many different features, which can be categorised into two major categories: temporal and non-temporal. In particular, temporal features are the most common category and include features such as keystroke durations and digram latencies. In this research, keystroke durations were chosen to be examined. The software implemented for feature extraction was implemented using the Python programming language, and its purpose is to calculate the average duration of keystrokes by each user. After its execution, files of appropriate format are generated to be readable by the WEKA 3.8.6 software.

3.3. Classifier Selection and Model Evaluation

After a thorough evaluation, the four models that emerged with excellent accuracy and low time complexity were naïve Bayes, SVM (support vector machine), MLP (multilayer perceptron), and random forest. The model validation stage seeks to assure the correct operation and application of the models. There are a variety of techniques that can be used to verify the reliability of a model, and several of these were applied to validate the four models.

In the context of evaluating machine learning algorithms, there are several metrics used to compare results between different approaches. One of the most common metrics is accuracy, which refers to the percentage of correctly classified instances in relation to the

total number of instances. In addition to this, there are other metrics that offer additional interest in the evaluation of algorithms.

In particular, the time to build model is critical, as it reflects the time required to train the model. In addition, the F-measure (F1), which is the harmonic mean between precision and recall [35] and is a safe measure even for unbalanced datasets, and the area under the ROC curve (AUC), which is the area under the receiver operating characteristic curve [36], were used. These additional metrics deepen the evaluation of models and help in the proper selection of machine learning methods.

4. Results

For each of the four models (naïve Bayes, SVM, MLP, and random forest), a large number of experiments were conducted to find the values of the classifiers' parameters that lead to the best performance of the system. The first priority evaluation was based on the performance with the highest accuracy, while the criterion with the lowest time complexity (TBM—time to build model) was then taken into account. This was followed by the criterion with the highest area under the ROC curve (AUC) and the criterion with the highest F-score (F1).

The experimental results for finding the age group of the users are summarised in Table 2.

Table 2. The best results of the four models, in terms of accuracy, time to build model (TBM), F1 score, and area under the ROC curve, according to age group classification.

Model	TBM (in s)	Accuracy	F1-Score	AUC
Naïve Bayes	<0.01	82.61%	0.826	0.858
SVM	<0.01	91.30%	0.913	0.909
Multilayer perceptron	2.89	80.43%	0.804	0.813
Random forest	0.01	93.48%	0.935	0.987

The results in Table 2 show that random forest performs better than all other models in finding the age group of users, and outperforms in terms of accuracy, F1, and AUC. As the second ranked model in performance is judged to be SVM, which shows the second best accuracy, F1, and AUC, but performs faster than random forest. In some additional observations, naïve Bayes is the fastest, while multilayer perceptron seems to have the worst performance among the four models in user classification in terms of finding the age group of users.

The experimental results for finding the educational level of users are summarised in Table 3.

Table 3. The best results of the four models, in terms of accuracy, time to build model (TBM), F1 score, and area under the ROC curve, according to educational level classification.

Model	TBM (in s)	Accuracy	F1-Score	AUC
Naïve Bayes	<0.01	60.87%	0.613	0.803
SVM	<0.01	73.91%	0.736	0.716
Multilayer perceptron	2.31	86.96%	0.871	0.841
Random forest	0.07	86.96%	0.868	0.875

From the results of Table 3, it can be seen that the two models multilayer perceptron and random forest have the same accuracy, i.e., 86.96%. However, the multilayer perceptron took about 2.31 s to train, while the random forest only took 0.07 s. Also, between these two models, random forest outperforms in AUC, while multilayer perceptron outperforms in terms of F1-Score. Naïve Bayes and SVM models show lower accuracy but are faster. Overall, the classification of Bulgarian-speaking volunteers based on their educational level showed lower accuracy rates than the classification by age.

Evaluating the results of Tables 2 and 3, it becomes clear that the successful prediction of the age group and educational level of Bulgarian-speaking users can be achieved at a rate of more than 90% and approximately 87%, respectively, by using more than one machine learning model. This is an indication that the high accuracy achieved is not just an outlier presented by a classifier. Also, it is observed that this high accuracy is achieved in a very short time to build model, which shows that a system that would aim to find user characteristics using keystroke dynamics would produce results in a very short time.

