Koltai et al. *EPJ Data Science* (2025) 14:60
https://doi.org/10.1140/epjds/s13688-025-00578-2

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

# Classifying social position with social media behavioral data

Júlia Koltai[1,2*†], Zsófia Rakovics[1,3†], Zoltán Kmetty[4,5], Kata Számel[1,6], Borbála Ungvári[1], Bendegúz Váradi[4] and Ákos Huszár[1,7]

Handling Editor: Kokil Jaidka

*Correspondence:
koltai.julia@tk.hun-ren.hu
[1]MTA–TK Lendület "Momentum" Digital Social Science Research Group for Social Stratification, HUN-REN Centre for Social Sciences, Tóth Kálmán utca 4, Budapest, 1097, Hungary
[2]Department of Social Research Methodology, Faculty of Social Sciences, ELTE Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, 1117, Hungary
Full list of author information is available at the end of the article
†Equal contributors

## Abstract

The main question of our study is how far social position can be predicted solely based on digital behavior. The phenomenon that offline inequalities are reflected in the digital space has been heavily researched since the digital revolution. Nevertheless, there are few data, which both measure social inequalities and digital behavior: scientists either have information on the social status of people, or on their observed digital behavior, but not on both. When analyzing digital behavioral data, however large scale it is, information on the social position of the users is hardly available. In the current paper, we analyze a special dataset collected with a data donation technique, which contains information on both the social position and the observed digital behavior of participants, and which is representative for the internet user population of Hungary. In the analysis, using diverse models, we explored how well basic indicators measuring digital behavior on Facebook can classify users' social class measured by the 5-category version of the European Socio-economic Classification (ESeC). The results show that based on basic quantitative indicators of digital behavior and usage the models cannot classify users' social position with a high degree neither in the classification of social class, nor in the case of socio-economic status. Nevertheless, the inclusion of socio-demographic characteristics as features increased the predictive power of the models, that could differentiate between the lowest and highest social position with a high degree. The models based on purely observed digital behavior could identify those in the lowest social position with the highest performance. Among those features, that played an important role in this classification, usage time, frequency network size and language characteristics (especially the diversity of the used language and punctuation) should be highlighted, while diverse Facebook activities and detected interest categories also played a role. These results are in line with the results of previous studies derived from smaller-scale, non-representative, or self-reported survey-based data on the same topic.

**Keywords:** Observed digital behavior; Social position; Social inequalities; Social media; XGBoost; Classification

## 1 Introduction

Social media has radically transformed social reality over the past 20 years. People spend hours on different social media platforms on a daily basis. The resulting data enabled re-



Springer

searchers to analyze social processes and phenomena, as well as social network theories on several order of magnitude larger scale. Since the appearance of these platforms, an ocean of studies have analyzed X (Twitter), Facebook, or Instagram data – nevertheless, in the latest years, access to these platforms has become limited [1]. However, due to privacy regulations, we know little about the social background of the users whose data is analyzed unless we involve the social media platform itself in the research, which has data on its own users. Accordingly, our knowledge on how different social groups behave in these online spaces is limited. This paper contributes to the understanding of the relationship between social position and inequality and social media usage with a special dataset collected by a data donation technique, which includes detailed information on both the participants' observed behavior on social media and their socio-demographic characteristics. With diverse classifier models we aim to explore how well social position can be classified by different features of digital behavior and interpret the characterization of such features in the classification.

### 1.1 Digital behavior and social inequality

Research on the relationship between digital behavior and social characteristics dates back to the advent of the internet. While early research focused on inequality in access to online spaces and digital tools, the focus shifted towards usage patterns as internet penetration became more widespread. Research on the second- and third level digital inequalities – which targets such ways of usage, like competencies and skills, and also opportunities and risks [2] – suggests that different groups of the society use the digital space differently. Even at the beginning of the century it was proven that socio-economic background of the users heavily affects internet usage [3]. Since then, many studies came to the same conclusion, highlighting the relationship of structural social inequalities (especially socio-demographic and financial ones) with digital skills and type of usage (see e.g., [4–7]). Focusing on social media platforms, various research showed that users with different social characteristics use the platforms differently. According to a meta-analysis based mostly on surveys with self-reported behavior and profile analysis, men and women use social media platforms with different consciousness and for different motivations: women are more likely to activate privacy settings and untag themselves from photos than men [8]. At the same time, a research based on a smaller scale survey showed that women mostly use these sites for keeping up with close friends and relatives, and gather social information, while men are keener for general information seeking [9]. An analysis based on a population-level survey (which measured digital behavior by self-report) found that with respect to education, higher educated people use social media for work and politics more, and for participating in fan clubs less, compared to lower educated ones [10]. Based on the self-reporting of respondents of a representative sample, age affects digital media usage as well, such as younger generations spend more time online, especially with media consumption and social use [11]. A research conducted on a smaller scale, non-representative sample of users with estimation on the users' age, suggests that varying feature usage can be observed between users of different generations, for example, younger users are more likely to write comments and use the reply function, while older ones tend to write directly on someone's wall and post hyper media [12]. According to the same research, comparing the size of social network, younger users have substantially larger network than older ones. Nevertheless, these studies examined social media usage only by main sociodemographic variables and did not analyze it by complex concepts such as social position.

Sociology has long used the concept of social class to operationalize social positions, as it captures multiple dimensions of inequalities and measures the social status of people well. Previous research showed that even within the same socio-demographic groups (e.g. age group), those belonging to different social classes are different in their technology usage [13]. Therefore, using the concept of social class in the understanding of the relationship between digital behavior and social inequalities is important, as it allows for a more nuanced analysis of how digital behaviors are shaped by underlying structural inequalities and varying access to resources. Earlier research showed that people in different social classes also differ in their time of usage and their focus on social media in digital consumption. Class affects self-reported digital media usage with lower classes spending more time online [11] and are more likely to primarily engage with social media platforms [14]. The same research showed that composition of activities in social media platforms differs among classes as well: those belonging to higher classes are more likely to carry out various cultural and personal activities on social media sites, such as participating in offline activities or joining groups. [14]. Regarding interest categories, they have found that lower classes may have less engagement with professional contexts on social media, while higher classes are more interested in culture and sport related discussions and events [14]. Additionally, the ocean of written text on social media provides an efficient source of class differentiation as well. Since Bernstein, we know that different classes use different languages, namely lower working class people tend to use restricted code with limited range of alternatives, while middle class people choose their language use from a more extensive range [15]. His theory has since been confirmed by many, and newer waves of research even detect those particular linguistic elements, which differ across social classes (e.g. [16–18]).

### 1.2 Measurement of digital behavior and social position

Working with observed digital behavioral data, a frequent solution for having information on the social background of the users is the inference of social characteristics (e.g., based on users' metadata, or other, platform related information), frequently with machine learning algorithms [19]. We can find inference of gender, race/ethnicity, age, socioeconomic status, or regional origin of users [20–28], but they lack the ground-truth about the real social characteristics of the users.

Another solution to collect information both about people's social characteristics and their digital behavior is conducting surveys. Still, self-reported surveys can only capture digital behavior to a limited extent, as they are more capable of measuring attitudes than real behavior [29].

Among the studies where researchers have precise information both on some social characteristics of users and their real behavior, we can find analyses based on early social network sites (e.g. [30]), data from times when privacy regulations towards private profiles were not as strict as they currently are (e.g. [31]), or studies, which applied data donation techniques (through web-browser plugins, or Data Download Packages (DDP's) and have used either a limited number of digital behavioral characteristics (e.g. [1] - study 1, [32–34]), or a smaller scale sample (e.g. [35]). Nevertheless, these efforts generally focus only on a very limited number of social characteristics and thus can barely grab the complexity of users' real social position.

The lack of studies on the correlation between social position and online behavior is also coming from a disciplinary boundary. Most studies that combine digital and survey

data are associated with media and communication studies or the field of political science - this approach is relatively rare in sociological research.

Summarizing the previously presented studies targeting the relationship between social characteristics or position and digital behavior are based on smaller scale (e.g. [7]) or non-representative research (e.g. [35]), measure digital behavior with self-reported survey data [8], or only measure a limited number of social characteristics (e.g. [1] - study 1, [32–34]). All of these have limitations regarding their conclusions: smaller scale, non-representative data is hardly generalizable to a larger population, self-declared data on digital behavior is biased compared to real behavior, while limited number of social characteristics does not allow for the measurement of complex social positions. For the solid analysis of the relationship between social position and digital behavior, it is crucial to have generalizable and valid data without self-report bias. Thus, having a representative sample of people with detailed knowledge on their real complex social position as well as their observed digital behavior is needed for such analysis. To the best of our knowledge, no such extensive digital- and survey data have been collected from the same respondents prior to this research.

In this study, the solution we applied is a data donation approach, which has become a promising research direction in the latest years. This process allows researchers to access social media data in a clean, legal environment [18]. When data donation is combined with a survey of the same participants, as is the case in the current research, both digitally observed data and information on the participants' social positions become available. Having information on the social position and the observed digital behavior of the same participants of a representative sample lets us test how well social position can be classified solely based on digital behavior.

