Personal characteristics prediction with behavioral chronometric data: Does the time you take to recognize words disclose your gender, age group, and educational level?

Rastislav Hronský
STUDENT NUMBER: 2034432

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Dr. Emmanuel Keuleers
Dr. Michal Klincewicz

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
January 2020

Preface

This work has been made possible thanks to the very attentive supervision by Dr. Emmanuel Keuleers, his profound expertise in psycholinguistic research and very stimulating input. Furthermore, I would like to express my gratefulness to my parents, who provided me with all the necessary mentoring and financial means, that a university student needs, in order to successfully complete the studies.

Personal characteristics prediction with behavioral chronometric data: Does the time you take to recognize words disclose your gender, age group, and educational level?

Rastislav Hronský

Suppose a person labels sequentially presented words as positive in case they know the word, negative otherwise. Given a series of such responses, can we predict the person's personal characteristics using the recognition times? Firstly, this is an interesting topic regarding privacy issues and personal information contained in chronometric behavioral data. Secondly, it is worth investigating the usefulness of machine learning methods applied to lexical decision data, given what we know about lexical processing from a prior research. This problem has been tackled before by developing a Pearson correlation based classifier. In the present work, we explored the performance of a probabilistic (Bayesian) classifier, motivated by the ability to reflect the recognition time distributions better and thus classify better. We developed models to predict participants' gender, educational level and age group, based on data from the English Crowdsourcing Project. Eventually, we arrived at a performance similar to the prior work when tested on unseen data (<3% above the majority baseline). However, the probabilistic classifier classified the training data with a remarkable accuracy exceeding 95%. Given the dataset used, the results lead us to conclude, that the proposed Bayesian classifier does not predict the aforementioned personal characteristics reliably, and an overall scepticism regarding the predictive capabilities of lexical decision data.

1. Introduction

The time it takes us to articulate sounds, the time it takes us to recognize a written word, the pauses we make between keystrokes when typing, footstep timing, the long circadian rhythms - all of these are examples of chronometric information (duration of an action) about some kind of human behavior. The chances are that partly the data consist of noise, and partly the data can be explained by certain circumstances. When dealing with behavioral data, the latter component is often related to some personal characteristic of an individual. Given such a relationship, it is reasonable to expect that the information on one end may lead to the information on the other one. In other words, certain behavioral data may be predictive of the person's characteristics, or vice versa. The goal of the current work is to exploit such a relationship in an effort to develop a machine learning classifier based on it. On the contrary, proving an existence of the relationship by means of hypothesis testing is not intended. The relationship is assumed to be justified based on prior work in related fields.

The present work operationalizes this reasoning in terms of lexical decision data based gender, age group and educational level prediction task. The choice of this data modality is motivated by previous findings of the psycholinguistic research, and the

potential societal implications regarding personal privacy. In the following section, the nature of lexical decision data is introduced.

1.1 Lexical Decision

A lexical decision task is an experimental procedure in psychology to quantitatively analyze semantic memory and lexical access (Meyer and Schvaneveldt 1971). When performing a lexical decision, participants are asked to identify a presented string of letters as word or non-word, and response times (RTs) are measured. Eventually, a session consisting of series of lexical decisions results in a sequence of RTs and corresponding words. A vocabulary test can be implemented as a type of a lexical decision task in which the participants are asked whether they know the presented word or not, and the non-words provide a way to control for dishonest responses. Additionally, performing a vocabulary test may provide the participant with some extra motivation: receiving an estimate of her vocabulary size at the end. This makes it more suitable for studies that do not acquire data under laboratory conditions, such as online-based studies. The dataset we used origins from such a vocabulary test and will be introduced in greater detail later.

1.2 Scientific Motivation

One of the common efforts in psycholinguistics is explaining the variance of lexical decision performance. The word frequency effect is an observation that people recognize higher frequency words faster than lower frequency words (Monsell, Doyle, and Haggard 1989). However, the way people use and are exposed to language differs across multiple factors. The age, gender, education, occupation, religion, etc. are common ways of describing socioeconomic background of an individual. Some of the aforementioned factors have been shown to modulate people's experience with language by investigating their lexical processing. For example, the education and age were shown to be important factors influencing the vocabulary size (Keuleers et al. 2015), and higher vocabulary size leads to faster lexical decisions (Keuleers et al. 2012). The Section 2.3 describes the prior findings regarding lexical decision performance further.

In the recent decade, the term *filter bubbles* started to resonate in terms of an important societal issue. This phenomenon has emerged from the trend of highly personalized social media and content providers, which utilize data mining methods, providing the user with content based on the specific profile or history of activities (Pariser 2011). In fact, evidence for narrowing the scope of viewpoints over time by a content recommender system has been demonstrated (Nguyen et al. 2014). This finding extends the idea of societal environments and language exposure by events happening on social media.