Results Comparing

As it became evident from the "Background" section there are other studies that have dealt with finding the age group a user belongs to. Although each study used a different methodology, collecting data in a different way, dividing the users into a different number of groups, and using different classifiers, a comparison of results that can be made between them is shown in Table 4.

Table 4. Comparison of results between keystroke dynamics studies dealing with users' age group.

Work	Logfiles	Age Groups	Accuracy
Tsimperidis et al. [16]	387	4	90.0%
Buriro et al. [17]	1500	6	87.7%
Schler et al. [26]	Unknown	3	76.2%
Rao et al. [27]	1000	2	74.0%
This study	46	2	93.5%

Although the results cannot be directly compared between the studies presented in Table 4, due to the fact that, for example, only the present work and the work of Tsimperidis et al. [16] collected data with a free-text method, it seems that the highest accuracy is achieved by this study.

Regarding the classification of users according to their educational level, only one relevant study is found, that of Tsimperidis et al. [37], in which users are divided into five classes and the accuracy reaches 86.8%, which is similar to this study.

5. Conclusions

It is part of everyday life for people to communicate over the Internet, and usually via text messaging. One of the major threats with this way of communication is those users who hide their personal characteristics, such as age and gender, and aim to deceive unsuspecting users. Due to the nature of online communication, hiding such information is an easy task. In order to protect unsuspecting users, various methods have been proposed to reveal some of the characteristics of anonymous users. To solve this problem, in this paper, a method based on keystroke dynamics is proposed.

For the purpose of this research, 46 logfiles containing data from keyboard usage were collected. After extracting the appropriate keystroke dynamics features, the performance of four different classifiers, naïve Bayes, support vector machine, multilayer perceptron, and random forest were examined. The experimental results proved that it is possible to create fairly reliable systems that can identify the age and educational level of an unknown Internet user with an accuracy of 93.47% and 86.95%, respectively.

The ability to identify the age and educational level of a typing user is valuable in many areas, such as digital forensics, targeted advertising, and behavioural biometrics for user profiling. However, it is important to note that the development of such a system must comply with the current legal framework, as the analysis of unauthorised keystrokes may violate privacy, with potential implications for human security.

The contributions of the paper are therefore two-fold: first, the creation of a free-text keystroke dynamics dataset, which is not often found on the Internet, due to the risk of leaking volunteers' personal data, and second, the novelty of using keystroke dynamics to detect certain features of unknown users.

In future research, it could be possible to record keystrokes from more users and to collect keystroking data from more devices, such as tablets and smartphones. In this way, the dataset will be extended resulting in any findings that are extracted being more valid. This also would allow more keystroke dynamics features to be examined and more classifiers to be tested. With these data from a future study, a complete profile of the user will be created and will be easily identified each time a user tries to enter a system or unauthorised access is attempted and rejected.

