Keystroke Dynamics Features for Gender Recognition

Abstract

This work attempts to recognize the gender of an unknown user with data derived only from keystroke dynamics. Keystroke dynamics, which can be described as the way a user is typing, usually amount to tens of thousands of features, each of them enclosing some information. The question that arises is which of these characteristics are most suitable for gender classification. To answer this question, a new dataset was created by recording users during the daily usage of their computer, the information gain of each keystroke dynamics feature was calculated, and five well-known classification models were used to test the feature sets. The results show that the gender of an unknown user can be identified with an accuracy of over 95% with only a few hundred features. This percentage, which is the highest found in the literature, is quite promising for the development of reliable systems that can alert an unsuspecting user to being a victim of deception. Moreover, having the ability to identify the gender of a user who types a certain piece of text is of significant importance in digital forensics. This holds true, as it could be the source of circumstantial evidence for "putting fingers on the keyboard" and for arbitrating cases where the true origin of a message needs to be identified.

Keywords: Keystroke dynamics, gender classification, data mining, feature selection, information gain, digital forensics.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1. Introduction

Forensic behavioral biometrics include, among others, gait, voice, signature, and keystrokes. The field of study of the latter is called *keystroke dynamics*, which is concerned with an individual's unique typing rhythms and style on a computer-based keyboard device. Over the past decade keystroke dynamics have been the subject of considerable research and their use as a tool for authentication has shown promising results (Vinayak and Arora, 2015).

Biometric technologies based on user behavior offer a number of advantages over traditional/physical ones. The information can be collected non-obtrusively without interfering with users' ongoing work by continuously requiring their consent. Collection of such data often does not require any additional hardware and thus is cost effective (Yampolskiy and Govindaraju, 2008). Of course, the efforts in searching user characteristics using behavioral biometric techniques are focused mostly to gender or age, and that is because these two characteristics are the first we want to know about an unknown user and they are the most frequently modified by malicious users. Some representative work in this field is presented below.

Li et al (2008) tried to identify the gender of a person by gait components. They conducted their experiments to find the combination of features which perform best. Using a linear SVM classifier and a voting scheme they reached accuracy of over 90%. Regarding the human voice, Barkana and Zhou (2015) proposed a new feature set for age and gender classification. They used a voice activity detection algorithm and separated their data into 4 age groups, 3 of which were divided according to gender, thus forming 7 gender-age classes. A k-Nearest Neighbor and SVM classifiers were involved in the process and the results showed an overall accuracy of 63%. Peersman et al (2011) were motivated by dangers emerging in social networks, such as pedophilia, by the users who hide their characteristics. They collected data from a Belgian social network and exploited features derived from words, emoticons, and punctuations. Using an SVM classifier they achieved an 88% success rate in separating adults from adolescents and 66% success rate in gender-age classification over 4 different classes. Similarly, Sboev et al (2016) used a corpus of Russian language texts to identify the gender of the author and in addition to recognize the sentiment of the text. For this purpose a convolutional neural network was used and their experiments concluded in 86% success rate in gender classification.

Finding the gender and/or age of an individual can be achieved by other methods, such as extracting features from facial images. Thus, Eidinger et al (2014) used a collection of facial images, presented a robust face alignment technique which explicitly considers the uncertainties of facial feature detectors, and achieved an accuracy of 67% over 7 classes in age classification and 88% in gender classification. Kalansuriya and Dharmaratne (2014) also conducted gender and age classification from facial images using neural networks, and reached a success rate of 86% for gender recognition and 74% for age classification over 4 different classes. In similar lines, Alowibdi et al (2013) conducted gender classification in their novel work. They harvested colors from user profiles on a social network, and then applied a reduction and quantization procedure which resulted in five color-base features. The best results of their experiments came from a probabilistic neural network and naïve Bayes/decision-tree hybrid classifiers which yielded an accuracy of 71%.

All the aforementioned approaches rely on machine-learning models and they showed some limitations in finding the characteristics of a user who tries to hide. For example, some of the proposed methods can only be applied if the target user has an account in a social network or there is an available facial picture. Some others use features derived from certain phrases, words, N-grams, and characters of a language, and therefore are incapable of dealing with the variety of languages in today’s Internet. Consequently, in cases where some characteristics of a user who is attempting to maintain anonymity are sought, in a forensic investigation for example, probably none of the above methods would be applicable. Clearly, the use of keystroke dynamics information could be seen as more suitable for such a problem. That is because textual communication on the Internet, such as e-mails, blog posts, and instant messages, is still the predominant mode of communication. In addition, communication between a criminal and a victim, in cases such as seduction of minors, financial frauds, cyber threats, etc., is usually done through text.

