Michał MALINOWSKI[1], Zuzanna KRAWCZYK-BORYSIAK[1]

Warsaw University of Technology, Faculty of Electrical Enigneering (1)

doi:10.15199/48.2024.06.03

Recognizing User Emotion Based on Keystroke Dynamics

Abstract. *The paper presents a study concerning recognizing user emotion based on keystroke dynamics of the written text. At first, the analysis of the dataset used in the task is performed. Followed by the training and the effectiveness assessment of classical methods: Naive Bayes, K-Nearest Neighbours, Random Forest, and Multilayer Perceptron applied to the classification of provided samples to one of four emotions: anger, calm, happiness, sadness. The precision, recall, F1-score and time performance are evaluated. The Random Forest and MLP classifiers performed best, with an overall F1 measure of 84.83% and 80.47%, respectively. The scenario for extending the data set is proposed, along with the analysis of classification results of new data.*

Streszczenie. *Artykuł przedstawia badania dotyczące rozpoznawania emocji użytkownika na podstawie dynamiki naciśnięć klawiszy wpisywanego tekstu. W pracy przeprowadzono analizę wykorzystywanego zbioru danych, wytrenowano oraz dokonano oceny skuteczności klasycznych metod takich jak: naiwny klasyfikator Bayesa, metoda najbliższych sąsiadów, las losowy oraz perceptron wielowarstwowy, zastosowanych do przyporządkowania danych do jednej z czterech emocji: złości, spokoju, radości lub smutku. Uzyskane wyniki zostały ewaluowane z wykorzystaniem miar precyzji, czułości oraz F1, oceniono również wydajność czasową. Las losowy oraz perceptron wielowarstwowy osiągnęły najlepsze wyniki, z wynikiem F1 równym odpowiednio 84.83% i 80.47%. Zaprezentowano również scenariusz rozszerzenia zbioru danych, razem z analizą wyników klasyfikacji nowych danych.* ***(Rozpoznawanie emocji użytkownika na podstawie dynamiki naciśnięć klawiszy)***

Keywords: keystroke dynamics, emotion recognition, MLP, decision trees, machine learning
Słowa kluczowe: dynamika naciśnięć klawiszy, rozpoznawanie emocji, perceprtron wielowarstwowy, drzewa decyzyjne, uczenie maszynowe

Introduction

The keyboard has become a tool used daily by people in the 21st century, and how we use it can be a source of much information. The rhythm, tempo and other aspects of the way we type allows for authenticating a user [1] in place of traditional methods. It can even be employed for continuous authentication during computer usage [2], enabling recognition of the situation when an unauthorized individual starts to use an unlocked machine. The paper is focused on another possible application of the keystroke dynamics analysis: recognition of emotions felt by users. The authors compare the effectiveness of different machine learning algorithms for the classification task of emotions felt by a user while inputting a given text. The second aim of this paper is to expand the chosen data set with additional data gathered from the gamers' community.

The next chapter of the paper presents existing solutions and how keyboard dynamics are used in similar tasks. Then, the training dataset used in the task is described. The fourth chapter gives insight into methods used for classification and is followed by the classification methodology. In the following chapter, the methods' effectiveness is compared. In the final chapters, the authors demonstrate the process of gathering and applying trained models to additional data and discuss the results.

State of the art

Many studies have been performed based on recognizing human emotion. For a long time most of these studies focused on deducing emotion from facial expressions [8] or other visual sources such as poses of the entire body [9]. The possibility of using keystroke dynamics for recognizing emotion has recently gained interest, but has been proposed and validated as a possibility as far back as 2013 [10]. A study [11] used keystroke dynamics combined with text analysis to great effect. A two class models meant to discern a specific emotion from all others was proposed and proven as viable for the task [12]. Recently, models for user authentication capable of recognizing a user in different emotional states were proposed in [13] and the use of emotional features extracted from keystroke dynamics increased model accuracy significantly.