Author Contributions: Conceptualization, D.G. and I.T.; methodology, I.T.; software, D.G. and I.T.; validation, D.G. and I.T.; formal analysis, D.G.; investigation, I.T.; resources, D.G. and I.T.; data curation, D.G.; writing—original draft preparation, D.G; writing—review and editing, I.T.; visualization, D.G.; supervision, I.T.; project administration, I.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The logfiles contain sensitive and/or personal data of the volunteers who participated in the typing recording and are therefore not available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, L.; Li, Z.; Shen, C. Performance Evaluation of an Anomaly-Detection Algorithm for Keystroke-Typing Based Insider Detection. *Tsinghua Sci. Technol.* **2018**, *23*, 513–525. [CrossRef]
2. Douhou, S.; Magnus, J.R. The reliability of user authentication through keystroke dynamics. *Stat. Neerl.* **2009**, *63*, 432–449. [CrossRef]
3. Videnov, M. The present-day Bulgarian language Situation: Trends and prospects. *Int. J. Sociol. Lang.* **1999**, *1999*, 11–36. [CrossRef]
4. Monrose, F.; Rubin, A.D. Keystroke Dynamics as a Biometric for Authentication. *Future Gener. Comput. Syst.* **2000**, *16*, 351–359. [CrossRef]
5. Spillane, R. Keyboard apparatus for personal identification. *IBM Tech. Discl. Bull.* **1975**, *17*, 3346.
6. Forsen, G.E.; Nelson, M.R.; Staron, R.J. *Personal Attributes Authentication Techniques*; Pattern Analysis and Recognition Corporation: Dallas, TX, USA, 1977; p. 0331.
7. Gaines, R.S.; Lisowski, W.; Press, S.J.; Shapiro, N. Authentication by Keystroke Timing: Some Preliminary Results. 1980. Available online: https://apps.dtic.mil/sti/pdfs/ADA484022.pdf (accessed on 18 August 2023).
8. Umphress, D.; Williams, G. Identity Verification through Keyboard Characteristics. *Int. J. Man-Mach. Stud.* **1985**, *23*, 263–273. [CrossRef]
9. Leggett, J.; Williams, G. Verifying Identity via Keystroke Characterstics. *Int. J. Man-Mach. Stud.* **1988**, *28*, 67–76. [CrossRef]
10. Canales, O.; Monaco, V.; Murphy, T.; Zych, E.; Stewart, J.; Tappert, C.; Castro, A.; Sotoye, O.; Torres, L.; Truley, G. A Stylometry System for Authenticating Students Taking Online Tests. In *Proceedings of Student-Faculty Research Day*; CSIS Pace University: New York, NY, USA, 6 May 2011; pp. B4.1–B4.6.
11. Zhong, Y.; Deng, Y.; Jain, A.K. Keystroke Dynamics for User Authentication. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; IEEE: Providence, RI, USA, 2012; pp. 117–123. [CrossRef]
12. Monrose, F.; Rubin, A. Authentication via Keystroke Dynamics. In Proceedings of the 4th ACM Conference on Computer and Communications Security—CCS '97, Zurich, Switzerland, 1–4 April 1997; ACM Press: Zurich, Switzerland, 1997; pp. 48–56. [CrossRef]
13. Ayotte, B.; Banavar, M.; Hou, D.; Schuckers, S. Fast Free-Text Authentication via Instance-Based Keystroke Dynamics. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 377–387. [CrossRef]
14. Acien, A.; Morales, A.; Monaco, J.V.; Vera-Rodriguez, R.; Fierrez, J. TypeNet: Deep Learning Keystroke Biometrics. *IEEE Trans. Biom. Behav. Identity Sci.* **2022**, *4*, 57–70. [CrossRef]
15. Sahu, C.; Banavar, M.; Schuckers, S. A Novel Non-Linear Transformation Based Multi User Identification Algorithm for Fixed Text Keystroke Behavioral Dynamics. *IEEE Trans. Biom. Behav. Identity Sci.* **2023**, *5*, 277–287. [CrossRef]
16. Tsimperidis, I.; Yucel, C.; Katos, V. Age and Gender as Cyber Attribution Features in Keystroke Dynamic-Based User Classification Processes. *Electronics* **2021**, *10*, 835. [CrossRef]
17. Buriro, A.; Akhtar, Z.; Crispo, B.; Del Frari, F. Age, Gender and Operating-Hand Estimation on Smart Mobile Devices. In Proceedings of the 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 21–23 September 2016; IEEE: Darmstadt, Germany, 2016; pp. 1–5. [CrossRef]