### 1.3 Goal and contribution

The goal of this paper is twofold. First, we explore how well the social position of users can be classified purely on basic quantitative data on their digital behavior. This exploration could contribute to existing literature as the data we test the question on is representative of a whole society, based on real behavior, not on self-report, while it also includes (not inferred) data on complex social characteristics. We use multiple model types to address this question, such as Random Forest, XGBoost, and TabNet. For these classifiers, we selected features based on the previously introduced literature, which showed meaningful relationship with social class and position. Therefore, the time of usage, the language that users use in their posts or comments, the size of the social network and the way users manage their contacts, the different types of activities users execute on the platform, and their interest categories were operationalized as indicators, which were then applied as features in the classifiers. The second goal of this paper is to identify those features in more detail, which play the most important role in the classification of different social positions. These analyses could contribute to the understanding of how different aspects of digital behaviors are related to diverse social positions.

## 2 Data and methods

This study uses a donation-based data collection, which collected survey responses and digital behavioral data from the same participants. In a donation-based research design, researchers ask participants to share their data stored by different digital platforms with them.

The implementation of the General Data Protection Regulation (GDPR) in 2016 made data donation more accessible by mandating that data controllers offer individuals electronic access to their data [1, 36, 37]. In response, all major digital platforms now provide users with downloadable "data packages" containing their information. In the context of data donation research, participants are generally recruited using standard survey sampling methods. After being selected, users download their personal data packages, review the accompanying privacy notice and consent form, and then submit their data to the research team. More details on data collection, the invitation forms, and provided instructions are available in the Supporting Information. These data can be downloaded and shared with the researchers by the user with informed consent, and thus, make it available for analysis.

### 2.1 Dataset

The dataset we analyze was collected on a nationally representative sample of 758 people representing the Hungarian 16 years old or older internet user population in terms of gender and age. The data collection period spanned from February 2023 to June 18, 2023. Participants were recruited via the online access panel of an online polling company, NRC. The recruitment process began with an invitation email sent by NRC to each potential participant. This email provided information about the project's objectives, the incentives offered and included a link to the project's dedicated webpage.

Participants who decided to participate in the study had to follow three consecutive steps:

- Initial Screening Questionnaire and Consent: Participants first completed a brief online questionnaire that included eligibility questions and a consent form. This initial screening was designed to exclude those who did not have or regularly use Google and Facebook profiles. Participants were required to read a detailed description of the research and agree to the consent form to proceed further. Without agreeing to the consent form, they could not go forward to the next steps.
- Data Download and Upload: Eligible participants were then directed to a webpage with comprehensive guides for each platform (Google, Facebook, Instagram, Twitter, and TikTok). These guides provided step-by-step instructions for exporting and downloading data from these platforms, supplemented by a video tutorial with the same content. After downloading their data to their computers, participants uploaded the unaltered files to the project's website. Uploading data from Google and Facebook was mandatory, while those who also provided data from Instagram, TikTok, and Twitter received additional incentives. Participants could contact the researchers via a dedicated email address for any technical issues or research related questions. If the uploaded data format was incorrect, researchers could reach out to the participants via email as well.
- Detailed Survey Questionnaire: Following the data upload, participants were required to complete a 30-minute questionnaire addressing various topics. This survey included detailed questions about the participants' labor market position, financial situation, and different socio-demographic characteristics.

The donated datasets are stored in a safe server at the research center. In the data preprocessing phase, all names and identifiers were hashed from the datasets. The parsed, sanitized, and anonymized data were uploaded to a highly integrated SQL database. The

processed dataset is only accessible from a dedicated server with a strict access protocol in place. Only aggregated and fully anonymized data has been deposited in a data repository.

By the end of the data collection period, 758 participants had successfully completed all the steps. To address sampling biases, we applied individual weighting to minimize discrepancies between the population and sample distributions concerning socio-demographic variables. These weights were derived using iterative proportional fitting [38]. Ultimately, the weights adjusted the distributions of all variables to closely match the population distribution (within a tolerable margin of error), where the population is defined as Hungarian internet users aged 16 and older, who uses the internet for communication with chats or emails. The dimensions considered for weighting included gender, age, education, settlement type, and geographic regions. The data donation study was fully complying with the actual European and Hungarian privacy data regulations and was conducted with the IRB approval of the Centre for Social Science Ethical Board (1-FOIG/130-37/2022) and with the informed consent of the participants. The hybrid technique of donation and survey made it possible to both have information on the participants' social position, and their observed behavioral data.

In this paper, from the digital behavioral data available in the dataset, Facebook data will be analyzed. Although in some countries' trends show a decreasing Facebook penetration, especially among the younger generations, Hungary is a country, where 7.43 million people used Facebook in 2022 [39] out of the 8.56 million internet users (which is 89 percent of the total population) [40]. This 86.7 percent penetration of Facebook among internet users does not only mean a high rate, but according to [41], the platform is equally popular among users of various age groups and genders.

### 2.2 Measures

Social class serves as a fundamental variable in this study. Adhering to the tradition of empirical class analysis, we measure it based on individuals' occupations and other labor market characteristics. We used the five-category concept of the European Socio-economic Classification (ESeC) [42] for the measurement of social positions, which includes the following categories: higher-level service class, lower-level service class, intermediate, skilled workers, and unskilled workers. The approach relies on the concept of social class developed by John Goldthorpe and his colleagues [43, 44]. We derived the variable from the survey data based on detailed occupational codes and labor market information. In the case of respondents who were not employed at the time of the survey, the most recent occupation was considered.

As a robustness test, we also tested models, in which we classified the 4-category version of the participants' socio-economic status (SES) instead of the ESeC5. The 4-category SES was created through a principal component analysis by using the education level (measured in years spent in the education system), the household income per capita based on the (OECD2 equivalent) income variable and the occupational prestige. The total variance explained by the first principal component was 62.973.

From the observed digital behavioral Facebook data, based on the previously introduced studies, we created indicators for the main concepts, about which previous research showed to have a relationship with social inequalities or social position. These concepts were *time and frequency of usage*, *language characteristics* of users used in their posts or comments, users' *social network and the way they manage their contacts*, the *different features users use* on the platform, and the *users' interest categories*. In the Supporting

Information, Table S1 presents all indicators used, including their previously mentioned category (type) and detailed description, along with a description on the creation of the interest categories. While creating the indicators, we applied a restriction on the database to reflect on the time-embeddedness of various activities. For the basic models, we filtered for the last 5 years of the available data as we assumed that older behaviors do not necessarily relate to the users' social position measured at the time of the data collection. Additionally, we normalized for the months spent on Facebook by each user in the sample. In more detail, we checked when the registration happened and computed the active months on Facebook within the five-year long period for each user. This number was used as a divisor to normalize the observed digital behavior. In the model used for robustness check, we conducted the same procedure, but instead of the 5-year-long timeframe, we filtered the data only for 2 years before the data collection for the creation of the time-normalized indicators.

### 2.3 Classification models

In our basic model, we classified the 5-category social class with an XGBoost model, with features calculated from the users' behavior in the last 5 years compared to the data collection. Nevertheless, as the performance of the model is in the focus of our first research question, we conducted multiple robustness tests to make sure the results are solid. First, we tested whether the performance of the model changes if the time frame, in which the features are created, is limited to 2 years before the data collection. Second, we ran the models with an alternative measurement of social position, which is based on socio-demographic variables more directly, namely the socio-economic status of the respondent. Third, we repeated all these classifications with multiple models: additionally, to the XGBoost model, we used Random Forest and a deep learning-based transformer model, TabNet model. Finally, to see how the inclusion of socio-demographic characteristics increases model performance, we also added the participants' gender [male/female], age, education level in 3 categories [Primary/Secondary/Tertiary], and type of settlement as features to the existing model. The different models tested for the analyses are presented in Table 1.

In the basic model, for the classification of social position, measured with the 5-category version of the European Socio-economic Classification (ESeC5), we used fine-tuned and ten-fold, stratified cross-validated XGBoost models [45] with 5-years ranged indicators of the main concepts as features. We have applied ten-fold cross-validation in order to handle the problem of over-fitting. As the ESeC5 variable had missing cases (46 participants), we used a smaller dataset, which contained 712 participants, all with sufficient social class values. We fine-tuned the hyperparameters, such as the number of trees, learning rate, and max depth of the trees, using the default parameters of the xgboost v.1.7.7.1. R-package [46] with train-test split validation (80%-20% proportions respectively), using the multi-class logloss as the evaluation metric. As the target 5-class ESeC variable had an unbalanced distribution, we used the Synthetic Minority Oversampling Technique (SMOTE, [47]). We applied the same procedure for the robustness check, when features of the XGBoost models were limited for users' behavior in the last 2 years from the time of the data collection; as well as in the case when socio-demographic characteristics of the users were added as features. Also, a very similar procedure was conducted for the robustness check when an alternative measurement of the classified variable was used in the XGBoost model: instead of ESeC5, 4-category version of the socio-economic status (SES)

**Table 1** Summary of the classifiers conducted for the research. The models are described by three characteristics: classified variable, features, and type of model. Altogether 12 models were executed, taking the first model (marked with bold) as the base model and running all other models with alternative specifications as robustness tests. Alternative specifications include limitation of the time frame the features are derived from, alternative classified variable for the measurement of social position, the extension of features with socio-demographic variables, and the usage of Random Forest and TabNet models additionally to the XGBoost

| Classified variable | **ESeC5** | ESeC5 | ESeC5 | ESeC5 | ESeC5 | ESeC5 |
|---|---|---|---|---|---|---|
| Features | **5-years behavior-based indicators** | 5-years behavior-based indicators | 5-years behavior-based indicators | 2-years behavior-based indicators | 2-years behavior-based indicators | 2-years behavior-based indicators |
| Model | **XGBoost** | Random Forest | TabNet | XGBoost | Random Forest | TabNet |
| Classified variable | ESeC5 | ESeC5 | ESeC5 | SeS4 | SeS4 | SeS4 |
| Features | 5-years behavior-based indicators + socio-demographic characteristics | 5-years behavior-based indicators + socio-demographic characteristics | 5-years behavior-based indicators + socio-demographic characteristics | 5-years behavior-based indicators | 5-years behavior-based indicators | 5-years behavior-based indicators |
| Model | XGBoost | Random Forest | TabNet | XGBoost | Random Forest | TabNet |

was considered. The only exception here was that we did not apply the Synthetic Minority Oversampling Technique as the distribution of the SES was balanced.