Keystroke dynamics may be defined as the detailed and precise timing information that describes when each key was pressed and when it was released as a user types on a keyboard and is first introduced in 1970’s. Since then, many methods based on keystroke dynamics have been proposed to replace password-based authentication. The features used for analyzing keystroke dynamics can be categorized into temporal and non-temporal. The most accepted temporal features are keystroke duration (the time duration a key is pressed) and digram latency (the time elapsed between two consecutive keystrokes). The latter can be expressed in four different ways, which are down-down, up-up, down-up, and up-down digram latency (Al-Jarrah, 2012), depending on whether the hit or release time of a keystroke is considered. Other temporal features, including the time associated with the trigrams, tetragrams, and generally n-grams, are referred in the work of Giot et al (2011). Common non-temporal features are typing speed (*e.g.* words per minute), frequency of errors, error correction mode (*e.g.* with “Backspace”, “Delete”, and the use of arrow-keys), the use of specific keys such as “Shift” (Bartlow and Cukic, 2006), “Alt”, “Ctrl”, “Caps Lock” (Devare, 2013), “Num Lock”, arrow keys, etc, the percentage of use of certain keys when there are two or more alternative options (Monaco et al, 2012), and others. Another non-temporal feature is the pressure applied to keys, which was used for example by Yan et al (2016). Of course, in this case, a specialized keyboard is needed for collecting data.

From the above, it becomes clear that keystroke dynamics are accompanied by a large number of features. For example, from a keyboard with n keys, using only keystroke durations and down-down digram latencies which are the most frequent used features in literature, there can be extracted n^2+n features. For n=104, which is the number of keys that most companies use for PC keyboards, the number of features amount to 10,920. Therefore, it is necessary for every keystroke dynamics study to follow a feature selection procedure so that to reduce the complexity of the produced system and to avoid high computational cost. This is one of the crucial stages of a keystroke dynamics study where a suitable and effective selection should be done. The type of classification problem and feature selection are directly associated, and the reason is that different typing features are more important when user authentication is performed than when a person's age is estimated. Thus, for example, Tsimperidis et al (2015) selected the most frequently appeared features when they tried to identify the gender of a user from smallest possible text.

In this paper we deal with gender recognition of a completely unknown user with data coming from the way he/she types. The purpose is to develop a system that will alert an unsuspecting user to the real characteristics of his/her interlocutor, as well as provide useful information in a forensic investigation where a computer crime has been committed. To optimize the proposed system, the best trade-off between accuracy and training time is sought, which is achieved by the number of keystroke dynamics features that are involved. The effectiveness of the proposed approach is demonstrated by a set of experiments with a new dynamic keystroke dataset created from recording users during the daily usage of their computers. To the best of our knowledge, this is the approach with the highest accuracy in the literature that predicts the gender of a user based on keystroke dynamics information only.

The rest of the paper is organized as follows. Section 2 describes the data acquisition, the keystroke dynamics features extraction, and the feature selection procedure. Section 3 summarizes the results obtained by comparing the performance of five well known machine-learning models, namely, support vector machine (SVM), random forest (RF), naïve Bayes (NB) classifier, multi-layer perceptron (MLP), and radial basis function network (RBFN). Section 4 discusses the results and finally Section 5 concludes the paper.

2. Method

Our experiments consist of three consecutive phases. In the first phase, we collected free-text data from the volunteers who agreed to participate in the experiment of extracting the real-life keystroke dynamics features. In the second phase, we ran a feature selection algorithm to sort the features according to their contained information. In the third phase, the performances of SVM, RF, NB, MLP, and RBFN, fed by different feature sets, are compared in terms of model accuracy, time complexity (CPU time to build model, TBM), F-score, and model robustness (ROC index).

2.1 Keystroke Dynamics Dataset

Data acquisition for the purpose of analysis of keystroke dynamics requires deploying a keylogger on a volunteer’s computing device. The volunteer may either be recorded on a fixed-text or a free-text. The term “fixed-text” implies the typing of a specific text usually in some closed environment, while “free-text” indicates the recording of volunteer during the typical daily use of his/her computer. On the one hand, by using fixed-text, researchers can focus to particular features and the sensitive data of the user remain secure. On the other hand, using free-text may reveal features which contain more information. In this work, the free-text approach is preferred as it integrates with the subject’s regular typing activities better and is less intrusive.

The continuous recording of a volunteer's typing over an extended duration of time, however, may introduce risks of disclosing passwords and personal messages to a third person. This is the main reason why there is a lack of existence of such free-text data in the literature (Vural et al, 2014), especially data that were not collected in a closed environment. For the purpose of creating a new keystroke dynamics dataset we designed and developed a free-text keylogger, called “IRecU”. This can be installed on any Microsoft Windows based devices.

A few hundred of people were approached to participate in the data acquisition process. Each of them was briefed in detail on the research objectives and imminent dangers. The volunteer recruitment process was completed with 117 people accepting to participate, “IRecU” was installed on their devices and the volunteers were asked to use the keylogger during the daily use of their device so that their natural way of typing to be recorded as well as possible. Only 75 volunteers returned log files that were saved on their devices, either physically or with an Internet service (email, shared folders, etc). The number of log files returned by a participant varied slightly with the average value being 3.3 files per volunteer.

To mitigate the effect of the aforementioned risks, firstly a signed statement was given to volunteers that the data would only be used for this research, secondly was given an option to use “IRecU” whenever they want, and thirdly an opportunity was given to review the recorded data so they can decide whether to share it or not. During the first use of “IRecU” volunteers were asked to provide their gender, among other characteristics such as age, educational level, dominant hand, etc.