Motivated by the previous arguments, the current work develops a naïve Bayes classifier that predicts age group, educational level and gender, based on lexical decision response times. To the best of our knowledge, prior to the current work, similar efforts were only made once, using a Pearson correlation based k-NN classifier (Qin 2018). The method we propose is a Bayesian, probabilistic classifier, making use of a more fine-grained knowledge about the underlying recognition time distributions.

1.3 Societal Motivation

Behavioral data utilization for personal profiling is worth decent societal awareness in general. The emerging trend of companies utilizing user data for commercial purposes recently caused the society to take various measures, in order to regulate the usage of personal data, e.g., the GDPR [1]. The exploration of recent-most computational possibilities of personal profiling is not only important because of its practical applications, but also to identify of personal information encoded in human behavior.

1.4 Research questions

Eventually, the research questions addressed are the following:

1. How well can a naïve Bayes classifier predict individuals' gender, age and education level based on their lexical decision recognition times?
2. How does the performance from the naïve Bayes classifier compare to the performance found in prior work using correlation-based classification?

1.5 Findings

The proposed model showed remarkable performance when evaluated with training data, outperforming the Pearson correlation based model, and exceeding classification accuracy of 95%.

However, large gap between unseen and training data was discovered in terms of classification accuracy. When evaluated on unseen data, the accuracy did not exceed the majority baseline by more than 3%. These values are comparable to the results of the prior work, thus do not improve the currently known gender and educational level prediction capabilities based on lexical decision data.

1.6 Agenda

In the upcoming section, the topic will be framed into the broader context of research regarding demographic profiling based on behavioral and chronometric data, including areas such as continuous authentication or customer profiling, and narrower context composed mainly of psycholinguistic papers. In the method section, we take a closer look at the dataset and introduce the naïve Bayes classifier in conjunction with a kernel density estimation method. Afterwards, the results are presented in tabular form along with corresponding performance of the prior work approach. Lastly, the results are discussed and framed in the wider debate.

2. Related Work

The research related to the present work can be divided in two categories: papers studying lexical processing by means of experimental psychology and papers using behavioral data to train a machine learning classifier for demographics prediction. There is a large variety of papers building personal characteristic classifiers based on human

1 General Data Protection Regulation: https://gdpr-info.eu/

behavior, however, the data are not always chronometric and not always related to language (such as lexical decision data). In fact, the research intersecting the question of classifier development and lexical decision data usage is so scarce that there is only one prior paper addressing this very issue, discussed in the section 2.4.

2.1 Customer profiling efforts

Recommender systems based on user's behavior started gaining popularity in the late 1990's (Schafer, Konstan, and Riedl 1999; Whittle and Foster 1989). Since then, customer profiling has become a hot research topic, commonly dealt with in the data mining literature. An area that can profit from advances in customer profiling is e-commerce (Wiedmann, Buxel, and Walsh 2002), due to the nature of behavioral data on the internet - they are easier to collect than information directly declared by the user. With increasing societal awareness regarding privacy issues in the online sphere, reluctance to explicitly providing demographic information by the user is to be expected.

Several studies have attempted to predict demographics of online users in a supervised manner, based on various sources of behavioral data, e.g., visited websites (Hu et al. 2007), opened mobile apps (Malmi and Weber 2016; Zhong et al. 2013), visited places (Zhong et al. 2015), Twitter Follows (Culotta, Kumar, and Cutler 2015), purchased products (Wang et al. 2016), Facebook Likes (Kosinski, Stillwell, and Graepel 2013), etc. Despite having a similar prediction target (e.g., age, gender), all of the aforementioned studies vary in the nature of behavioral data used. In none of the studies the discriminative essence of the training data was assumed to be the timing information itself, which is how they mainly differ from the current work.

2.2 Authentication and biometrics

Personal characteristics prediction is a relevant problem in the field of user authentication and continuous authentication. Continuous authentication is an ongoing process during an active user session with the goal to continuously assess the user's authenticity (Niinuma, Park, and Jain 2010). User experience and an application's security are factors that often balance at the expense of each other. In order to minimize such a trade-off, innovative methods are increasingly researched and implemented into applications, e.g., biometrics. A behavioral biometric usually provides an especially unobtrusive way to get some information about a user, which is why it is suitable for many situations in continuous authentication.

Keystroke timing is an example of behavioral biometric that has been researched for decades (Monrose and Rubin 1997; Obaidat and Sadoun 1996; Joyce and Gupta 1990). It has shown a promising performance in authentication applications. Later on, researchers also inspected features derived from keystroke timing and their predictive power of a typist's gender (Tsimperidis, Arampatzis, and Karakos 2018) using entropy and information gain based feature ranking. It was found that it is possible to reduce a typically high dimensional keystroke dynamics feature-set by roughly a factor of 10 without a major sacrifice in classification performance; the study reports gender prediction accuracy of more than 95%. Another study compared touchscreen-obtained keystroke dynamics and swipe based gender prediction, achieving only a slightly better than chance classification accuracy, suggesting more difficulties linked to using data from mobile devices (Antal and Nemes 2016).