The outcomes of these models can reveal the distinct dimensions that are important for one class, but not for others, and provide information on the direction of these effects. To aid the interpretation of our results, we used SHAP values [48] displayed on beeswarm plots and calculated feature importance values as mean absolute SHAP values.

As additional robustness tests, we conducted all previously described analysis with Random Forest and TabNet classifiers. Tree-based models - like the aforementioned XGBoost and Random Forest - are still dominant in the task of tabular data classification. Similarly to the XGBoost model, we used a fine-tuned and ten-fold, stratified cross-validated Random Forest model [49]. TabNet is a novel deep learning architecture that uses sequential attention to select the most important features at each processing step and according to the authors [50, 51], outperforms XGBoost and Random Forest classifiers, and contrary to previous neural networks used for classification tasks, TabNet deploys learnable masks to help feature interpretation. In the case of TabNet, an additional global feature standardization was applied to the inputs in addition to the internal batch normalization of the model. The TabNet models were hyperparameter optimized for the width of the decision layer (n_d), width of the attention embedding (n_a) and the gamma parameter of the model. Additionally, similarly to the XGBoost models, a ten-fold cross-validation was applied. The same train-test split was used as in the XGBoost model. Adam optimizer with entmax mask type and logloss evaluation metric was used during the fine-tuning process. The TabNet model was built in Python using Pytorch, pytorch-tabnet 4.1.0. [50, 51].

## 3 Results

First, we aim to answer our first research question, namely how well the social position of users can be classified purely by basic quantitative data on their digital behavior. In our basic model, where ESeC5 was classified with XGBoost models using digital indicators

derived from the participants' behavior from the last 5 years, the balanced accuracy of the model was 32.7% (see Table 2). Although this performance does not seem that strong, we must consider some circumstances by which we should evaluate this value. First, as a result of Synthetic Minority Oversampling Technique, we had equal group sizes. Therefore, based on the 5-class setup, we expect the accuracy of 20% according to the random chance, and should evaluate the model performances compared to this measure, which, in this case means a 12.7 percentage point increase. A 5-category classification is a more difficult problem than the most commonly used binary classification problem: This is reflected in the complexity of the XGBoost decision boundaries, which are increasingly fragmented and irregular as the number of classes increases [52]. We can see in a number of previous studies using state-of-the-art model architectures for classification that - also depending on the complexity of the data - as the number of classes increases, the accuracy scores can drop substantially (ex. [53]).

Checking the balanced accuracy measures, which reveal information about the classification of each category of ESeC individually, one can observe that the classes are not similarly difficult to classify. The balanced accuracies of classes from top to bottom - starting with the highest class and ending with the lowest one were: 0.458, 0.499, 0.472, 0.518, and 0.545. These results suggest that the model has the weakest performance in the case of the classification of the highest social class. Nevertheless, its best performing task was to classify the lowest category of ESeC, the unskilled workers category, where the balanced accuracy was 54.5%. Comparing this accuracy to the random chances of one category out of the 5 (20%) means a 34.5 percentage points increase in the case of this social class.

### 3.1 Analyses of the alternative models' performances

To check whether the not so strong performance of our basic model was due to the model used, or the selection of the features, we came up with various checks for robustness and complementary analyses as well (as we described in the Data and Methods chapter).

First, we built further models with the same settings, complementing the results of the XGBoost classifier. We built a simple machine learning model - a Random Forest classifier -, and also a deep learning-based transformer model - a TabNet classifier. Based on the comparison of the performances of various classifiers. It turned out that for this feature set (containing qunatitative Facebook digital behavior indicators of the last 5 years compared to the data collection) the machine learning-based XGBoost classifier performed the best. Both Random Forest and TabNet provided weaker performance with 24.9% and 20% overall accuracy - compared to the 32.7% accuracy of the XGBoost model.

As a second type of robustness test, we created a different subset containing only the last 2 years right before the time-period of the data collection process (see Table 3). We created the same indicators as we used in the original modeling but on the data collected from July 1, 2021. In this way, we were able to compare the robustness of our models from the perspective of time: besides the original 5-year-long-period (data from July 1, 2018) we had data for the 2-year-long-period as well. According to the model performances, the classifiers built on the data restricted to the last 2 years were quite similar and mostly over-performed by the classifiers built on the initially used 5-year-long period, containing data from July 1, 2018. These results can be explained both from a theoretical and from a methodological point of view. Based on the theoretical literature, social position and the theory-driven class structures are quite robust over time [54, 55]. From a methodological

**Table 2** Summary of classification performances of 5-category ESeC by digital behavior indicators, regarding the 5-year-long period and the classification model type (Random Forest, XGBoost, and TabNet classifiers). The table contains the general accuracy of each model, along with the class-specific performance metrics: balanced accuracy, F1 score, Negative Predictive Value, Positive Predictive Value, Sensitivity, and Specificity for each ESeC5 class. The ESeC classes (ranging from Class 1 to 5) are as follows: higher-level service class, lower-level service class, intermediate class, class of skilled workers, class of unskilled workers, respectively

| Classification of ESeC5 by digital indicators | | Accuracy | Balanced Accuracy | F1 | Negative Predictive Value | Precision (Positive Predictive Value) | Sensitivity (True Positive Rate) | Specificity (True Negative Rate) |
|---|---|---|---|---|---|---|---|---|
| Digital indicators of last 5 years - Random Forest | Class 1 | 0.249 | 0.366 | 0.260 | 0.586 | 0.189 | 0.415 | 0.318 |
| | Class 2 | | 0.390 | 0.217 | 0.568 | 0.214 | 0.220 | 0.560 |
| | Class 3 | | 0.444 | 0.167 | 0.561 | 0.263 | 0.122 | 0.767 |
| | Class 4 | | 0.464 | 0.346 | 0.578 | 0.350 | 0.341 | 0.587 |
| | Class 5 | | 0.498 | 0.218 | 0.563 | 0.429 | 0.146 | 0.849 |
| Digital indicators of last 5 years – XGBoost | Class 1 | 0.327 | 0.458 | 0.311 | 0.671 | 0.258 | 0.390 | 0.526 |
| | Class 2 | | 0.499 | 0.316 | 0.655 | 0.343 | 0.293 | 0.705 |
| | Class 3 | | 0.472 | 0.250 | 0.644 | 0.290 | 0.220 | 0.725 |
| | Class 4 | | 0.518 | 0.391 | 0.681 | 0.353 | 0.439 | 0.598 |
| | Class 5 | | 0.545 | 0.358 | 0.655 | 0.462 | 0.293 | 0.797 |
| Digital indicators of last 5 years – TabNet | Class 1 | 0.200 | 0.521 | 0.255 | 0.809 | 0.226 | 0.293 | 0.750 |
| | Class 2 | | 0.537 | 0.253 | 0.814 | 0.263 | 0.244 | 0.829 |
| | Class 3 | | 0.527 | 0.253 | 0.811 | 0.239 | 0.268 | 0.787 |
| | Class 4 | | 0.491 | 0.145 | 0.797 | 0.179 | 0.122 | 0.860 |
| | Class 5 | | 0.591 | 0.346 | 0.836 | 0.350 | 0.341 | 0.841 |