Our keylogger creates .csv files with the following data for each volunteer.

```
84,#2014-06-20#,32794136,"dn"
79,#2014-06-20#,32794168,"dn"
84,#2014-06-20#,32794188,"up"
79,#2014-06-20#,32794275,"up"
32,#2014-06-20#,32794299,"dn"
32,#2014-06-20#,32794327,"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 indicates the date the action took place in the format of yyyy-mm-dd. The third field is the elapsed time since the beginning of that day (12:00am) in milliseconds, and the forth field is the action, “dn” for key-press and “up” for key-release. From these text files, it is feasible extract most of the features of keystroke dynamics. Due to the hundreds of thousands of features and in order to keep the complexity low, we calculated the most frequently used keystroke dynamics features, namely the keystroke durations and down-down digram latencies. Even now, considering *n* keys on the keyboard, we extracted n^2+n features. The duration of keystroke is calculated from the subtraction of milliseconds that correspond to the "up" action minus the ms that correspond to the "dn" action, for the same key. Similarly, the down-down digram latency results from the subtraction of ms of a "dn" action minus the ms of the previous "dn" action.

The selection of volunteers was anything but random, and this was because the sample was attempted to be representative to the general population. Table 1 shows the number of volunteers in each class, the number of the corresponding log files produced, and the shares in terms of their age, their educational level, and daily use of their device.

Table 1
Number of volunteers and logfiles per age, educational level, and daily use of device.

Characteristic	**Class**	**Female**		**Male**	
		Vol.	**Logs**	**Vol.**	**Logs**
Age	18-25	5	11	6	21
	26-35	20	66	9	36
	36-45	12	39	17	54
	46+	2	7	4	14
Educational Level (According UNESCO)	ISCED-3	5	11	9	29
	ISCED-4	2	6	3	9
	ISCED-5	7	24	4	19
	ISCED-6	16	52	9	33
	ISCED-7-8	9	30	11	35
Daily Usage in hours	0-1	5	15	3	8
	1-2	7	19	9	27
	2-4	8	23	8	31
	4-6	4	13	6	25
	6+	15	53	10	34
Total		39	123	36	125

Table 1 is a strong indication that the created dataset is balanced, both quantitatively and qualitatively. This is because, on the one hand, the number of volunteers and log files is almost equal for the two classes. On the other hand, there is a similar distribution of the number of volunteers and log files between the two classes in terms of age, educational level and daily use of computer.

After a period of 10 months of recruiting volunteers and collecting data, a dataset with 248 log files (125 labeled as “male” and 123 labeled as “female”) was created. The log files varied in size from 170 KB to 271 KB and contained data from 2,800 to 4,500 keystrokes. There are two reasons for this range in file sizes. First, "IRecU" was designed to record data of a certain size in Bytes. Therefore, depending on the time of the day the volunteer was recorded, and depending on the keys he/she was using, the number of recorded keys could have a difference of ±5%. Secondly, as stated in the consent form, no volunteer was obliged to complete the recording process, thus created files deficient in size. Eventually, files exceeding a threshold in size became accepted.

From all log files, 226 come from right-handed users, 16 from left-handed, and 6 from ambidextrous. Moreover, 240 log files created by native Greek speakers and 8 by Bulgarian native speakers, but it is not known which language they used when they were recorded. However, this is not one of our concerns, since keystroke dynamics methods prove to be language independent (Tsimperidis et al, 2015).

2.2 Feature Selection

To extract features from the log files, we developed “ISqueezeU”, a software application which reads the text files created by “IRecU” and calculates the average values of keystroke durations or down-down digram latencies. In order to avoid outliers, only the keys that have at least 5 appearances and the digrams with at least 3 appearances have been taken into account. For the other ones, the values were marked as unknown. Within this specification, a reasonable number of features was extracted, with known and unknown values, which are over 10,000 assuming the average computer’s keyboard. This huge number of features leads to systems with high time complexity and therefore a feature selection procedure is needed.

A crucial variable in this procedure is the entropy *H(x)* of the system *x*, given by:

$$H(x) = -\sum_{i=1}^{m} P(x_i)\ \mathrm{ln}P(x_i) \qquad (1)$$

where *m* is the length of vector *x*, which in the classification problem is the number of classes. Where *P*(x_i) is the probability of x_i. In the case of gender classification there are 2 classes and the probabilities of “male” and “female”, according to demographics (Statista, 2017), are both almost equal to 1/2. This means that the entropy of the system is 0.693, *i.e.* the result of Eq.1 using the class probabilities of our own dataset.