18. Ulinskas, M.; Woźniak, M.; Damaševičius, R. Analysis of Keystroke Dynamics for Fatigue Recognition. In *Computational Science and Its Applications—ICCSA 2017*; Gervasi, O., Murgante, B., Misra, S., Borruso, G., Torre, C.M., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O., Stankova, E., Cuzzocrea, A., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 235–247. [CrossRef]
19. Pentel, A. Predicting User Age by Keystroke Dynamics. In *Artificial Intelligence and Algorithms in Intelligent Systems*; Silhavy, R., Ed.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2019; Volume 764, pp. 336–343. [CrossRef]
20. Yan, X.; Yan, L. Gender Classification of Weblog Authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*; American Association for Artificial Intelligence: Washington, DC, USA, 2006; pp. 228–230.
21. Jones, R.; Kumar, R.; Pang, B.; Tomkins, A. 'I know what you did last summer': Query logs and user privacy. In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, Lisbon, Portugal, 6–10 November 2007; pp. 909–914. [CrossRef]
22. Tsimperidis, I.; Peikos, G.; Arampatzis, A. Classifying Users Through Keystroke Dynamics. In *Data Analysis and Rationality in a Complex World*; Chadjipadelis, T., Lausen, B., Markos, A., Lee, T.R., Montanari, A., Nugent, R., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 311–319. [CrossRef]
23. Pentel, A. Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns. In Proceedings of the Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, Bratislava, Slovakia, 9–12 July 2017; ACM: Bratislava, Slovakia, 2017; pp. 381–385. [CrossRef]
24. Cascone, L.; Nappi, M.; Narducci, F.; Pero, C. Touch Keystroke Dynamics for Demographic Classification. *Pattern Recognit. Lett.* **2022**, *158*, 63–70. [CrossRef]
25. Raul, N.; Shankarmani, R.; Joshi, P. A Comprehensive Review of Keystroke Dynamics-Based Authentication Mechanism. In *International Conference on Innovative Computing and Communications*; Khanna, A., Gupta, D., Bhattacharyya, S., Snasel, V., Platos, J., Hassanien, A.E., Eds.; Advances in Intelligent Systems and Computing; Springer: Singapore, 2020; Volume 1059, pp. 149–162. [CrossRef]
26. Schler, J.; Koppel, M.; Argamon, S.; Pennebaker, J. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*; American Association for Artificial Intelligence: Washington, DC, USA, 2006; pp. 199–205.
27. Rao, D.; Yarowsky, D.; Shreevats, A.; Gupta, M. Classifying Latent User Attributes in Twitter. In Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, Toronto, ON, Canada, 30 October 2010; ACM: Toronto, ON, Canada, 2010; pp. 37–44. [CrossRef]
28. Roy, S.; Roy, U.; Sinha, D.D. Protection of Kids from Internet Threats: A Machine Learning Approach for Classification of Age-group Based on Typing Pattern. In Proceedings of the International MultiConference of Engineers and Computer Scientists 2018, Hong Kong, China, 14–16 March 2018.
29. Roy, S.; Sinha, D.; Roy, U. Identifying Age Group and Gender Based on Activities on Touchscreen. *IJBM* **2022**, *14*, 61. [CrossRef]
30. Deutschmann, I.; Nordstrom, P.; Nilsson, L. Continuous Authentication Using Behavioral Biometrics. *IT Prof.* **2013**, *15*, 12–15. [CrossRef]
31. Yaacob, M.N.; Idrus, S.Z.S.; Ali, W.N.A.W.; Mustafa, W.A.; Jamlos, M.A.; Wahab, M.H.A. Soft Biometrics and Its Implementation in Keystroke Dynamics. *J. Phys. Conf. Ser.* **2020**, *1529*, 022086. [CrossRef]
32. Roy, S.; Roy, U.; Sinha, D.D. Analysis of Typing Pattern in Identifying Soft Biometric Information and Its Impact in User Recognition. In *Information Technology and Applied Mathematics*; Chandra, P., Giri, D., Li, F., Kar, S., Jana, D.K., Eds.; Advances in Intelligent Systems and Computing; Springer: Singapore, 2019; Volume 699, pp. 69–83. [CrossRef]
33. Lin, C.H.; Liu, J.C.; Lee, K.Y. On Neural Networks for Biometric Authentication Based on Keystroke Dynamics. *Sens. Mater.* **2018**, *30*, 385–396. [CrossRef]
34. International Standard Classification of Education (ISCED). Retrieved. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCED) (accessed on 18 August 2023).
35. Chauhan, J.; Rajasegaran, J.; Seneviratne, S.; Misra, A.; Seneviratne, A.; Lee, Y. Performance characterization of deep learning models for breathing-based authentication on resource-constrained devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–24. [CrossRef]
36. Feng, D.; Cortese, G.; Baumgartner, R. A comparison of confidence/credible interval methods for the area under the ROC curve for continuous diagnostic tests with small sample size. *Stat. Methods Med. Res.* **2017**, *26*, 2603–2621. [CrossRef] [PubMed]
37. Tsimperidis, I.; Yoo, P.D.; Taha, K.; Mylonas, A.; Katos, V. R^2BN: An Adaptive Model for Keystroke-Dynamics-Based Educational Level Classification. *IEEE Trans. Cybern.* **2020**, *50*, 525–535. [CrossRef] [PubMed]