Another data modality containing timing information and used for biometric purposes is the human gait. Such data can be acquired by a smartphone equipped with an

inertial-based accelerometer worn in a person's pocket. The researchers demonstrated limited, but promising authentication capabilities (Derawi et al. 2010; Sprager and Juric 2015). Attempts to predict personal characteristics with human gait are also common. Inertial sensor based statistical features extracted from the time-domain of human gait showed predictive power for both gender and age (Khabir et al. 2019). A more atomic behavior was also investigated, specifically, a single step (Riaz et al. 2015). The researchers managed to predict gender, age and height of the participants with accuracy ranging from 80% to 90%.

The data used in these papers often have the character of time-series. Although the complexity of the commonly used data is higher than a lexical decision, the results do suggest that timing information about behavior can be predictive of personal characteristics.

2.3 Findings of psycholinguistics

Multiple aspects of the word frequency effect have been investigated in the recent years. The strength of the effect has been shown to differ across a number of commonly used corpora, e.g., Celex, SUBTLEX-DE, Google Books, posing a limitation to the corpus based (objective) word frequency measures in explaining the lexical decision performance (Brysbaert et al. 2011). Consequently, an attempt has been made to prevent the problem by collecting subjective ratings of word familiarity across differently experienced readers (Kuperman and Van Dyke 2013). Later on, the measure of word prevalence was introduced as an alternative to the word frequency (Brysbaert et al. 2016). The measure was described as a proportion of people that know the word, and shown to explain additional variance to the word frequency. Numerous word characteristics have been further accounted for affecting the lexical decision performance: word length (Ferrand 2011), similarity to other words (Yarkoni, Balota, and Yap 2008) or age of acquisition (when the word was learned) (Kuperman, Stadthagen-Gonzalez, and Brysbaert 2012), etc.

Several factors with regard to language exposure can be accounted for modulating the lexical decision performance. Age, education and multilingualism were identified as important factors influencing the vocabulary size (Keuleers et al. 2015). An individual's increasing vocabulary size can be associated with faster recognition times (Kuperman and Van Dyke 2013; Yap et al. 2012). The educational level, vocabulary size and age relate to a common denominator: the language exposure. It it justifiable to assume that individuals with higher education are exposed to more reading materials throughout the life, or older people are exposed to language for a longer time, which impacts their vocabulary. Additionally, the participant's age alone has been shown to have an effect on both RTs and accuracy: older people usually take more time to make the decision, but they are more accurate (Ratcliff et al. 2004).

Differences in language use among males and females have been previously investigated as well. A study asking whether women are more talkative than men did not manage to prove the hypothesis by conducting an experiment requiring the participants to wear voice recorders in daily life (Mehl et al. 2007). However, a recent study, which builds word prevalence norms, claims that, depending on interests, some words are more common among one gender over the other, e.g., games and weapons for males, flowers for females (Brysbaert et al. 2018).

The previously mentioned psycholinguistic papers are related to the present work in terms of demonstrating a relationship between lexical decision performance and some of the participants' characteristics. The exploitation of the relationship for classifi-

cation purposes is the major value that the present work attempts to provide. The efforts to develop a classifier for one or more of the individual's characteristics using lexical decision data are rare and thus create a research gap that the current work attempts to fill.

2.4 Lexical decision based gender and educational level classifier

A master's thesis written at the Tilburg University (Qin 2018), which is the most related paper by a big margin, asks, if gender and educational level can be predicted using lexical decision performance (accuracy and RT) data. The method used is a Pearson-correlation based classifier. The classifier is built by computing a reference vector for each target feature value (e.g., male, female, bachelor, etc.), consisting of word-wise average RTs within this subset, and subsequently determining the class with maximum Pearson correlation to the current set of RTs. For gender prediction, the accuracy using RTs and response accuracies were approximately 3% and 5% (correspondingly) higher than the majority baseline. The classifier for educational level of four classes achieved accuracy roughly 6% and 7% below the majority baseline using RTs and response accuracy correspondingly. The main difference to the current work is in the classification algorithm; a minor difference is that the current work also evaluates a classifier for age group in addition to gender and educational level prediction. However, the research questions are tightly bound to the current work. To the best of our knowledge, this is the only prior work attempting to build a machine learning model using lexical decision data for personal characteristics classification.

3. Method

3.1 Dataset

The lexical decision data used to develop the classifier originate in a large-scale megastudy: the English Crowdsourcing Project (ECP) (Mandera, Keuleers, and Brysbaert 2019). The goal in a megastudy is commonly the analysis of a continuous variable, which is why large amounts of data are gathered. The ECP gathers the data online (rather than in a laboratory) using a public web-page with an embedded vocabulary test. Thanks to the accessibility of such an approach the dataset has grown to an unprecedented size, allowing for novel research, e.g., application of machine learning methods. Not only does the ECP contain a large number of trials, but it contains a large enough number of participants for profiling purposes. This is a deciding attribute for dataset choice regarding the current research questions.