The challenge is to find the features that contain the most information, *i.e.* reduce the most the entropy of the system. The information gain (*IG*) is the measure that illustrates the ability of a feature to reduce the entropy, and it is expressed as:

$$\mathrm{IG}(x,\mathrm{feature}) = H(x) - H(x \vee \mathrm{feature}) \qquad (2)$$

To get the *H(x|feature)* it is needed to split the dataset into groups according to the value of the particular feature. Then the entropy of each group is calculated and the *H(x|feature)* is given by:

$$H(x \vee \mathrm{feature}) = \frac{1}{N}\sum_{j=1}^{k} n_j H(x_j) \qquad (3)$$

where *N* is the number of instances of the initial dataset, where *k* is the number of groups that the initial dataset was split, where n_j is the number of the *j*-th group, and where *H*(x_j) is the entropy of the *j*-th group.

If this procedure, which is also described in the work of Sharma and Dey (2012), is followed for each extracted feature, a list with the amount of information that every feature carries will emerge. In our case a section of this list is shown in Table 2, where the first 45 features with the highest *IG* are presented (features with one number indicate keystroke duration and with two numbers indicate down-down digram latency). For example, the feature with the highest *IG* is the average time needed from the pressing of “N” key to the pressing of “A” key. Similarly, the second feature in this list is the down-down digram latency of “M-O”, third is the down-down digram latency of “K-A”, etc. Respectively, in terms of keystroke durations, the feature with highest *IG* is the “A” key (7th in the list), second is “D” key (9th in the list), third is “W” key (17th in the list), etc.

Table 2
Keystroke dynamics features with the highest *IG* in gender classification.

#	Feature	Keys	*IG*	#	Feature	Keys	*IG*	#	Feature	Keys	*IG*
1	78-65	N-A	0.0897	16	78-73	N-I	0.0458	31	89-77	Y-M	0.0350
2	77-79	M-O	0.0815	17	87	W	0.0452	32	97	1 from numpad	0.0334
3	75-65	K-A	0.0706	18	77-69	M-E	0.0452	33	73-77	I-M	0.0320
4	82-73	R-I	0.0647	19	75-79	K-O	0.0445	34	83-65	S-A	0.0312
5	77-65	M-A	0.0612	20	73-79	I-O	0.0431	35	75-69	K-E	0.0304
6	84-79	T-O	0.0593	21	80	P	0.0431	36	164-16	Left Alt-Shift	0.0298
7	65	A	0.0584	22	84-73	T-I	0.0424	37	69-186	E-;:	0.0289
8	65-83	A-S	0.0553	23	68-73	D-I	0.0412	38	82-186	R-;:	0.0266
9	68	D	0.0550	24	77-73	M-I	0.0411	39	188	(comma)	0.0246
10	73-65	I-A	0.0545	25	82-69	R-E	0.0389	40	76-89	L-Y	0.0226
11	69-73	E-I	0.0543	26	79-77	O-M	0.0384	41	8-186	(backspace)-;:	0.0206
12	79-78	O-N	0.0536	27	32-75	(space)-K	0.0371	42	80-76	P-L	0.0198
13	69-32	E-(space)	0.0526	28	76-65	L-A	0.0368	43	88-79	X-O	0.0171
14	80-65	P-A	0.0503	29	71-73	G-I	0.0359	44	86-84	V-T	0.0164
15	84-69	T-E	0.0458	30	71-82	G-R	0.0353	45	79-8	O-(backspace)	0.0128

2.3 Validation of Models and Testing

The objective of model validation and testing is to ensure that the implementations of the models (SVM, RF, NB, MLP, and RBFN) are correct and work as desired. This is of particular importance when dealing with dynamic systems that are dependent on observing the data – rather than relying on a static set of equations to give decisions – which is the case in machine-learning models. There are many techniques that can be utilized to verify a model. In this work, several techniques were adopted to validate the five models used by using theoretical relationships and experimental analyses.

Firstly, to assess the performance of the five models fairly, we use the well-known cross-validation, which divides the data into 10-folds without replacement, and uses 9 folds for training and the remaining one for testing (Arlot and Lerasle, 2016). In our case, where we have 248 log files, each fold consists of 24-25 files. Also, the vast majority of volunteers delivered 3-4 log files. With these numbers it was easy to include all participants' files in one fold, so that to avoid overfitting in the case that one log file from a person could end up in the training set while another one ends up in the testing set.

Secondly, to evaluate the effectiveness of the feature selection procedure we additionally use F-score, as a combined measurement of precision and recall, because accuracy alone cannot fully give the picture of the overall performance of a model, and that is because F-score is a measurement of how balanced is the prediction between classes. For example, in our problem, there is a difference when the total accuracy of the system is 80% with successful prediction for "male" to be 100% and for "female" 60%, and when the total accuracy is 80%, again, with successful prediction for both “male” and “female” to be 80%. In the latter case, the F-score is higher. Moreover, to assess the ranking ability of the classifiers we use the area under the ROC curve (AUC) or ROC index (Guvenir and Kurtcephe, 2013).

3. Experiments and Results