**Table 3** Summary of classification performances of 5-category ESeC by digital behavior indicators, regarding the 2-year-long periods and the classification model type (Random Forest, XGBoost, and TabNet classifiers). The table contains the general accuracy of each model, along with the class-specific performance metrics: balanced accuracy, F1 score, Negative Predictive Value, Positive Predictive Value, Sensitivity, and Specificity for each ESeC5 class. The ESeC classes (ranging from Class 1 to 5) are as follows: higher-level service class, lower-level service class, intermediate class, class of skilled workers, class of unskilled workers, respectively

| Classification of ESeC5 by digital indicators | | Accuracy | Balanced Accuracy | F1 | Negative Predictive Value | Precision (Positive Predictive Value) | Sensitivity (True Positive Rate) | Specificity (True Negative Rate) |
|---|---|---|---|---|---|---|---|---|
| Digital indicators of last 2 years – Random Forest | Class 1 | 0.259 | 0.400 | 0.299 | 0.607 | 0.221 | 0.463 | 0.337 |
| | Class 2 | | 0.502 | 0.419 | 0.603 | 0.400 | 0.439 | 0.565 |
| | Class 3 | | 0.390 | 0.198 | 0.577 | 0.200 | 0.195 | 0.584 |
| | Class 4 | | 0.448 | 0.167 | 0.571 | 0.263 | 0.122 | 0.774 |
| | Class 5 | | 0.440 | 0.107 | 0.568 | 0.200 | 0.073 | 0.806 |
| Digital indicators of last 2 years – XGBoost | Class 1 | 0.322 | 0.440 | 0.253 | 0.647 | 0.239 | 0.268 | 0.611 |
| | Class 2 | | 0.516 | 0.319 | 0.647 | 0.393 | 0.268 | 0.764 |
| | Class 3 | | 0.437 | 0.192 | 0.634 | 0.219 | 0.171 | 0.702 |
| | Class 4 | | 0.481 | 0.274 | 0.644 | 0.313 | 0.244 | 0.718 |
| | Class 5 | | 0.576 | 0.500 | 0.736 | 0.403 | 0.659 | 0.494 |
| Digital indicators of last 2 years – TabNet | Class 1 | 0.178 | 0.500 | 0.184 | 0.800 | 0.200 | 0.171 | 0.829 |
| | Class 2 | | 0.595 | 0.358 | 0.841 | 0.315 | 0.415 | 0.774 |
| | Class 3 | | 0.482 | 0.097 | 0.793 | 0.143 | 0.073 | 0.890 |
| | Class 4 | | 0.491 | 0.202 | 0.796 | 0.188 | 0.220 | 0.762 |
| | Class 5 | | 0.570 | 0.318 | 0.829 | 0.298 | 0.341 | 0.799 |

**Table 4** Summary of classification performances of 4-category SES by digital behavior indicators, regarding the 5-year-long period and the classification model type (Random Forest, XGBoost, and TabNet classifiers). The table contains the general accuracy of each model, along with the class-specific performance metrics: balanced accuracy, F1 score, Negative Predictive Value, Positive Predictive Value, Sensitivity, and Specificity for each SeS class. The SeS categories (ranging from Quartile 1 to 4) are as follows: First (lowest) quartile, Second quartile, Third quartile, and Fourth (highest) quartile, respectively

| Classification of the 4-category SeS by digital indicators | | Accuracy | Balanced Accuracy | F1 | Negative Predictive Value | Precision (Positive Predictive Value) | Sensitivity (True Positive Rate) | Specificity (True Negative Rate) |
|---|---|---|---|---|---|---|---|---|
| Digital indicators of last 5 years - Random Forest | Quartile 1 | 0.349 | 0.529 | 0.436 | 0.632 | 0.425 | 0.447 | 0.610 |
| | Quartile 2 | | 0.467 | 0.341 | 0.619 | 0.318 | 0.368 | 0.565 |
| | Quartile 3 | | 0.428 | 0.243 | 0.603 | 0.250 | 0.237 | 0.620 |
| | Quartile 4 | | 0.510 | 0.371 | 0.615 | 0.406 | 0.342 | 0.678 |
| Digital indicators of last 5 years – XGBoost | Quartile 1 | 0.382 | 0.523 | 0.371 | 0.643 | 0.406 | 0.342 | 0.703 |
| | Quartile 2 | | 0.565 | 0.482 | 0.679 | 0.444 | 0.526 | 0.603 |
| | Quartile 3 | | 0.475 | 0.325 | 0.643 | 0.310 | 0.342 | 0.608 |
| | Quartile 4 | | 0.501 | 0.338 | 0.639 | 0.364 | 0.316 | 0.687 |
| Digital indicators of last 5 years – TabNet | Quartile 1 | 0.276 | 0.531 | 0.196 | 0.763 | 0.385 | 0.132 | 0.930 |
| | Quartile 2 | | 0.544 | 0.354 | 0.777 | 0.293 | 0.447 | 0.640 |
| | Quartile 3 | | 0.610 | 0.416 | 0.805 | 0.410 | 0.421 | 0.798 |
| | Quartile 4 | | 0.544 | 0.325 | 0.773 | 0.310 | 0.342 | 0.746 |

point of view – according to our results –, using a greater amount of data (5 years instead of 2 years) improved the performance. Therefore, after the robustness of the time component, we became determined to use the 5-year-long time-period for the later and final analyses in this paper.

Third, as another type of robustness test, we conducted classifications, where the variable to be classified was changed from ESeC5 to another measure of social position connecting more strongly to the socio-demographic characteristics of the participants. Accordingly, we built a classifier for a 4-category socio-economic status (4-category SES).

For this analysis, we used the original time-constraints (5 years prior to the data collection) and the same model types (Random Forest, XGBoost, TabNet). The model performances are detailed in Table 4.

Considering the accuracy of each model, the XGBoost classifier is proved to be again the best performer, with the accuracy of 38.2%. Compared to the random chance-based classification, in case of which the expected accuracy would be 25%, there is a 13.2% increase when using the digital behavior indicators to classify the 4 categories of SES. This increase is very similar to the increase of 12.7% that we could observe in the case of ESeC5 classification using the same model and same settings. Therefore, the performance of this classifier with the alternative measurement of the social position shows similar results compared to the basic model.

As a last complementary analysis, we extended the list of features of the basic model and involved socio-demographic characteristics into the set of features as well. The variables – gender, age, education level, type of settlement – were included to examine the potential change in the model performance compared to the initial setup. Therefore, the classification problem was the same, we predicted the outcome of the five-category ESeC variable, and applied the methodology as described above, using a hyperparameter fine-tuned, cross-validated XGBoost model on a balanced dataset. We tested the 5-year-long

**Table 5** Summary of classification performances of 5-category ESeC by using socio-economic and digital behavior indicator features. The summary was made regarding the 5-year-long period, and the classification model types (Random Forest, XGBoost, TabNet). The table contains the general accuracy of each model, along with the class-specific performance metrics: balanced accuracy, F1 score, Negative Predictive Value, Positive Predictive Value, Sensitivity, and Specificity for each ESeC5 class. The ESeC classes (ranging from Class 1 to 5) are as follows: higher-level service class, lower-level service class, intermediate class, class of skilled workers, class of unskilled workers, respectively

| Classification of ESeC5 by digital indicators and socio-demographic characteristics | | Accuracy | Balanced Accuracy | F1 | Negative Predictive Value | Precision (Positive Predictive Value) | Sensitivity (True Positive Rate) | Specificity (True Negative Rate) |
|---|---|---|---|---|---|---|---|---|
| Digital indicators of last 5 years - Random Forest | Class 1 | 0.395 | 0.673 | 0.583 | 0.803 | 0.509 | 0.683 | 0.663 |
| | Class 2 | | 0.501 | 0.257 | 0.692 | 0.310 | 0.220 | 0.783 |
| | Class 3 | | 0.489 | 0.247 | 0.692 | 0.281 | 0.220 | 0.758 |
| | Class 4 | | 0.547 | 0.400 | 0.750 | 0.328 | 0.512 | 0.583 |
| | Class 5 | | 0.600 | 0.424 | 0.713 | 0.560 | 0.341 | 0.859 |
| Digital indicators of last 5 years – XGBoost | Class 1 | 0.502 | 0.757 | 0.653 | 0.878 | 0.574 | 0.756 | 0.758 |
| | Class 2 | | 0.557 | 0.308 | 0.750 | 0.417 | 0.244 | 0.869 |
| | Class 3 | | 0.501 | 0.083 | 0.721 | 0.286 | 0.049 | 0.953 |
| | Class 4 | | 0.695 | 0.560 | 0.852 | 0.475 | 0.683 | 0.708 |
| | Class 5 | | 0.745 | 0.627 | 0.888 | 0.525 | 0.780 | 0.710 |
| Digital indicators of last 5 years – TabNet | Class 1 | 0.297 | 0.555 | 0.295 | 0.823 | 0.277 | 0.317 | 0.793 |
| | Class 2 | | 0.558 | 0.293 | 0.823 | 0.293 | 0.293 | 0.823 |
| | Class 3 | | 0.503 | 0.154 | 0.801 | 0.208 | 0.122 | 0.884 |
| | Class 4 | | 0.588 | 0.348 | 0.838 | 0.314 | 0.390 | 0.787 |
| | Class 5 | | 0.585 | 0.337 | 0.834 | 0.333 | 0.341 | 0.829 |

time-period with various classifiers detailed above (Random Forest, XGBoost, TabNet). The classification performances are summarized in Table 5.

According to the results of Table 5, when using digital behavior indicators and social characteristic features for classifying the 5-category ESeC, the XGBoost classifier performs the best: the general accuracy is 50.2%, which means a 30.2 percentage point increase compared to random chance. This model has definitely stronger performance compared to the basic model (where general accuracy was 32.7% and increase was 12.7 percentage points), which suggests the continued importance of social demographic characteristics in the determination of social position.