The main entities in the dataset are participants, sessions and trials. A session refers to one run of the vocabulary test and consists of 100 trials. One trial corresponds to a response to one lexical stimuli. A participant can perform multiple sessions.

The publicly available dataset underwent a comprehensive filtering pipeline in order to eliminate undue influences, e.g., not more than three sessions are included from a single IP address. Additionally, the dataset is limited to native English participants only. Eventually, the available dataset consists of roughly 700 thousand sessions and includes reactions to more than 62 thousand English words.

The participants' were asked to indicate whether they know the presented word or not, and to press the 'j' or 'f' key correspondingly. Within each session, the subject responded to 70 words and 30 non-words in an arbitrary order. The participants were motivated by testing their vocabulary size. They were informed ahead, that a positive

response to a non-word results in a heavy penalization. The non-words, which follow the phonotactic rules of English language, were selected from a list of pseudowords generated by the Wuggy (Keuleers and Brysbaert 2010).

Besides completing the vocabulary test, the participants filled a short questionnaire. Questions were asked regarding the following information: where they grew up, highest obtained degree (or worked towards), gender, age, number of languages they spoke apart from English and their mother tongue, and which one of these they command the best. The fact that this dataset contains information about the participants' gender, education and age enables us to tackle the proposed research questions.

The resulting dataset consists of rows corresponding to a single lexical decision and relevant columns for the research. Every row contains the following information: session identifier, spelling of the presented stimuli, lexicality of the stimuli (binary for word/non-word), accuracy of the decision (binary; 0 for non-words identified as known or words identified as not known, 1 otherwise), recognition time, z-score of the recognition time according to the current session, participant's gender, age and educational level. Due to computational limitations, a subset of the dataset has been used: the first 10 million lexical decisions; 100 thousand test sessions. The average number of RTs per word is proportional to the dataset size. Given the subset used, the average number of RT observations per word is 77.

Lastly, the relevance of using the z-scored RTs instead of plain RTs is worth pointing out. Within every session, calculating the z-score of the RT ensures it is mean-centered and the standard deviation is 1. Therefore, the z-scores of RTs are freed from participants' individual biases and variances in lexical decision performance. The pattern has been previously justified for investigating effects of word prevalence (Brysbaert et al. 2018). Thus any further mentioning of a RT in the context of the classification algorithm refers to the z-scored variant.

3.2 Data Filtering

A very trivial and obvious filtering step is to remove sessions missing the information about the currently predicted personal characteristic on a per-model base. Next, two kinds of lexical decisions are not informative for the proposed classifier: trials with non-words, and inaccurate decisions. The following paragraphs argue, why these should be excluded from the training data.

Firstly, there is a technical argument for excluding trials with non-words. During the classifier's learning phase, we need to estimate the probability density functions (PDFs) of RTs given a specific word and a specific demographic feature value, e.g., female (gender). The amount of distinct non-words is very large compared to the amount of distinct words, and they represent a minority of trials in every session (30 non-words compared to 70 words). Therefore, the sets of RTs for estimating the PDFs for non-words would be dramatically smaller and less representative than the ones for words. Additionally, including trials with non-words would not be reasonable with regards to the research question. The research objective is to demonstrate how well do RTs in a lexical decision predict personal characteristics, because of the assumption that they relate to the person's exposure to the words. People are not exposed to the non-words, so it is not reasonable to include them.

Secondly, the trials are limited to accurate ones only. Information about inaccurate lexical decisions does not connect to our research question, because we are seeking personal information solely in the chronometric component of the data.

Table 1
Target label distributions in the dataset (comprised of first 10 million trials - 100 thousand experimental sessions). In case of the 3 age-related target variables, the values correspond to the intervals resulting from the quantile-based split. The frequencies correspond to the values with the same order.

Variable	*Values*	*Frequencies [%]*
Age (Binary)	(0, 32], (32, 100]	50.1, 49.9
Age (Ternary)	(0, 27], (27, 39], (39, 100]	29, 37, 34
Age (Quaternary)	(0, 24], (24, 32], (32, 44], (44, 100]	22.6, 27.4, 25.3, 24.7
Gender	Male, Female	49.2, 50.8
Education level	HS, BC, MA, PH	21.2, 47.4, 21.7, 9

3.3 Model Design

The chosen approach to answer the research question is the evaluation of a machine learning classifier trained on lexical decision data. The problem we are facing can be stated as binary or multiclass classification, depending on the predicted personal characteristic. A separate model has to be trained for each personal characteristic, but all models follow the same method.

The model for age-group prediction requires to additionally discretize the age variable; we worked with division into two, three and four bins. Meaningful bin-thresholds have been determined by computing the quantiles of the age variable according to the desired number of bins. Originally, the education attribute had five levels: primary school, high school, bachelor, master and PhD. However, due to it's low representation (563 participants), the primary school class will not be further considered. All of the resulting target variables, values, discretization intervals and corresponding frequencies are depicted in the table 3.3.