Training data

The publicly available data set – EmoSurv [3] was used as training data. The data was gathered as part of an experiment studying the viability of using keystroke dynamics for emotion recognition [4]. Emotions were induced in users by way of presenting them with specially chosen emotionally charged videos and then keystroke data was gathered immediately afterwards. The result is a data set containing a wide array of cases for four emotions: anger, sadness, happiness and calm.

Demographic data: The data set includes extensive data on the participants. Figures 1 and 2 present a subset of the demographic information taken into consideration in this study.

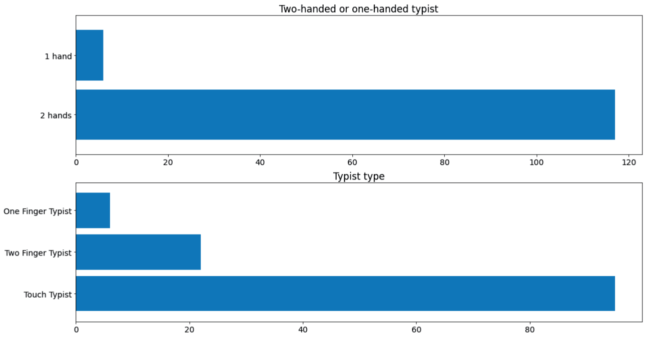

Fig. 1. Typing style distribution

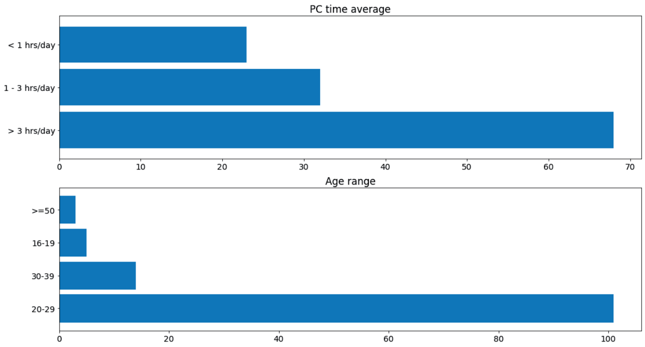

Fig. 2. Age and PC Time distribution

Figure 1 shows the distribution of how participants use a keyboard. Very few of them used one hand for text input which is to be expected from users nowadays. Most also declared being touch typists, meaning they use all of their fingers for writing. Figure 2 illustrates personal data relevant for the study: age range and the average daily time using a computer. Over 50% of participants were in the 20-29 age range, as that is the group easiest to reach with this kind of study. And while most of the participants are avid computer users, spending more than 3 hours each day before a computer, around half of the people polled declared lesser averages.

In summary, the data used comes from a fairly diverse survey group. Despite a certain degree of predominance of younger people and those using the computer more, the data set should perform well in this task, as we can expect a similar distribution of respondents later in the survey.

Classification data: The most crucial section of the data set is the part of the data that are used to classify the emotions of the subjects. The first subset of this is frequency data. It contains parameters describing the frequency of use of specific keys, for example those related to typing errors like backspace or delete or the use of capital letters as well as typing speed. The data set contains frequency data for each participant grouped by induced emotion. The second, more numerous group is timing data, describing the rhythm and pace of typing. The most important parameters are:

- Dwell time - time between pressing and releasing a key.
- Flight time - time between releasing one key and pressing another. Describes the time it take the user to decide on the next key to press.
- N-graphs - time between N presses, releases or combination of both actions. We're typically most interested in digraphs (happy -> ha,ap,pp,py) and trigraphs (happy -> hap,app,ppy).

The method for calculating these attributes is described in an addendum for the data set [3]. The data was recorded separately for participants inputting a given text (Fixed Text) and typing whatever they wanted (Free Text).

Methods used

The purpose of this work is to compare the effectiveness of different methods in the task of classifying emotions based on the above-described data set. Four methods from different branches of machine learning were chosen in order to determine the best direction for further research – Naive Bayes, K-Nearest Neighbors (KNN), Random forest and Multilayer Perceptron (MLP).