### 3.2 Analyses of the features

In the second part of the Results chapter, we aim to answer the second research question of the paper, namely, to identify those features in more detail, which play the most important role in the classification of different social positions. In the basic model, we only included the digital behavioral features from the 5-years period classifying the ESeC5 based social classes with XGBoost models. Although the performance of this model was not high, we believe the analyses of the strongest patterns observed among the features can bring us closer to the understanding of the relationship between digital behavior and social position and could serve as a potential starting point for further research.

As social class condenses multiple inequality-related characteristics, which can be important in the interpretation of the features, we first introduce the description of the classes by multiple socio-demographic dimensions (see Table 6). In the main text of the article, the weighted values are demonstrated (Table 6), while in the Supporting Information, we present the unweighted table (Table S2) as well. In general, by interpreting the

**Table 6** Description of social classes based on the European Socio-economic Classification (ESeC) by gender, age, domicile, education, marital status, household size, and occupational prestige. The table displays weighted column percentages

| | Higher-level Service Class | Lower-level Service Class | Intermediate Class | Skilled Workers | Unskilled Workers |
|---|---|---|---|---|---|
| Gender | | | | | |
| Male | 53.4% | 38.8% | 42.8% | 54.6% | 45.7% |
| Female | 46.6% | 61.2% | 57.2% | 45.4% | 54.3% |
| Age | | | | | |
| 16–29 | 15.9% | 11.6% | 19.9% | 20.4% | 15.2% |
| 30–39 | 28.9% | 24.0% | 24.6% | 20.4% | 28.3% |
| 40–49 | 21.6% | 19.4% | 24.7% | 29.6% | 26.1% |
| 50–59 | 14.8% | 14.0% | 15.8% | 14.2% | 17.4% |
| 60–69 | 12.5% | 26.3% | 12.3% | 12.3% | 13.0% |
| 70+ | 6.3% | 4.7% | 2.7% | 3.1% | 0.0% |
| Type of Settlement | | | | | |
| Capital | 36.3% | 20.2% | 16.4% | 20.4% | 14.1% |
| County Town | 18.8% | 22.5% | 23.3% | 25.3% | 22.8% |
| City | 29.0% | 36.4% | 37.0% | 24.7% | 25.0% |
| Village | 15.9% | 20.9% | 23.3% | 29.6% | 38.1% |
| Education Level | | | | | |
| Primary | 6.8% | 9.4% | 33.6% | 43.6% | 70.6% |
| Secondary | 20.5% | 43.0% | 46.5% | 51.5% | 27.2% |
| Tertiary | 72.7% | 47.6% | 19.9% | 4.9% | 2.2% |
| Marital Status | | | | | |
| Single | 17.7% | 18.9% | 29.9% | 26.4% | 22.8% |
| Married/Partnered | 74.3% | 70.1% | 56.2% | 65.6% | 68.5% |
| Divorced/Widowed | 8.0% | 11.0% | 13.9% | 8.0% | 8.7% |
| Household Size | | | | | |
| 1 | 20.6% | 20.0% | 22.1% | 17.8% | 15.4% |
| 2 | 29.7% | 36.2% | 30.3% | 35.0% | 35.1% |
| 3 | 21.7% | 23.8% | 22.1% | 27.6% | 23.1% |
| 4+ | 28.0% | 20.0% | 25.5% | 19.6% | 26.4% |
| Prestige | | | | | |
| Tier 1 (Lowest) | 2.8% | 1.6% | 14.7% | 28.8% | 98.9% |
| Tier 2 | 5.1% | 2.3% | 48.9% | 60.1% | 1.1% |
| Tier 3 | 19.9% | 64.1% | 34.3% | 11.1% | 0.0% |
| Tier 4 (Highest) | 72.2% | 32.0% | 2.1% | 0.0% | 0.0% |
| Total | 25.0% | 18.3% | 20.7% | 23.0% | 13.0% |

weighted values it is – from a methodological point of view – more correct to generalize for the target population, however as in the modeling the unweighted values are used, we entail that information in the Supporting Information section.

Regarding the European Socio-economic Classification based social classes, the ratios of different social classes in the sample were as follows: higher-level service class 25%, lower-level service class 18%, intermediate class 21%, skilled workers 23%, and unskilled workers 13%. The most important socio-demographic variable in the description of social classes is education and its distribution among classes is in line with expectations based on earlier research. In the two highest classes, those with tertiary education are over-represented, while in the intermediate class, those with secondary education have a higher ratio than the sample average. Skilled workers are characterized by having dominantly secondary education, while unskilled workers tend to have a mainly primary but also secondary education. Compared to the whole sample, the highest class is more likely to live in the cap-

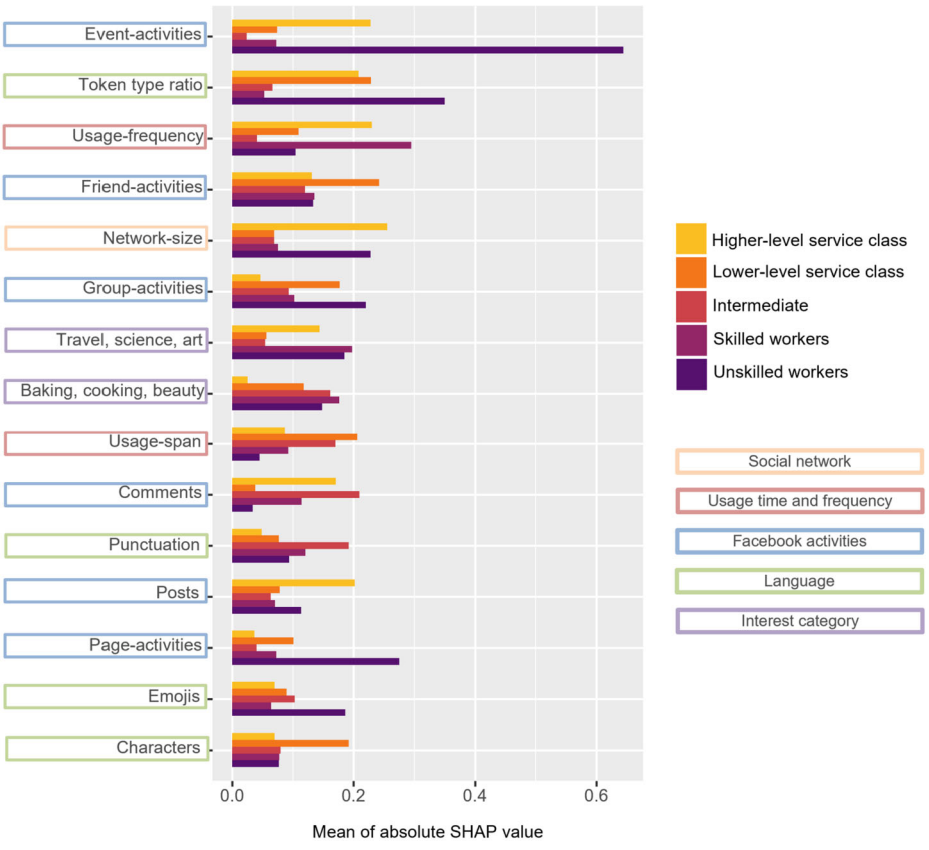


**Figure 1** General feature importance for the 5-category ESeC. The means of absolute SHAP values – the feature importance – are displayed in the order of the features: the higher a feature is positioned, the more it contributes to the classification. The color-coding, as shown in the upper legend, reveals the feature importance for each ESeC category (higher-level service class, lower-level service class, intermediate, skilled workers, and unskilled workers), while lower legend marks the feature categories (Social network, Usage time and frequency, Facebook activities, Language, and Interest category)

ital, while the lowest is in villages. Lower-level service class consists of more women, and there are more men among skilled workers. The prestige of the occupations is in line with class membership. (The detailed analysis of social classes by socio-demographic variables is presented in the Supporting Information, under the Description of Social Classes by socio-demographic indicators.)

Based on the overall importance of features, we examined which ones contribute to the classification of the social classes the most, for each of the 5-category of ESeC (see Fig. 1). Overall, the most important features belong to the initial concepts of language, usage time, interest categories, network-size, and different types of Facebook-activities.

Considering the diversity of language used (token type ratio) by the respondents, the observed amount of punctuation and length of the texts, as well as the number of emojis are among the top 15 features (out of the 21 used in the model). The features measuring various Facebook-activities are also dominant; the number of event-related, friend-related, group-related, page-related activities along with the number of comments and posts written. Usage time related indicators, such as the frequency and span of usage were also among the most important ones, and the network-size of the Facebook friends is also considered an important feature. From the different interest-categories, the most classify-

ing power is found in travel, science, and art, and in baking, cooking and beauty, according to these categories it is easier to differentiate between social classes.

As we defined a classification problem for the 5-category ESeC by using digital behavior indicators and all classes were involved in the modeling, the results can be compared to each other and represent the relative feature importance of each class. In addition to the general importance of features, it is important to note that the same features are not necessarily important for all classes. These results show that there are no general features that are important for all classes in the classification. Rather, each class has different features that are distinctive.