Every machine learning model needs a baseline for a relevant performance comparison. The most trivial baseline we consider is the chance accuracy. Chance accuracy is determined as $1/K$, where K is the amount of possible classes. Secondly, the majority baseline is a commonly used method. It is computed as n_k/N, where n_k refers to the amount of data points with the most common label and N refers to the total amount of data points. Lastly, we compare the model's results to the results obtained by the Pearson correlation based classifier proposed in prior work (Qin 2018).

3.4 Algorithm

The proposed algorithm is a Bayesian probabilistic classifier. Bayesian probability is the interpretation of the probability concept as *reasonable expectation* (Cox 1946). The Bayesian theorem relates the conditional probabilities of two events as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In other words:

$$posterior = \frac{likelihood * prior}{evidence}$$

Given the events A and B, the *prior* is the initial degree of belief for A; the *posterior* is the degree of belief for A, having accounted for B. The *likelihood* is the probability of B occurring given A is true; the *evidence* is the probability of B.

Suppose, we want to classify an instance of one of the C possible classes, described by a set of features $X = x_1, ..., x_n$. The naïve Bayes is a method of modelling such a problem with the Bayes theorem, thus in terms of conditional probabilities. The simplest form of naïve Bayes works by computing the conditional probabilities $P(C_k|X)$ for all the possible classes C. The class with the highest computed conditional probability is the one chosen as a result. The interpretation in the current context is the following: the personal characteristic value with the highest probability, given a set of RTs, will be chosen.

Although the classifier makes a decision based on probabilities, the actual values of those are usually not very informative. As is clear from the formula of naïve Bayes, the final value of the conditional probability heavily depends on a product of n probabilities. It follows, that the more features there are (or the larger the value of n), the smaller the final value tends to get. Therefore, the value is usually only used for decision making and not interpreted further.

There is a small caveat, however: what is the likelihood $P(C_k|X)$, if X is a *set* of features? This is where the approach's naïvness is explained: it lies within the independence assumption between the individual feature occurrences x_i. Assuming the independence $P(x_i|x_{i+1}, ..., x_n, C_k) = P(x_i|C_k)$, the method can be dramatically simplified and modelled by the following formula:

$$P(C_k|X) = P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

The training process of the classifier consists of computing the prior probabilities for every class and the likelihood for every combination of class and feature. The former is a simple computation of K probabilities, but the latter is more exhaustive: it requires an amount of computations that is proportional to the number of classes and the number of features.

In case the features x_i are categorical, $p(x_i|C_k)$ can be easily computed as the ratio between the number of instances of class C_k having feature x_i and amount of instances of class C_k in total. However, in case x_i is an observation on a continuous scale, things get more complicated. In our case, the features correspond to n words of one test session, and feature values are the RTs. In order to determine $p(x_i|C_k)$, we first need to determine the probability distribution of RTs given the word i and one of the classes C_k, which is the final missing piece to the method.

A rather trivial way of obtaining a PDF of a continuous variable is a histogram (discretization of the variable into a finite number of bins). However, we chose to use the kernel density estimation (KDE), which is a more sophisticated way to estimate the distribution. The KDE works by superimposing a set of probability kernels corresponding to the positions of a finite set of observations, and subsequently normalizing the resulting distribution so that the PDF integrates to one. The KDE has hyperparameters, however: the choice of the kernel type and kernel bandwidth. Usually, there is no reason to not choosing the Gaussian kernel, which is also the one we are working with. The bandwidth is a numerical parameter having a fairly strong influence on the resulting shape of the distribution. The smaller the bandwidth, the "spikier"; the higher the bandwidth, the smoother the final probability distribution. It is a way to over-

, and underfit the estimation. However, there are heuristics to determine an optimal bandwidth value, such as the Scott's rule (Scott 1992). Therefore, the bandwidth will be determined as the Scott's factor:

$$SF = n^{-1/(d+4)}$$

where n is the amount of observations the estimate is based on and d the number of dimensions (1 in our case).

The quantification process of $P(x_i|C_k)$ deserves some attention, because normally it is not equivalent to evaluating the PDF at a point x. Any continuous PDF $g(x)$ has 2 important properties. Firstly, the area under the curve sums to 1: $\int_{-\infty}^{\infty} g(x)dx = 1$. Secondly, the probability of the occurrence of any real numbered value is equal to 0. It follows, that the process of probability evaluation based on a continuous PDF has to be an integration between two locations a and b. Therefore, the evaluation of the probability of $p(X = x)$ with a continuous probability density function $PDF(x)$ is defined by the following integral:

$$p(x \leq X \leq x + d) = \int_{x}^{x+d} PDF(x)dx$$

where d is a constant determining the size of the interval to integrate. However, if d becomes very small, given the definition of a derivative, we can come to the expression

$$p(X = x) \approx PDF(x) \times d$$

implying the conditional probability $p(x_i|C_k)$ can be substituted by $PDF_{k,i}(x_i)$. This follows from the fact, that the multiplication by d in naïve Bayes also appears in the denominator ($evidence$ is also based on an evaluation of a continuous probability density function), thus eventually cancels out with multiplications by d within the $likelihood$ (a product of conditional probabilities $p(x_i|C_k)$) (John and Langley 1995).