Naive Bayes classifier is chosen as baseline model. As features of keystroke dynamics are deeply interrelated we expect poorer performance of the model than in case of other proposed solutions.

K-Nearest Neighbors method was chosen due to the easiness of training step. For the discussed problem the authors used standard Euclidean distance measure. The value of K determining the number of neighbors analyzed, is an important parameter in this method. If the data has a lot of noise and outliers, a larger K number will probably perform better, but it may lead to over fitting the classifier. Due to this the authors decided on a K value equal to 7, to consider a possibility of noise while avoiding the loss of generalization ability.

Decision trees are widely used in classification problems based on tabular data. Although they are easy to interpret and quite flexible in their applications, in their base version, they are prone to over-fitting and have high variance in the results. For this reason, it was decided to use the Random Forest variant - an ensemble version of the algorithm. It is based on a combination of multiple decision trees trained on a randomly selected subset of the learning data, then each tree individually performs sample classification based on its own rules. The resulting classification decision is achieved using majority voting from among the trees in the forest.

The last method used is a simple Multilayer Perceptron (MLP) neural network. For described task we used a variant chosen experimentally with 3 hidden layers featuring 50, 25 and 4 neurons respectively, using the ReLu activation function and the Adam optimizer. The model training ended after 161 epochs, due to fulfillment of early-stopping criterion (the accuracy of the validation set did not increase for 20 epochs). The Fig. 3 shows loss function value for training data set along with accuracy value computed for validation set (10% of data randomly put aside from training data set).

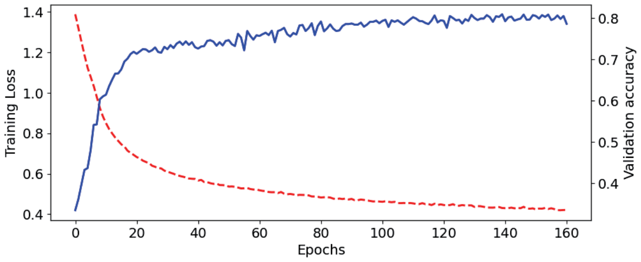

Fig. 3. Training process of the MLP classifier

Methodology

The data set we use for training divided the emotional spectrum into four categories [4], based on the distribution relative to arousal: low (characteristic for sadness and calm) and high (anger and happiness), and attitude: positive (happiness, calm) and negative (anger, sadness). The emotion categories are depicted in Fig. 4 . As the base for training we used the Fixed Text subset of data [3]. After removing elements related to the neutral emotion, which were those gathered before inducing any emotion, as well as those where the participant answered incorrectly to a question regarding what the video he watched was about, we were left with 17883 usable samples. The distribution of these samples by emotion is shown in table 1 Next according to a suggestion from the original work [4] standardization of parameters was performed for each user separately. The goal of this approach is to focus on variance and changes in parameters, to obtain better generalization of the problem. The final step in preparing the data set for learning was to supplement the timing data with frequency and demographic data. These param-

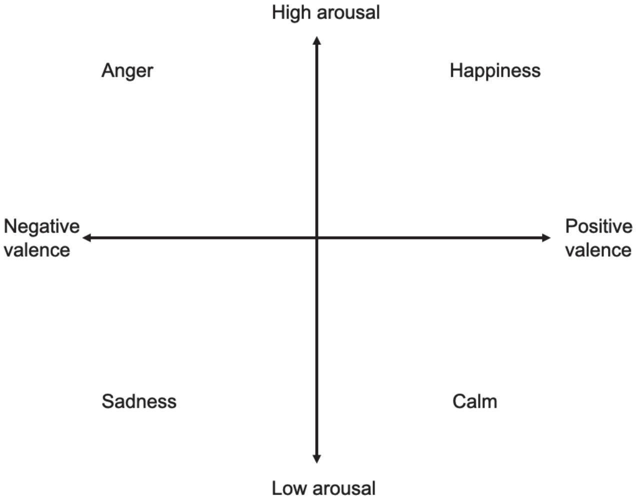

Fig. 4. Division of emotions based on the paper [4]