Besides the features we selected in Subsection 2.2, perhaps the most recognizable feature in keystroke dynamics is the typing speed, and one might wonder if it is enough to classify users, for example male-female. Despite the fact that there is no published research (to our knowledge) that attempts something like that, we tried to find out if it is possible to recognize the gender of an unknown user only by typing speed. For this purpose we extracted the mean from all digram latencies in each log file, because this corresponds to the user's typing speed. Then the average values and the standard deviations for the “male” and for the “female” log files were calculated and compared. The results showed that the average value of all digram latencies of males was 373.04 ms with a standard deviation of 135.26 ms, while the corresponding values of females were 375.71 ms and 116.86 ms. The two samples have almost equal average value and very large standard deviation and it is obvious that typing speed cannot be used to recognize the gender of a user. It seems that the patterns to classify the users according the gender are hidden and probably sophisticated classifiers, such as neural networks, combinations of decision trees, etc, are more suitable to extract them.

For this reason, the performance of the five well-known machine-learning models, namely support vector machine (SVM), random forest (RF), naïve Bayes (NB) classifier, multi-layer perceptron (MLP), and radial basis function network (RBFN) is evaluated using the keystroke dynamics dataset and several different sets of features. The metrics calculated are the model accuracy (Acc.), which is the percentage of correctly classified instances, the time complexity (TBM), which is the CPU time has taken to build model, the F1-score (F1), which is the harmonic mean of recall and precision, and the ROC index (AUC), which is the area under the ROC curve, in order to find out the optimal set of keystroke dynamics features that lead to the system with the best performance.

We conducted numerous of experiments on the five above-mentioned models using a different number of keystroke dynamics features each time, starting at 50 features and ending at 400 features, with a step of 50, and the optimum configurations were found for all models in each of the eight datasets (50 features to 400 features). The best performance of SVM for different number of features, along with the optimal C value which is a classifier parameter set during the training phase, is shown in Table 3.

Table 3
The performance of SVM over different number of features.

# of Feats.	Statistical Values				Classifier Parameters	
	Acc.	TBM	F1	AUC	C	Kernel Type
50	73.8%	0.16	0.738	0.738	20.0	Polykernel
100	81.9%	0.13	0.819	0.818	2.0	Polykernel
150	**85.1%**	0.16	0.851	0.851	1.0	Polykernel
200	83.9%	0.19	0.839	0.839	1.0	Polykernel
250	84.3%	0.22	0.843	0.843	1.0	Polykernel
300	84.7%	0.33	0.847	0.847	1.0	Polykernel
350	**85.1%**	0.31	0.851	0.851	1.0	Polykernel
400	84.7%	0.42	0.847	0.847	1.0	Polykernel

As it is shown in Table 3, the Polykernel (polynomial kernel) is used as kernel type in every case, because it works better than the other types, such as RBFKernel (radial basis function kernel), string kernel, PUK (Pearson VII function-based universal kernel), and normalized Polykernel (Trivedi and Dey, 2013). Another finding is that the accuracy stays high for a wide range of parameter values of classifiers. For example, in the 50 features' dataset, C value can be set in range from 7 to 25 with a loss to accuracy of at most 2%. Also, in the 200 features' dataset, accuracy remains above 80.7% for a C value range of 1 to 15. These observations are also made in the other datasets.

Similarly, the best performance of RF and the corresponding optimal number of trees used over the eight datasets is shown in Table 4.

Table 4
The performance of RF over different number of features.

# of Feats.	**Statistical Values**				**Classifier Parameters**
	Acc.	**TBM**	**F1**	**AUC**	**# of trees**
50	77.0%	1.00	0.770	0.847	120
100	79.8%	2.65	0.798	0.872	240
150	80.7%	2.03	0.806	0.869	140
200	78.2%	4.60	0.782	0.874	260
250	**81.9%**	6.65	0.818	0.886	320
300	80.7%	6.15	0.806	0.888	250
350	80.7%	8.22	0.806	0.891	300
400	79.0%	8.14	0.789	0.894	260

The robustness observed in SVM was also observed in the RF. For example, in 100 features' dataset, the number of decision trees may varied from 180 to 400 with a loss to accuracy of at most 2%, while similar behavior is observed also in the other datasets.

The results for the NB classifier are in Table 5.

Table 5
The performance of NB over different number of features.

# of Feats.	**Acc.**	**TBM**	**F1**	**AUC**
50	69.0%	0.03	0.689	0.688
100	77.0%	0.02	0.770	0.847
150	77.4%	0.18	0.774	0.830
200	77.0%	0.02	0.770	0.798
250	**78.6%**	0.09	0.786	0.804
300	76.6%	0.08	0.766	0.766
350	76.6%	0.02	0.766	0.763
400	76.6%	0.02	0.766	0.767

The best performance results for RBFN are in Table 6, which also presents the corresponding optimal number of clusters for K-Means and the optimal minimum standard deviation for the clusters.

Table 6
The performance of RBFN over different number of features.