3.5 Evaluation

The resulting dataset of 100,000 experimental sessions has been divided into a training and test set in a $90:10$ ratio, preserving the target label distributions in both sets. Random sample of 10,000 sessions from the training set has been created for evaluation of the performance on the training data. For the sake of model evaluation, both the training and test data have been filtered in such way, that the target label is present in all session instances. Therefore, in some situations the support for the evaluation is equal to a little less than 10,000.

Given the current research question, the most relevant evaluation metric is *accuracy*: the ratio between the amount of correctly classified instances and all of the instances. There are several reasons for this. Firstly, we are not coping with a dramatic class imbalance that would require paying more attention to one class over another. Secondly, there is no particularly crucial importance of minimizing either false positives or false negatives, as particular problems pose, e.g., disease diagnosis, public threat detection, etc. Therefore, recall and precision are equally important for the sake of the current work. In case the approach would have proven to be applicable in practical situations,

Table 2
Summary of the results in tabular form. The table is divided in classification problem type and it's baselines, evaluation data type (train vs. test), and approach (PC for Pearson Correlation, NB for naïve Bayes). Results are reported as accuracy per cent.

				Training acc.		*Test acc.*	
Target	*No. of classes*	*Chance BL*	*Majority BL*	*PC*	*NB*	*PC*	*NB*
Age	2	50%	50.1%	79.2%	95.3%	55.3%	52%
Age	3	33%	37%	72.1%	95.4%	38.5%	37.7%
Age	4	25%	27.4%	71%	96.2%	28.4%	27.9%
Gender	2	50%	50.8%	77.9%	97.2%	53.5%	50.5%
Education	4	25%	47.4%	70%	97.7%	31.8%	39.1%

eventual biases would have to be further investigated, in order to ensure algorithmic fairness accordingly.

3.6 Software

The code was written in Python. The Scipy 1.3.1 package was used for computing the PDFs, by doing the kernel density estimation with $scipy.stats.gaussian\_kde$.

4. Results

The results presented in Table 4 can be discussed from various points of view. Firstly, there is a distinction between the classification problem type (predicted target feature), corresponding to table rows. The age group occupies three rows, because it was discretized in three ways. The column following the target variable refers to the number of possible classes of the prediction target. This number implies the value of the chance baseline in the next column. The majority baseline values are in the next two columns. Next, the actual results are divided in training data, test data, NB (naïve Bayes) and PC (Pearson correlation), contained in the last 4 columns. These cells report the performance by means of accuracy per cent.

More light can be shed upon the bare values in terms of average accuracy across groups. The average accuracy achieved on test data (41.5%) was more than twice as low as the average accuracy achieved with an equally sized random sample of training data (85.2%). Overall, the mean accuracy achieved with the naïve Bayes and Pearson correlation based classifier was 68.9% and 57.78% respectively. When it comes to predictive abilities (i.e, the test accuracy), the methods are on par in terms of average values, i.e., 41.5% achieved with the correlation, 41.4% achieved with the naïve Bayes. Evaluation with training data resulted in an average accuracy of 74.04% for Pearson correlation and 96.36% for naïve Bayes.

In the scope of the training data performance, the Bayesian approach outperforms the Pearson correlation at all prediction tasks. On the other hand, when evaluated on unseen data, the Pearson correlation based approach marginally outperforms naïve Bayes at age group and gender prediction. The educational level was better predicted with naïve Bayes in all situations.

The chance baseline was outperformed under all circumstances. The majority baseline was not outperformed at educational level prediction and naïve Bayes based gender

prediction (the cases with unseen data). In all the other prediction tasks it was outperformed marginally, although the Pearson correlation based approach dominated more.

In case of the naïve Bayes approach evaluated with training data, the classification accuracy did not suffer from an increasing amount of possible target labels. In fact, the highest accuracy was achieved with education level having 4 possible classes. Contrarily, the Pearson correlation based model was the most accurate with gender classification.

5. Discussion

Disappointingly, the model performs rather poorly on unseen data. The naïve Bayes based approach outperforms the Pearson correlation based approach only at education level prediction. Both approaches predict education with an accuracy below the majority baseline. Gender and age prediction models do outperform the majority baseline, however, not by a very high margin. Moreover, the performance of the naïve Bayes classifier is comparable to the Pearson correlation based classifier (with unseen data). These figures suggest, that neither one of the methods predict the participants' personal characteristics reliably.

A comparison of the results achieved with training set and test set might be more insightful. Interestingly, the naïve Bayes based approach achieves relatively high accuracy on the training set. Accuracy values of magnitude above 95% were providing us with a good dose of hope at the the beginning of the research, suggesting a solid ability to model the underlying process. However, a little later it turned out that the model does not generalize well, as the performance with unseen data was dramatically worse.