Table 1. Emotion distribution in training data

Emotion	Number of samples	Percentage of all samples
Anger	4024	22.5%
Calm	5097	28.5%
Happiness	4088	22,9%
Sadness	4674	26.1%

Table 2. Table comparing the results for all classifiers

Classifier	Precision	Recall	F1-Score	Execution time
All classes				
Naive Bayes	0.2803	0.3709	0.3138	0.357s
KNN	0.8026	0.8017	0.8021	1.806s
Random Forest	0.8485	0.8481	0.8483	2.699s
MLP	0.8072	0.8029	0.8047	16.76s
Anger				
Naive Bayes	0.0	0.0	0.0	n/a
KNN	0.7982	0.7932	0.7957	n/a
Random Forest	0.8587	0.8458	0.8522	n/a
MLP	0.8023	0.7631	0.7822	n/a
Calm				
Naive Bayes	0.3688	0.3831	0.3758	n/a
KNN	0.7901	0.8069	0.7984	n/a
Random Forest	0.8427	0.8419	0.8423	n/a
MLP	0.7808	0.8089	0.7946	n/a
Happiness				
Naive Bayes	0.3792	0.6945	0.3906	n/a
KNN	0.8352	0.8300	0.8326	n/a
Random Forest	0.8711	0.8743	0.8727	n/a
MLP	0.8635	0.8263	0.8445	n/a
Sadness				
Naive Bayes	0.3729	0.4059	0.3887	n/a
KNN	0.7867	0.7767	0.7817	n/a
Random Forest	0.8213	0.8301	0.8252	n/a
MLP	0.7821	0.8130	0.7972	n/a

eters have a great impact on the way the subjects write and are a source of information that classifiers can use effectively.

Effectiveness comparison

The data was split into training and testing sets in a ratio of 80 to 20. The precision, recall and F1-score metrics were used to evaluate the performance of the classifiers. Also execution time was considered, to compare computational costs of the considered approaches. The resulting metrics for each classifier are shown in Table 2. As expected, the Naive Bayes classifier performed by far the worst. The examined data set does not fit the limitations of this method, hence the F1-Score lower than 0.5 is not surprising.

The K-Nearest Neighbors method performed much better. Despite its relative simplicity, it managed to achieve an F1-Score of 0.8. This is a satisfactory result, providing opportunities for further development.

By far the best of the methods was Random Forest. The method achieved an F1-Score of 0.85, with a slight increase in execution time relative to the KNN method.

The Multilayer Perceptron achieved a result worse than Random Forest, and at a much higher computational cost - the execution time here was almost four times that of the previous classifier. However, MLP's result is still promising, and the authors does not yet exclude the possibility of using some variant of this method in further research.

In summary, the most promising approach is Random Forest algorithm. MLPs and neural networks in general also deserve consideration here, as potentially effective classification methods for this problem. Naive Bayes should be rejected altogether. Analysis of its performance on this set shows that the use of this method is not right for this task.

Gathering additional data

The next stage of the study was performing an experiment on self-acquired data. Due to personal interests and access to a wide set of potential respondents, the authors decided to perform a study of the impact of multiplayer video games on players' emotions. A simple web application with graphical interface was prepared for this purpose.

The application consists of two sections. In the first, respondents are asked to transcribe the given text into the text box below. Using keydown and keyup events, the pressing and releasing of specific keys is recorded while subjects are typing.

After the text is transcribed, respondents are asked to fill out a short form collecting additional information crucial to our study. Participants are asked to self-identify the emotion they felt when filling out the questionnaire. This information will give us the opportunity to compare classical methods of emotional self-assessment with the machine-defined prediction of our classifiers. The second element of the form is data on the subject's gaming experience. We find out what the person played, how long the game lasted, and whether they lost or won more during the game. We also ask whether the last game played ended in a win or a loss. This knowledge can be useful when analyzing the impact of recency bias on emotions. Finally the demographic data of the same type as presented in the basic data set are gathered. During the course of the study, we managed to collect close to 8000 samples from 20 participants.