# of Feats.	**Statistical Values**	**Classifier Parameters**

	Acc.	**TBM**	**F1**	**AUC**	**# of clusters**	**min std dev**
50	81.9%	0.53	0.818	0.864	50	1.5
100	88.3%	0.73	0.883	0.891	50	1.0
150	92.7%	2.43	0.927	0.974	150	2.4
200	93.2%	2.95	0.931	0.978	150	2.4
250	93.2%	3.68	0.931	0.972	150	2.6
300	94.4%	4.31	0.943	0.981	150	2.4
350	**<u>95.6%</u>**	4.89	0.956	0.983	150	2.3
400	94.8%	5.46	0.948	0.981	150	2.4

Once again, accuracy is not highly dependent to classifier parameters. For example, in 150 features' dataset, it remains above 89.5% for a number of clusters range of 50 to 150, and for a minimum standard deviation for clusters range of 1 to 3. In the other datasets, there is a large range of values of the classifier parameters where the accuracy is close to the highest value.

Finally, the optimal configuration of MLP in terms of learning rate (L) and momentum (M) yielding the best performance and the corresponding results are presented in Table 7.

Table 7
The performance of MLP over different number of features.

# of Feats.	**Statistical Values**				**Classifier Parameters**	
	Acc.	**TBM**	**F1**	**AUC**	**L**	**M**
50	73.8%	8.55	0,738	0,781	0.5	0.2
100	81.9%	31.33	0,818	0,912	0.5	0.8
150	**<u>85.1%</u>**	73.47	0,851	0,915	0.8	0.4
200	84.7%	120.43	0,847	0,924	0.3	0.4
250	82.7%	181.93	0,827	0,906	0.7	0.2
300	81.9%	274.20	0,819	0,880	0.8	0.4
350	83.9%	373.90	0,839	0,886	0.3	0.6
400	79.8%	509.65	0,796	0,870	0.7	0.4

Like the other models, accuracy remains close to the highest value in each different dataset, for a wide range of values of learning rate and momentum of MLP.

In Tables 3 to 7, the best accuracy is bolded and underlined.

Figure 1 visualizes the accuracy of the five models on various datasets with different number of features.

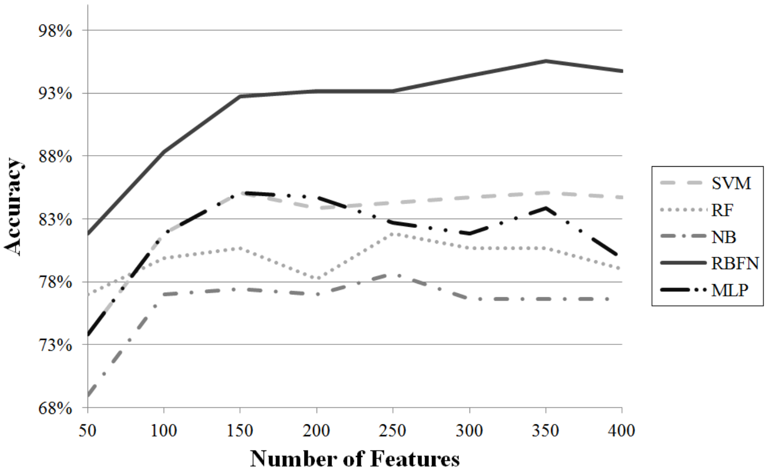

Fig. 1. Accuracy of five models on various datasets.

From Figure 1 it is derived that RBFN has always the best accuracy and NB has always the worst, among the tested models. The purpose of this paper is not to determine why one model performs better than the other, nor try to prove a classifier as the best. In literature there are a rising number of papers which claim that random forests, support vector machines or neural network methods are the predominant. One can cite significant pros and cons for each of the models. For example, neural networks work very well in learning important features from any kind of data without having to manually derive features. With regard to SVMs, they use fewer hyperparameters and in general need less grid search to achieve a reasonable accuracy. However, both NNs and SVMs, are treated as black boxes and cannot be completely interpreted. In contrary, random forests are very interpretable and relatively robust to selection bias. Finally, naïve Bayes is a linear or, in some cases, quadratic classifier. In the problem of gender classification with keystroke dynamics features, neural networks and SVMs seem to perform better, and apparently RBFN is even better, thanks to the way of minimizing the output error.

4. Discussion

From our experimental results, three main conclusions can be drawn. First, it seems that all models achieve their highest accuracy before using the maximum number of keystroke dynamics features. This is a very important indication since, as explained earlier, tens of thousands of features can be extracted from keystroke dynamics – in our case, using keystroke durations and down-down digram latencies, they were extracted 10,920 features. According to this finding, it is not necessary for a system to use a very large number of features to reach its maximum accuracy, or at least, an accuracy percentage near the maximum value. Therefore, it is possible to create gender recognition systems based on keystroke dynamics which need a short training time.

Second, the five tested models appear to have almost constant accuracy in the range of 150 to 350 features. Thus, in this range, SVM has accuracy 84.5±0.6%, RF has 80.0±1.8%, NB has 77.6±1.0%, RBFN has 94.2±1.4%, and MLP has 83.5±1.6%. Moreover, in every dataset, accuracy remains close to the highest value for a wide range of values of classifiers' parameters. These indicate that it is possible to implement robust systems that will perform reliably even if most keystroke dynamics features are absent from the available data, but even if there is no precise tuning of their parameters.

Third, in the case of 350 features, the RBFN model predicts correctly the gender of an unknown user in 19 times out of 20. To our knowledge, the accuracy of 95.6% is the highest prediction rate of user gender by using keystroke dynamics, in the literature. This means that it is possible to develop quite accurate systems which perform gender recognition with data coming only from the simplest form of communication between users, the text, without violating privacy rights, since all data are being extracted by the way a user types and not by what he/she types. Such systems will operate as shown in Figure 2. When user 2 sends a message to user 1, then keystroke dynamics data are simultaneously sent to a keystroke dynamics gender recognition system. By receiving sufficient amounts of data, the gender recognition system sends to user 1 the prediction of the gender of user 2. If user 2 is unknown to user 1, then the prediction of the system is a warning. Of course, the process can work and vice versa, by sending information to user 2 for the gender of user 1.