However, not only does the importance of digital features provide meaningful insights for the interpretation of the relationship between digital features and social class, but the direction of the features' effect on the classification. To interpret these relationships, we illustrated the SHAP values of each participant with color coding for different feature values in the classification of the five classes in Fig. 2. With such a presentation of the results, we can conclude on the relationship between different features and the given class position.

Based on the SHAP values (Fig. 2a), belonging to the *higher-level service class* increases with a larger Facebook network, a higher number of event-related activities, and interests in travel, science, and art. Additionally, less frequent posting and commenting, combined with more complex and diverse phrasing in texts, such as a higher token-type ratio and greater use of nouns and proper nouns with fewer emojis, also contribute positively. Conversely, activities related to managing Facebook friends – like adding, removing, following, unfollowing, and handling friend requests – reduce the probability, as do lower average Facebook usage spans and higher activity frequencies. Similarly, interests in music and films are less characteristic of this group.

Participants in the *lower-level service class* (Fig. 2b) are more characterized by the model (measured by the number of characters), use slightly more punctuation, emojis, nouns, and adverbs (but fewer proper nouns), and show interest in topics such as baking, cooking, and beauty. They also tend to be more active in relation to pages and reacting. On the other hand, they are less likely to have a high token type ratio, indicating less diverse textual content, and they write fewer posts, which aligns with their lower average usage span and less frequent platform use. Additionally, they are less active in group- and friend-related activities.

Members of the *intermediate class* (Fig. 2c) show interest in baking, cooking, beauty, European and American sports, and music and films. They also tend to use relatively higher amounts of punctuation and insert slightly more emojis into their posts and comments, which are often shorter. Additionally, they have larger networks, are more active in managing relationships and participating in groups and are active on Facebook over a wider timeframe during an average day. Conversely, they comment less, use fewer auxiliaries in their textual content, and have a lower token-type ratio, indicating less diverse phrasing, though this relationship is not necessarily linear.

Members of the *skilled working class* (Fig. 2d) use Facebook on a broader time horizon during the day, comment more frequently (albeit briefly), and have dominant interests in cars and men's sports. However, they tend to use Facebook less frequently overall, write with fewer proper nouns and less punctuation, and are less active in groups and friends-related activities. Additionally, they show a lower interest in travel, science, art, music, and films.

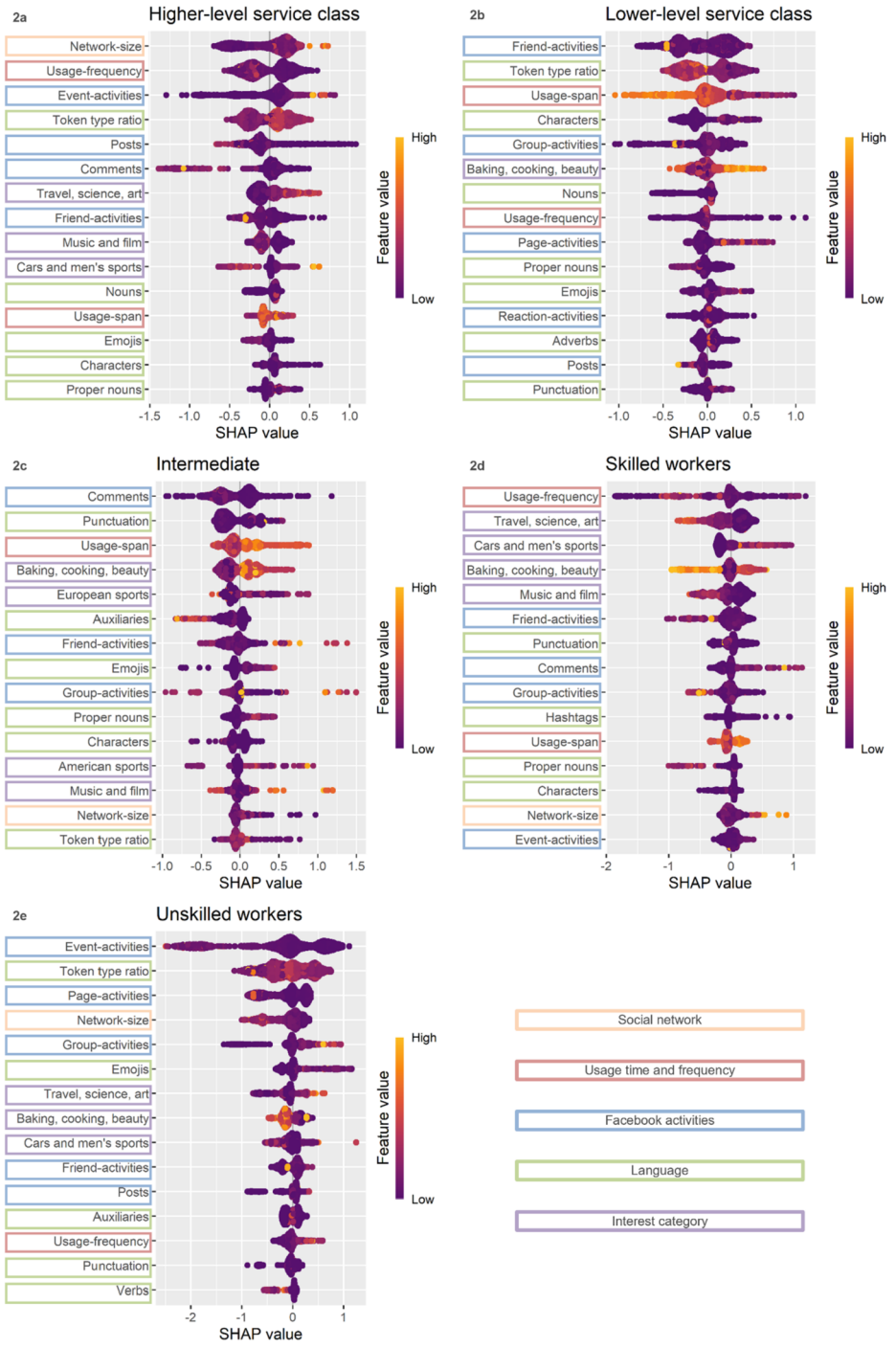


**Figure 2** SHAP beeswarm plot for summarizing the results of the XGBoost classification for the 5 categories of ESeC; 2a: higher-level service class, 2b: lower-level service class, 2c: intermediate, 2d: skilled workers, 2e: unskilled workers. The SHAP values belonging to the features are ordered hierarchically by feature importance (the higher a feature is positioned, the more important it is in the classification). The color-coding, as shown in the legends, reveals the values of the feature, while lower legend marks the feature categories (Social network, Usage time and frequency, Facebook activities, Language, and Interest category)

Members of the *unskilled workers* (Fig. 2e) tend to use Facebook more frequently, post more often, and be more active in groups. They also tend to have smaller networks and are less active in managing their friends (e.g., adding, removing, following). However, they participate less in event-related activities, follow pages, or respond to online or offline events. In terms of language, they use slightly fewer auxiliaries, verbs, and punctuation, and their class-belonging does not show a clear relationship with token-type ratio. They are more interested in cars, men's sports, and travel and science, but less so in baking, cooking, and beauty.

## 4 Discussion

The aim of this study was twofold. First, we investigated the extent to which users' social positions could be accurately classified using only basic quantitative data derived from their digital behavior. This approach provides contribution to the existing literature by utilizing a dataset that is representative of an entire society, based not on self-reported information but on actual behavioral traces, and which included directly observed—rather than inferred—social characteristics. To examine this, we applied a range of classification models, including Random Forest, XGBoost, TabNet (and for a complementary analysis for the classification of 5-category ESeC by digital behavior indicators along with socio-demographic characteristics). Feature selection was guided by prior literature that had identified meaningful associations with social class and position. As such, dimensions such as usage time, language used in posts or comments, the size and structure of users' social networks, patterns of contact management, activity types, and interest categories were operationalized with multiple digital behavioral indicators, which were used as features in the models.

The models, which only included the indicators of the observed digital behavior and classified the 5-class ESeC as a complex measure of social position did not show strong performance. Out of the three models tested, the XGBoost model showed the strongest overall accuracy with 32.7 percent. Using alternative measures for the operationalization of social position, like 4-category SES neither showed significantly better results (compared to the random distribution), nor those models, in which we limited the time constraint of the observed digital data closer to the measurement of social status. The similar results of the different model types, the varied timespan of the data, and the alternative measurement of social position suggest that the results are quite robust. Although this accuracy seems low, it is important to mention that in this case, a 5-class classifier was trained, which means that compared to random distribution, the model shows a 12.7 percentage point gain. As mentioned above, ESeC is a complex measure, and classifying it into 5 categories – as opposed to a binary classification – presents a significantly more challenging task. Therefore, we can conclude that basic indicators of digital behavior can only classify social position to a limited extent.