Applying trained models to gathered data

The collected data were analysed with two best classifiers trained on the basic data set - Random Forest and MLP.

Single sample analysis: The first approach is analyzing single samples, represented by time data supplemented by frequency data described in Section . The majority of respondents declared happiness, which was an expected result, given research conducted on the subject[5]. Happiness was also the emotion in which both classifiers agreed with self-assessment most often. Percentage-wise, at the individual sample level, Random Forest agreed with this assessment with an F1-Score equal to 57%, and MLP with value equal to 69%. It can be concluded that happiness is the emotion most accurately self-assessed by users.

Overall, the agreement of the MLP classifier with emotions self-assessed by respondents was higher (achieving an accuracy of 49%) than in the case of Random Forest, where the accuracy value was only 33%. However, we must remember that in this analysis, this characteristic does not inform us about the quality of the classification. It is merely an indicator of agreement between the self-assessment of the user completing the survey and the classifier. Metrics for both classifiers are shown in table 3.

The most interesting point here are the differences in the results of the two classifiers. In the case of the emotion of sadness, the MLP classifier mostly agreed with users' self-assessments, but the Random Forest classifier classified most of these samples as the emotion of anger. This disagreement may suggest two things. First, the multilayer

Table 3. Comparing the results on additional data

Classifier	Precision	Recall	F1-Score	Accuracy
All classes				
Random Forest	0.2689	0.2643	0.2619	0.3323
MLP	0.4249	0.4215	0.4154	0.4882
Anger				
Random Forest	0.0702	0.0967	0.0814	n/a
MLP	0.1944	0.1331	0.1580	n/a
Calm				
Random Forest	0.3671	0.2461	0.2946	n/a
MLP	0.4168	0.3203	0.3623	n/a
Happiness				
Random Forest	0.5248	0.6187	0.5679	n/a
MLP	0.7114	0.6685	0.6893	n/a
Sadness				
Random Forest	0.1134	0.0957	0.1038	n/a
MLP	0.3771	0.5641	0.4520	n/a

perceptron may be better suited to detect the emotion of sadness. The second option however, could suggest that the Random Forest, which performed better on the data from the original set, is right instead and users rating themselves sad were actually just angry more often than not, which would agree with people's existing aversion to admitting anger [6]. In the case of the emotion of calm, both classifiers behaved similarly. The results for these samples showed mostly a mix of sadness and happiness. The MLP method was more likely to classify these emotions as actual calmness. Once again, we see here a serious mismatch between self-assessment and classification. This time, however, both classifiers disagree with the users' assessment. One could conclude that the majority of respondents choosing the calm option actually felt a mixture of happiness and sadness - or even started typing while feeling sadness, which faded. Both classifiers struggled with samples where the user self-assessed anger. Machine predictions for both classifiers were split between all four possibilities, with anger being the option least often chosen by both MLP and Random Forest. MLP had a preference towards choosing Sadness, but Random Forest was split close to evenly between Calm, Happiness and Sadness. The classifiers may simply struggle with Anger as an emotion, since it impacts people in many different ways, or users may not actually be feeling angry when picking this option - just sad and frustrated.

Aggregate analysis: An alternative approach may be to analyze all samples aggregated while subject was writing surveys text (example on Fig. 5). We achieve such an effect by summing the classification probabilities of each sample to a particular emotion. In this way, we obtain sum values indicating the classifier's opinion on the probability of the entire survey belonging to each class of emotion. Such data can be interpreted as a percentage of an emotion felt, as the strength of a given emotion, or as simply the probability of that emotion occurring during the survey. The emotions we feel are not mutually exclusive [7] , so such an interpretation of the results may be reasonable.