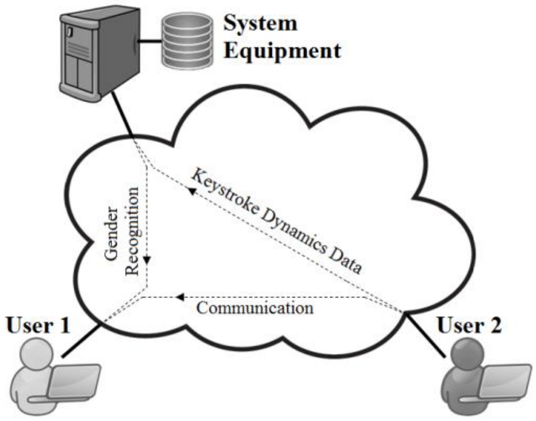

Fig. 2. The operation of keystroke dynamics gender recognition system.

A proposal for implementation of such a system is its embedment to operating systems. In this way, once the keystroke dynamics data collected, and once the desired features extracted, they are sent to a dedicated server which is responsible for deciding on the gender of the user. There are two points worth paying attention to. First, by sending keystroke dynamics features, instead of the data itself, it is not possible to mine sensitive or personal information such as passwords, personal messages, etc. Second, knowing that patterns are not as prominent as the typing speed, but rather they are quite hidden, it would be very difficult for a user to modify his/her typing rhythm so as to conceal his/her characteristics, especially when the proposed system dynamically adapts its parameters triggered by the availability of new training data. However, in order to prove the above allegations, further research will be needed, but this will go beyond the objectives of this study.

It is reminded that the log files include data from 3,650±850 keystrokes, which implies that the phrase "sufficient amount of data" mentioned above corresponds to approximately 610±140 words (since the average characters per word, including the space between words, is around 6) (Bochkarev et al, 2015), or to 55±12 instant messages according to O’Connor et al (2010), and Thurlow and Poff (2013). After the user has received this number of instant messages, the system will warn him/her about the gender of his/her interlocutor, with an accuracy that exceeds 95%. However, no further research has been carried out to identify the minimum number of messages at which this performance of the system is achieved; we will leave this for future work.

Other findings are the very low performance of MLP in terms of time complexity, where NB outperforms, but with low accuracy. SVM presents the second highest accuracy after RBFN and the second lowest time complexity after NB, among the five models. However, MLP has the second best AUC value after RBFN and the second best F1 value, together with SVM, after RBFN, again. From the above and from Tables 3 to 7 it follows that the RBFN has the best performance in terms of accuracy, F1 value, and AUC value, while it has comparable time complexity to the fastest models.

Finally, as shown in Tables 3 to 7, in each case the accuracy is equal to the value of F1 or they have a difference less than 0.002. Examining the equations that give accuracy and F1, which are Acc=(TP+TN)/(TP+FP+FN+TN) and F1=2TP/(2TP+FP+FN), it follows that true positives (TP) are equal, or almost equal, to true negatives (TN). When the dataset is balanced, as it happens to be in our case, then the above finding is interpreted as that there is no bias in prediction.

5. Conclusion

Many times it is necessary to know some characteristics of an Internet user, either for security reasons or for greater exploitation of the services offered. Existing methods that achieve gender recognition of a user either require specific data, such as facial images, or are intrusive. On the contrary, keystroke dynamics provide a non-intrusive low-cost method from data coming only from the simplest and most common way of communicating between users, *i.e.* text.

Based on this idea, this study presents a process in which the most suitable keystroke dynamics features are selected to identify the gender of an unknown user. To accomplish the objective, a new keystroke dynamic dataset was created from recording users during daily usage of their computers, and 248 log files were collected. The information gain of each attribute was then calculated and they were ranked according to the reduction of entropy of the system. Variations of the original dataset with different number of features were formed and then were fed into five well-known classification models. The experimental results on the keystroke dynamic datasets proved that it is possible to create quite reliable systems that can recognize the gender of an unknown Internet user with 95.6% accuracy using only a few hundred features.

Having the ability to identify the gender of a user who types a certain piece of text has significant value in digital forensics, targeted advertisement, and behavioural biometrics, as the utility of the proposed framework has been proven in dealing with the source of circumstantial evidence for “putting fingers on keyboard” and for profiling the characteristics of the users. However, we note that the deployment of such a system must be in accordance with the current legal and regulatory framework, as the unauthorized analysis of keystrokes may entail privacy violations, which might involve sensitive personal information (*e.g.*, in accordance to the EU legislation).