One reason for this performance can be the limited selection of features used in the model. Although we carefully selected the concepts which we included based on the results of earlier studies, other factors not included in the model can still play a role in the classification. This argument is strengthened by the results of those modeling scenarios, which complement digital behavior indicators with socio-demographic variables. Adding socio-demographic variables (such as gender, age, education level, and type of settlement) as features increased the models' performance. This suggests that we cannot properly classify the social class of social media users based solely on their digital behavior, without

considering traditional offline inequalities – as social class is strongly defined by socio-demographic characteristics.

For the selected set of features – exclusively digital behavior indicators – of the main analysis, it turned out that compared to a deep learning- and transformer-based model (TabNet) a more traditional machine learning model XGBoost performed better. Therefore, considering the nature of the features (purely quantitative, tabular data) involved in the modeling, and the classification performances of alternative models, for the main analysis we interpreted the results of the XGBoost model. Further solution for the extension of these modeling scenarios would be to extend the feature set with additional, not only quantitative, but also textual digital features. The weaker performance of TabNet classifier might be explained by the fact that traditional deep learning models – while performing well on image or textual data – are overparameterized for tabular data [51]. The vast amount of information that can be encoded in the high-dimensional representations of most neural networks cannot be effectively utilized. Additionally, unlike textual data, which benefits from a highly structured semantic embedding space where vectors carry semantic information that can be combined algebraically, tabular data generally lack such geometry [56]. Other studies showed that based on raw digital text data of Facebook posts and comments the social class of users can be detected with high performance [57]. Another opportunity for the extension of features could be the inclusion of indicators from other platforms, such as Google Search or YouTube. The usage of multiplatform data could also reflect on the diverse platform usage of different social groups [58] and could increase the performance of the classifiers. These methodological questions, including the data used, the features involved in the classification, and the architecture of the models applied, should be considered in future research. By examining and documenting various modeling scenarios, we hope that this article contributes to the debate on those methodological aspects, while providing insights about the social classes' digital behavior differences.

It is also worth mentioning that balanced accuracy varies a lot between the five social classes, suggesting that there are classes which members can be better classified than others. In the model, which only includes digital behavioral features, the lowest balanced accuracy can be observed at the highest class – the higher-level service class – and the model has the highest performance at the lowest class – the class of unskilled workers. These results suggest that based only on observed digital behavior, people with low status are the more distinguishable than others, their digital traces seem more remarkable than the ones of other classes. Although the description of the precise processes behind this phenomenon needs deeper research, this result indicates that the main fault line in second- and third level digital inequalities lies between the most vulnerable groups and the other parts of society.

The second objective of the study was to identify the features that played the most significant roles in classifying users into different social positions. Although the models did not perform very strongly, we believe that analyzing the most significant patterns of the features can add to our understanding of the relationship between various dimensions of digital behavior and social position because of the above-mentioned arguments on the goodness of a 5-category classification problem. The first result that is important to mention is that the most distinctive features are not necessarily the same for all classes, suggesting that different social positions can be distinguished by different types of behavior. In the following, we summarize the findings of the XGBoost model including features from

2018, and we focus on those results, which are relevant for the lowest class – namely the class of unskilled workers – that had the best performance in the classification models.

We found that the size of the network increases the probability of classification for almost all higher-level classes, while decreases it for the lowest class. This finding is in line with multiple social network related research (see e.g. [48, 59, 60]). Usage-frequency was also present among the most important features in almost all classes. The average daily frequency of usage decreased the classification to almost all classes, except the unskilled workers, where high frequency of usage was associated with higher group-belonging. These results are partly in line with previous studies, which suggest that lower classes spend more time in the digital space [11], as we detected that the breaking point between high and low frequency is between the lowest and the other classes. The results of the model show that the higher three classes post and comment less, while the two lower classes do it more. Regarding language characteristics, we found that more diverse language characterizes the highest class, while we can observe the opposite in almost all other classes. Interestingly, these results are only partially aligned with Bernstein's argument [15], namely that lower working class people use limited alternatives in their language, while individuals with higher status are characterized with more diverse language use. In our case, this distinction can be detected between the highest and other classes. Focusing on different Facebook activities, we should highlight that for the highest class, event related activities are frequent, while for the lowest class, such activities are rare. These results confirm the results of Yates et al. [14] that higher classes are more engaged with offline activities and interested in various events. Also, in each class, we can detect some interest-categories, which shape the classification.

Additionally to the results, the method of data collection and the findings of the paper rightly raise the question of social impact and privacy. Our approach respects user agencies by emphasizing the informed and consensual gathering of data. The data collection we conducted was preceded by careful planning of anonymization in order to ensure privacy of the participants, and through informed consent, users were informed about how and for what their data would be used. By openly studying these data, the predicting power and results of our models, we can spread awareness about how the predictive system based on digital footprints works. The average user might not realize how vast and in-depth the data gathered about them by the platform providers can be, and neither how the predictions-based algorithms work. By demonstrating how we can predict social position solely through the data that Facebook stores about them, we can highlight the importance of conscious online presence and user awareness in our increasingly digitalized world. For the impact on the academic scene, understanding how digital footprints can be used to predict social position when no tangible survey data is available can help improve models and reduce their bias, which can have cascading effects and social impact as well.

This study addressed the extent to which social position can be inferred solely from individuals' digital behavior. The study could contribute to the existing research by testing the question on a representative dataset that provided simultaneous access to both users' social status and their observed digital activity. By using various models with different timeframes and operationalizations, our findings indicate that models relying exclusively on digital behavior showed limited accuracy in predicting social class or socio-economic status. Nevertheless, they were more effective in identifying individuals in the lowest social strata. The addition of socio-demographic attributes substantially improved classifi-

cation performance, particularly in distinguishing between the most and least advantaged groups. The analysis of features aligned with prior findings from smaller or self-reported datasets, reinforcing the relevance of digital traces in understanding social inequalities.

**Abbreviations**
HUN-REN, Hungarian Research Network; XGBoost, eXtreme Gradient Boosting; DDP, data download packages; ESeC, European Socio-economic Classification; ESeC5, 5-category European Socio-economic Classification; GDPR, General Data Protection Regulation; Random Forest, RF; SHAP, SHapley Additive exPlanations; SMOTE, Synthetic Minority Oversampling Technique.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00578-2.

**Additional file 1.** (PDF 550 kB)

**Acknowledgements**
The authors thank Michelle Horváth and Anna Sára Ligeti for their useful comments and suggestions.

**Author contributions**
Author contributions: J.K. contributed equally to this work with Zs.R. Conceptualization: J.K., Zs.R., K.Sz., Á.H. Literature review: J.K., B.U., B.V., and Á.H. Data processing: J.K., Zs.R., Z.K., B.U., and B.V. Indicator creation: J.K., Zs.R., Z.K., B.U., and B.V. Analyses: J.K., Zs.R., Z.K., B.U., and B.V. Visualization: J.K., Zs.R. Supervision: J.K., Á.H. Writing—original draft: J.K., Zs.R., Z.K., K.Sz., B.U., B.V., Á.H. Writing—review & editing: J.K., Zs.R., Z.K., K.Sz., B.U., B.V., Á.H.

**Funding information**
Open access funding provided by HUN-REN Centre for Social Sciences. J.K., Zs.R., Á.H. acknowledges funding from the Hungarian Academy of Sciences Lendület Program: LP2022-10/2022. Z.K. acknowledges support from the Bolyai Scholarship, grant number: BO/834/22.

**Data Availability**
Data generated and analyzed during the current study, without sensitive data, is available in the repository of the HUN-REN Centre for Social Sciences Research Documentation Centre (https://openarchive.tk.mta.hu/629/). Repository access requires agreement from the corresponding author.

## Declarations

**Ethics approval and consent to participate**
The Ethical Approval of the research was prepared according to the standards of the HUN-REN Centre for Social Sciences, Budapest, Hungary in accordance with the Declaration of Helsinki. The data collection was fully complying with the actual European and Hungarian privacy data regulations and was approved by the Ethics Committee of the HUN-REN Centre for Social Sciences (resolution number 1-FOIG/130-37/2022).

**Consent for publication**
All authors have approved the manuscript and agree with the submission to EPJ Data Science.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]MTA–TK Lendület "Momentum" Digital Social Science Research Group for Social Stratification, HUN-REN Centre for Social Sciences, Tóth Kálmán utca 4, Budapest, 1097, Hungary. [2]Department of Social Research Methodology, Faculty of Social Sciences, ELTE Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, 1117, Hungary. [3]Department of Statistics, Faculty of Social Sciences, ELTE Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, 1117, Hungary. [4]CSS-RECENS, HUN-REN Centre for Social Sciences, Tóth Kálmán 13 utca 4, Budapest, 1097, Hungary. [5]Department of Sociology, Faculty of Social Sciences, ELTE Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, 1117, Hungary. [6]Doctoral School of Demography and Sociology, University of Pécs, Ifjúság útja 6, Pécs, 7624, Hungary. [7]Institute for Sociology, HUN-REN Centre for Social Sciences, Tóth Kálmán 13 utca 4, Budapest, 1097, Hungary.