Normally, a dramatic difference in training and test performance suggests model overfitting. This term refers to a common type of modelling error in machine learning, that occurs when the model learns beyond the patterns in the data relevant for the classification (specific details present in the training set only) in a naïve pursuit of higher training accuracy (Hawkins 2004). A typical case of overfitting is choosing a too small k in a kNN classifier, or fitting a polynomial of too high degree in a linear regression model. Although some classifiers are sensitive to overfitting, the naïve Bayes is not a particularly prone one.

The naïve Bayes with discrete probability distributions does not require any hyperparameters selection. However, the variant used in the present work uses kernel density estimation, bringing some hyperparameters to the table: the kernel shape and bandwidth. It is no secret, that reaction times in psycholinguistic studies tend to form an ex-Gaussian distribution (Balota and Spieler 1999), which is typical for cognitively demanding tasks. However, using a Gaussian kernel for the KDE is justifiable in this case, because the goal is to model the observation's error, rather than the cognitive process itself. On the other hand, the bandwidth selection might play a more important role, because it determines the smoothness/spikiness of the resulting distribution. In the present work, the Scott's rule has been used (Scott 1992), which depends on the amount of points the estimation is based on. We tried to both up-, and downscale the Scott's value, which did not lead to a significantly increased prediction accuracy.

The figures do not particularly favor the initial ideas of the research. It is clear, that there are statistical differences between RTs of certain groups of people. However, the results of the present work suggest, that the footprint of personal characteristics in the form of recognition times in a lexical decision is not sufficiently discriminative for gender, age group and educational level prediction purposes.

However, there is a limiting factor to the previous statement, that also points to a potentially future work: the dataset the findings are based on contains lexical decisions on tens of thousands of non-uniformly distributed word stimuli. The classification model might benefit from some precautions taken when choosing or building a suitable dataset. Firstly, using a smaller set of words would increase the amount of RT samples per word, which is beneficial for the probability density estimation. Secondly, a tailored selection of the stimuli for classification purposes may improve the classification accuracy, given there are enough observations per stimulus.

Nonetheless, researchers from various domains have shown that chronometric behavioral data can encode a solid amount of personal information, e.g., keystroke dynamics and human gait. The modality of keystroke timing data encodes personal information well enough even for identification purposes. In contrast to that, the recognition times in a lexical decision may be lacking the necessary complexity for this kind of classification problem.

From an ethical and societal point of view, this research does not imply any additional privacy concerns as a result of progressions in machine learning and it's applications. From a practical point of view, the research does not imply possible innovations in fields, where behavioral data can be well exploited, such as artificial agents or robotics.

6. Conclusion

We successfully designed, implemented and evaluated a probabilistic classifier using lexical decision data, an instance of behavioral chronometric data. Evaluation of the proposed model along with a Pearson correlation based model from prior work lead to surprising results. A significant performance gap has been revealed between training and test data based evaluation. This was not only the case for the naïve Bayes based solution, but the Pearson correlation based model as well. This observation was not communicated by the prior work.

Nonetheless, the proposed model classified unseen data with a very low accuracy (mostly <3% above the majority baseline), suggesting that the lexical decision data are not enough to reliably predict an individual's gender, age group and educational level. Furthermore, the prior work's performance was not outperformed. On a positive note, the finding implies no additional concerns regarding personal privacy issues.

However, personal profiling based on behavioral data is a fairly wide field, as the section about related work depicts. Various data modalities, e.g., keystroke dynamics, have been shown to be predictive of personal characteristics and usable in machine learning applications. A core data-related difference to the present work is in the atomicity of the lexical decision data compared to the more complex keystroke dynamics data. Therefore, we conclude that machine learning methods fueled with chronometric data generated by inner processes of the human brain can be useful, however, a certain level of data complexity may be necessary to achieve reasonable results. The lexical decision data have not been proven to have this attribute yet.

References

Antal, Margit and Gyozo Nemes. 2016. Gender recognition from mobile biometric data. In *SACI 2016 - 11th IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, pages 243–248, Institute of Electrical and Electronics Engineers Inc.

Balota, David A and Daniel H Spieler. 1999. Word Frequency, Repetition, and Lexicality Effects in Word Recognition Tasks: Beyond Measures of Central Tendency. *Journal of Experimental Psychology: General*, 128(1):32–55.

Brysbaert, Marc, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58(5):412–424.

Brysbaert, Marc, Paweł Mandera, Samantha F. McCormick, and Emmanuel Keuleers. 2018. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2):467–479.

Brysbaert, Marc, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. The impact of word prevalence on lexical decision times: Evidence from the Dutch lexicon project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):441–458.

Cox, R. T. 1946. Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14(1):1–13.

Culotta, Aron, Nirmal Ravi Kumar, and Jennifer Cutler. 2015. Predicting the demographics of twitter users from website traffic data. In *AAAI*, pages 72–78.

Derawi, Mohammad O., Claudia Nickely, Patrick Bours, and Christoph Busch. 2010. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In *Proceedings - 2010 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2010*, pages 306–311.