An extension of this work involves a second round of data collection by recording volunteers during the daily usage of their computers. With an enlarged dataset new studies can be done on other user characteristics, such as handedness, educational level, etc, since all classes will be adequately represented. Furthermore, the proposed method can be approached, as well, with Dempster-Shafer’s theory of evidence, considering each keystroke dynamics feature as a “view” which must be combined with other "views" to produce the final outcome.

References

Al-Jarrah MM. An anomaly detector for keystroke dynamics based on medians vector proximity. J Emerg Trends Comput Inf Sci 2012; 3(6): 988-93.

Alowibdi JS, Buy UA, Yu P. Language independent gender classification on Twitter. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara, Ontario, Canada: 2013. p. 739-43.

Arlot S, Lerasle M. Choice of V for V-fold cross-validation in least-squares density estimation. J Mach Learn Res 2016; 17: 1-50.

Barkana BD, Zhou J. A new pitch-range based feature set for a speaker’s age and gender classification. Appl Acoust 2015; 98: 52-61.

Bartlow N, Cukic B. Evaluating the reliability of credential hardening through keystroke dynamics. In: Proceedings of the 17th International Symposium on Software Reliability Engineering. Raleigh, NC, USA: 2006. p. 117-26.

Bochkarev V, Shevlyakova A, Solovyev V. Average word length dynamics as indicator of cultural changes in society. Soc Evol & Hist 2015; 14(2): 153-75.

Devare M. Mixing and matching human traits using hand typing. Int J Comput Appl 2013; 73(18): 1-5.

Eidinger E, Enbar R, Hassner T. Age and gender estimation of unfiltered faces. IEEE Trans Inf Forensics Secur 2014; 9(12): 2170-9.

Giot R, El-Abed M, Rosenberger C. Keystroke dynamics overview. In: Yang J, editor. Biometrics. InTech; 2011. p. 157-182.

Guvenir HA, Kurtcephe M. Ranking instances by maximizing the area under ROC curve. IEEE Trans Knowl Data Eng 2013; 25(10): 2356-66.

Kalansuriya TR, Dharmaratne AT. Neural network based age and gender classification for facial images. Int J Adv ICT Emerg Reg 2014; 7(2): 1-10.

Li X, Maybank SJ, Yan S, Tao D, Xu D. Gait components and their application to gender recognition. IEEE Trans Syst, Man, Cybern — Part C: Appl Rev 2008; 38(2):145-55.

Monaco JV, Bakelman N, Cha S, and Tappert CC. Developing a keystroke biometric system for continual authentication of computer users. In: Proceedings of the 2012 European Intelligence and Security Informatics Conference. Washington, DC, USA: 2012. p. 210-6.

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA. From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media. Washington, DC, USA: 2010. p. 122-9.

Peersman C, Daelemans W, Van Vaerenbergh L. Predicting age and gender in online social networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents. Glasgow, Scotland, UK: 2011. p. 37-44.

Sboev A, Voronina I, Litvinova T, Dmitry Gudovskikh D, Rybka R. Deep learning network models to categorize texts according to author's gender and to identify text sentiment. In: Proceedings of the 2016 International Conference on Computational Science and Computational Intelligence. Las Vegas, NV, USA: 2016. p. 1101-6.

Sharma A, Dey S. Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. Int J Comput Appl Special Issue Adv Comput Commun Technol HPC Appl 2012; 3: 15-20.

Statista [Internet]. Internet Usage Rate Worldwide in 2017, by Gender and Region; c2017 [cited 2017 Aug 18]. Available from: https://www.statista.com/statistics/491387/gender-distribution-of-internet-users-region/

Thurlow C, Poff M. Text messaging. In: Herring SC, Stein D, Virtanen T, editors. Handbook of the pragmatics of computer-mediated communication. Berlin: De Gruyter Mouton; 2013. p. 163-90.

Trivedi SK, Dey S. Effect of various kernels and feature selection methods on SVM performance for detecting email spams. Int J Comput Appl 2013; 66(21): 18-23.

Tsimperidis I, Katos V, Clarke N. Language-independent gender identification through keystroke analysis. Inf Comput Secur 2015; 23(3): 286-301.

Vinayak R, Arora K. A survey of user authentication using keystroke dynamics. Int J Sci Res Eng & Technol 2015; 4(4): 378-84.

Vural E, Huang J, Hou D, Schuckers S. Shared research dataset to support development of keystroke authentication. In: Proceedings of 2014 IEEE International Joint Conference on Biometrics. Clearwater, FL, USA: 2014. p. 1-8.

Yampolskiy RV, Govindaraju V. Behavioural biometrics: a survey and classification. Int J Biom 2008; 1(1): 81-113.

Yan Q, Wang W, Qin R, Jiang H, Yang B, Zhang B. Study on keystroke dynamic with feature of pressure. In: Proceedings of 2016 International Conference on Artificial Intelligence and Computer Science. Guilin, China: 2016. p. 475-80.