Received: 19 January 2025 Accepted: 21 July 2025 Published online: 15 August 2025

**References**
1. Breuer J, Kmetty Z, Haim M, Stier S (2023) User-centric approaches for collecting Facebook data in the 'post-API age': experiences from two studies and recommendations for future research. Inf Commun Soc 26:2649–2668
2. Helsper E (2019) Why location-based studies offer new opportunities for a better understanding of socio-digital inequalities?

3. DiMaggio P, Hargittai E, Celeste C, Shafer S et al (2004) From unequal access to differentiated use: a literature review and agenda for research on digital inequality. Soc Inequal 1:355–400
4. Macevičiūtė E, Wilson TD (2018) Digital means for reducing digital inequality: literature review. Informing Sci: Int J Emerg Transdiscipline 21:269–287
5. Gui M, Büchi M (2021) From use to overuse: digital inequality in the age of communication abundance. Soc Sci Comput Rev 39:3–19
6. Hargittai E (2021) Handbook of digital inequality. Edward Elgar, Cheltenham Glos
7. Eynon R (2023) Utilising a critical realist lens to conceptualise digital inequality: the experiences of less well-off Internet users. Soc Sci Comput Rev 41:1081–1096
8. Tifferet S (2019) Gender differences in privacy tendencies on social network sites: a meta-analysis. Comput Hum Behav 93:1–12
9. Krasnova H, Veltri NF, Eling N, Buxmann P (2017) Why men and women continue to use social networking sites: the role of gender differences. J Strateg Inf Syst 26:261–284
10. Koiranen I, Keipi T, Koivula A, Räsänen P (2020) Changing patterns of social media use? A population-level study of Finland. Univ Access Inf Soc 19:603–617
11. Yates S, Kirby J, Lockley E (2015) Digital media use: differences and inequalities in relation to class and age. Soc Res Online 20:71–91
12. Quinn D, Chen L, Mulvenna M (2011) Does age make a difference in the behaviour of online social network users? In: 2011 international conference on Internet of things and 4th international conference on cyber, physical and social computing. IEEE, pp 266–272
13. North S, Snyder I, Bulfin S (2008) Digital tastes: social class and young people's technology use. Inf Commun Soc 11:895–911
14. Yates S, Lockley E (2018) Social media and social class. Am Behav Sci 62:1291–1316
15. Bernstein B (2003) Class, codes and control: applied studies towards a sociology of language, vol 2. Psychology Press
16. Aliakbari M, Allahmoradi N (2014) On the effects of social class on language use: a fresh look at Bernstein's theory. Adv Lang Lit Stud 5:82–88
17. Shi Y, Lei L (2021) Lexical use and social class: a study on lexical richness, word length, and word class in spoken English. Lingua 262:103155
18. Kacewicz E, Pennebaker JW, Davis M, Jeon M, Graesser AC (2014) Pronoun use reflects standings in social hierarchies. J Lang Soc Psychol 33:125–143
19. Cesare N, Grant C, Nguyen Q, Lee H, Nsoesie EO (2017) How well can machine learning predict demographics of social media users? Preprint. Available at arXiv:1702.01807
20. Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist J (2011) Understanding the demographics of Twitter users. In: Proceedings of the international AAAI conference on web and social media, vol 5, pp 554–557
21. Vedres B, Vasarhelyi O (2019) Gendered behavior as a disadvantage in open source software development. EPJ Data Sci 8:25
22. Hinds J, Joinson AN (2018) What demographic attributes do our digital footprints reveal? A systematic review. PLoS ONE 13:e0207112
23. McCormick TH, Lee H, Cesare N, Shojaie A, Spiro ES (2017) Using Twitter for demographic and social science research: tools for data collection and processing. Sociol Methods Res 46:390–421
24. De Choudhury M, Sharma S, Kiciman E (2016) Characterizing dietary choices, nutrition, and language in food deserts via social media. In: Proceedings of the 19th acm conference on computer-supported cooperative work & social computing, pp 1157–1170
25. Filho RM, Borges GR, Almeida JM, Pappa GL (2014) Inferring user social class in online social networks. In: Proceedings of the 8th workshop on social network mining and analysis, pp 1–5
26. He Y, Tsvetkova M (2023) A method for estimating individual socioeconomic status of Twitter users. Sociol Methods Res 00491241231168665
27. Preoţiuc-Pietro D, Volkova S, Lampos V, Bachrach Y, Aletras N (2015) Studying user income through language, behaviour and affect in social media. PLoS ONE 10:e0138717
28. Rao D, Yarowsky D, Shreevats A, Gupta M (2010) Classifying latent user attributes in Twitter. In: Proceedings of the 2nd international workshop on search and mining user-generated contents, pp 37–44
29. Lazer D, Radford J (2017) Data ex machina: introduction to big data. Annu Rev Sociol 43:19–39
30. Lőrincz L, Koltai J, Győr A, Takács K (2019) Collapse of an online social network
31. Matz SC, Menges JI, Stillwell DJ, Schwartz HA (2019) Predicting individual-level income from Facebook profiles. PLoS ONE 14:e0214369
32. Haenschen K (2020) Self-reported versus digitally recorded: measuring political activity on Facebook. Soc Sci Comput Rev 38:567–583
33. Guess AM, Barberá P, Munzert S, Yang J (2021) The consequences of online partisan media. Proc Natl Acad Sci USA 118:e2013464118
34. Mangold F, Schoch D, Stier S (2024) Ideological self-selection in online news exposure: evidence from Europe and the US. Sci Adv 10:eadg9287
35. Zannettou S, Nemeth O-N, Ayalon O, Goetzen A, Gummadi KP, Redmiles EM, Roesner F (2023) Analyzing User Engagement with TikTok's Short Format Video Recommendations using Data Donations. https://doi.org/10.48550/arXiv.2301.04945
36. Boeschoten L, Ausloos J, Möller JE, Araujo T, Oberski DL (2022) A framework for privacy preserving digital trace data collection through data donation. Comput Commun Res 4:388–423
37. van Driel II, Giachanou A, Pouwels JL, Boeschoten L, Beyens I, Valkenburg PM (2022) Promises and pitfalls of social media data donations. Commun Methods Meas 16:266–282
38. Bishop YM, Fienberg SE, Holland PW (2007) Discrete multivariate analysis: theory and practice. Springer, Berlin
39. Statista.com, Number of Facebook users in Hungary from September 2018 to September 2023. https://www.statista.com/statistics/1029770/facebook-users-hungary/#:~:text=The%20number%20of%20Facebook%20users,2022%20at%207.43%20million%20people

40. DataReportal.com, DIGITAL 2022: Hungary. https://datareportal.com/reports/digital-2022-hungary
41. Statista.com, Social media usage in Hungary - statistics & facts. https://www.statista.com/topics/6592/social-media-usage-in-hungary/#topicOverview
42. Rose D, Harrison E (2010) Social class in Europe. An introduction to the European socio-economic classification
43. Erikson R, Goldthorpe JH (1992) The constant flux: a study of class mobility in industrial societies. (No Title)
44. Goldthorpe JH (2007) On sociology second edition volume two: illustration and retrospect, vol 2. Stanford University Press, Stanford
45. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat1189–1232
46. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J (2025). Xgb. contributors (base Xgb. implementation), xgboost: Extreme Gradient Boosting, version 1.7.11.1. https://cran.r-project.org/web/packages/xgboost/index.html
47. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
48. Johnson CL (1994) Differential expectations and realities: race, socioeconomic status and health of the oldest-old. Int J Aging Hum Dev 38:13–27
49. Breiman L (2001) Random forests. Mach Learn 45:5–32
50. Arik SO, Pfister T (2020) TabNet: Attentive Interpretable Tabular Learning. https://doi.org/10.48550/arXiv.1908.07442. arXiv:1908.07442 [Preprint]
51. Arik SÖ, Pfister T (2021) TabNet: attentive interpretable tabular learning. Proc AAAI Conf Artif Intell 35:6679–6687
52. Del Moral P, Nowaczyk S, Pashami S (2022) Why is multiclass classification hard? IEEE Access 10:80448–80462
53. Rahamim A, Uziel G, Goldbraich E, Anaby Tavor A (2023) Text augmentation using dataset reconstruction for low-resource classification. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the association for computational linguistics: ACL 2023, vol 466/. Association for Computational Linguistics, Toronto, pp 7389–7402. https://aclanthology.org/2023.findings-acl
54. Huszár Á, Füzér K (2023) Improving living conditions, deepening class divisions: Hungarian class structure in international comparison, 2002–2018. East Eur Polit Soc 37:740–763
55. Kolosi T Chap. 8 Transitions and Structural Distortions. Brill, 2019. https://brill.com/display/book/9789004400283/BP000008.xml
56. Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on tabular data? Preprint. https://arxiv.org/abs/2207.08815v1
57. Váradi B (2024) Kísérletek A Társadalmi Osztály És Nyelv Kapcsolatának Vizsgálatára Klasszifikációs Modellek Döntéseinek Felhasználásával
58. Auxier B, Anderson M (2021) Social media use in 2021, pew research center. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/
59. Moore G (1990) Structural determinants of men's and women's personal networks. Am Sociol Rev 55:726–735
60. Campbell KE, Marsden PV, Hurlbert JS (1986) Social resources and socioeconomic status. Soc Netw 8:97–117


## Publisher's note