Conclusion

Two classifiers emerged as the most suited for the task - Random Forest and Multilayer Perceptron. Though Random Forest achieved better results on the training data set, the MLP classifier seems to have better generalization capabilities on newly gathered data. The next important step would be adding a neutral emotion as additional class which would encompass all cases where game play was not engaging for the player enough to feel strong emotion.

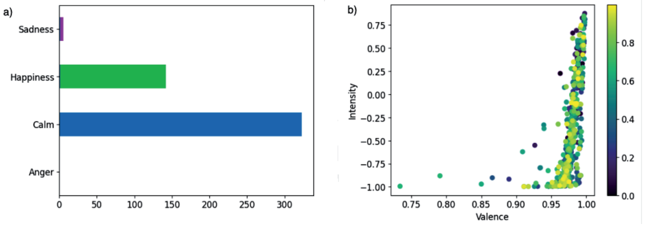

Fig. 5. a) Overall emotion strength b) Emotion over time

The results divergent from self-assessment may suggest that emotions felt can be a complex mixture of a few of them and thus not so straightforward to classify. In future work we plan to apply the LSTM network to the time keystroke data to consider the whole sample range.

An interesting idea for further development of this work would be to turn the site into a portal for evaluating one's own emotional state. The portal could assess the user's emotions based on a summative analysis.

The study showed that keystroke dynamics could be widely used as a source of information regarding the user.

REFERENCES

[1] https://github.com/jatanloya/KeystrokeAnalysis, accessed 21.05.2023

[2] Xiaofeng Lu et al., Continuous authentication by free-text keystroke based on CNN and RNN, Computers Security, Volume 96, 2020, 101861, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2020.101861.

[3] https://ieee-dataport.org/open-access/emosurv-typing-biometric-keystroke-dynamics-dataset-emotion-labels-created-using, accessed 21.05.2023

[4] Maalej, Aicha et al. Investigating Keystroke Dynamics and Their Relevance for Real-Time Emotion Recognition. http://dx.doi.org/10.2139/ssrn.4250964

[5] Johannes Niklas, Vuorre Matti and Przybylski Andrew K. 2021Video game play is positively correlated with well-beingR. Soc. open sci.8202049202049

[6] Fernandez, Ephrem, et al. "Social Desirability Bias Against Admitting Anger: Bias in the Test-Taker or Bias in the Test?" Journal of Personality Assessment 101.6 (2019): 644-52. Web.

[7] Berrios R, Totterdell P, Kellett S. Eliciting mixed emotions: a meta-analysis comparing models, types, and measures. Front Psychol. 2015 Apr 15;6:428. doi: 10.3389/fpsyg.2015.00428.

[8] Hamann, Stephan B, et al. "Recognizing Facial Emotion." Nature (London) 379.6565 (1996): 497. Web.

[9] Schindler, Konrad, Luc Van Gool, and Beatrice De Gelder. "Recognizing Emotions Expressed by Body Pose: A Biologically Inspired Neural Model." Neural Networks 21.9 (2008): 1238-246. Web.

[10] Tsui, Wei-Hsuan, Poming Lee, and Tzu-Chien Hsiao. "The Effect of Emotion on Keystroke: An Experimental Study Using Facial Feedback Hypothesis." 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2013): 2870-873. Web.

[11] Nahin, A.F.M. Nazmul Haque, Jawad Mohammad Alam, Hasan Mahmud, and Kamrul Hasan. "Identifying Emotion by Keystroke Dynamics and Text Pattern Analysis." Behaviour & Information Technology 33.9 (2014): 987-96. Web.

[12] Hippe, Zdzisław S, Juliusz L Kulikowski, and Teresa Mroczek. "Usefulness of Keystroke Dynamics Features in User Authentication and Emotion Recognition." Human-Computer Systems Interaction. Vol. 551. Switzerland: Springer International AG, 2018. 42-52. Advances in Intelligent Systems and Computing. Web.

[13] Jia, Weichen, et. al. "High Security User Authentication Based on Piezoelectric Keystroke Dynamics Applying to Multiple Emotional Responses." IEEE Sensors Journal 22.3 (2022): 2814-822. Web.