Ferrand, Ludovic. 2011. Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from Chronolex. *Frontiers in Psychology*, 2.

Hawkins, Douglas M. 2004. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12.

Hu, Jian, Hua Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. In *16th International World Wide Web Conference, WWW2007*, pages 151–160.

John, George H and Pat W Langley. 1995. Estimating continuous distributions in Bayesian classifiers. In *Eleventh conference on Uncertainty in artificial intelligence*, Stanford University, Morgan Kaufmann Publishers Inc.

Joyce, Rick and Gopal Gupta. 1990. Identity authentication based on keystroke latencies. *Communications of the ACM*, 33(2):168–176.

Keuleers, Emmanuel and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633.

Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2012. The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44(1):287–304.

Keuleers, Emmanuel, Michaël Stevens, Paweł Mandera Mandera, and Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental psychology*, 68(July 2015):1665–1682.

Khabir, Kanij Mehtanin, Md Sadman Siraj, Masud Ahmed, and Mosabber Uddin Ahmed. 2019. Prediction of gender and age from inertial sensor-based gait dataset. In *2019 Joint 8th International Conference on Informatics, Electronics and Vision, ICIEV 2019 and 3rd International Conference on Imaging, Vision and Pattern Recognition, icIVPR 2019 with International Conference on Activity and Behavior Computing, ABC 2019*, pages 371–376, Institute of Electrical and Electronics Engineers Inc.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15):5802–5805.

Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Kuperman, Victor and Julie A. Van Dyke. 2013. Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3):802–823.

Malmi, Eric and Ingmar Weber. 2016. You are what apps you use: Demographic prediction based on user's apps. In *Tenth International AAAI Conference on Web and Social Media*.

Mandera, Paweł, Emmanuel Keuleers, and Marc Brysbaert. 2019. Recognition times for 62 thousand English words: Data from the English Crowdsourcing Project. *Behavior Research Methods*.

Mehl, Matthias R., Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher, and James W. Pennebaker. 2007. Are women really more talkative than men?

Meyer, David E and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.

Monrose, Fabian and Aviel Rubin. 1997. Authentication via keystroke dynamics. In *Proceedings of the 4th ACM conference on Computer and communications security*, pages 48–56, Citeseer.

Monsell, S., M. C. Doyle, and P. N. Haggard. 1989. Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118(1):43–71.

Nguyen, Tien T., Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW âĂŹ14, page 677âĂŞ686, Association for Computing Machinery, New York, NY, USA.

Niinuma, Koichiro, Unsang Park, and Anil K. Jain. 2010. Soft biometric traits for continuous user authentication. *IEEE Transactions on Information Forensics and Security*, 5(4):771–780.

Obaidat, MS and B Sadoun. 1996. Keystroke dynamics based authentication. In *Biometrics*. Springer, pages 213–229.

Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Qin, Xue. 2018. Can Participants' Gender and Educational Level be Predicted from Their Lexical Decision Task Performance? Master's thesis, Tilburg University.

Ratcliff, Roger, Pablo Gomez, Anjali Thapar, and Gail McKoon. 2004. A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, 19(2):278–289.

Riaz, Qaiser, Anna Vögele, Björn Krüger, and Andreas Weber. 2015. One Small Step for a Man: Estimation of Gender, Age and Height from Recordings of One Step by a Single Inertial Sensor. *Sensors*, 15(12):31999–32019.

Schafer, J Ben, Joseph Konstan, and John Riedl. 1999. Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166, ACM.

Scott, David W. 1992. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sprager, Sebastijan and Matjaz Juric. 2015. Inertial Sensor-Based Gait Recognition: A Review. *Sensors*, 15(9):22089–22127.

Tsimperidis, Ioannis, Avi Arampatzis, and Alexandros Karakos. 2018. Keystroke dynamics features for gender recognition. *Digital Investigation*, 24:4–10.

Wang, Pengfei, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Your cart tells you: Inferring demographic attributes from purchase data. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 173–182, Association for Computing Machinery, Inc.

Whittle, Susan and Morris Foster. 1989. Customer profiling: getting into your customerâĂŹs shoes. *Management Decision*, 27(6).

Wiedmann, K-P, H Buxel, and G Walsh. 2002. Customer profiling in e-commerce: Methodological aspects and challenges. *Journal of Database Marketing & Customer Strategy Management*, 9(2):170–184.

Yap, Melvin J., David A. Balota, Daragh E. Sibley, and Roger Ratcliff. 2012. Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1):53–79.

Yarkoni, Tal, David Balota, and Melvin Yap. 2008. Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, 15(5):971–979.

Zhong, Erheng, Ben Tan, Kaixiang Mo, and Qiang Yang. 2013. User demographics prediction based on mobile data. In *Pervasive and Mobile Computing*, volume 9, pages 823–837, Elsevier B.V.

Zhong, Yuan, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You are where you go: Inferring demographic attributes from location check-ins. In *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 295–304, Association for Computing Machinery, Inc.