Language Resources and Evaluation (2023) 57:1547–1570
https://doi.org/10.1007/s10579-022-09612-5

ORIGINAL PAPER

The language of discrimination: assessing attention discrimination by Hungarian local governments

Jakab Buda[1] · Renáta Németh[1] · Bori Simonovits[2] · Gábor Simonovits[3,4,5]

Accepted: 5 August 2022 / Published online: 6 September 2022

Abstract

In our study we assess the responsiveness of Hungarian local governments to requests for information by Roma and non-Roma clients, relying on a nationwide correspondence study. Our paper has both methodological and substantive relevance. The methodological novelty is that we treat discrimination as a classification problem and study to what extent emails written to Roma and non-Roma clients can be distinguished, which in turn serves as a metric of discrimination in general. We show that it is possible to detect discrimination in textual data in an automated way without human coding, and that machine learning (ML) may detect features of discrimination that human coders may not recognize. To the best of our knowledge, our study is the first attempt to assess discrimination using ML techniques. From a substantive point of view, our study focuses on linguistic features the algorithm detects behind the discrimination. Our models worked significantly better compared to random classification (the accuracy of the best of our models was 61%), confirming the differential treatment of Roma clients. The most important predictors showed that the answers sent to ostensibly Roma clients are not only shorter, but their tone is less polite and more reserved, supporting the idea of attention discrimination, in line with the results of Bartos et al. (2016). A higher level of attention discrimination is detectable against male senders, and in smaller settlements. Also, our results can be interpreted as digital discrimination in the sense in which Edelman and Luca (2014) use this term.

Keywords Correspondence study · Controlled field experiment · Attention discrimination · Natural language processing · Machine learning · Metric of discrimination

✉ Renáta Németh
nemeth.renata@tatk.elte.hu

Extended author information available on the last page of the article

1 Introduction

Understanding why and on what grounds people discriminate based on observable group attributes has been one of the core interests of the social sciences for the past decades. Prior research evidence suggests that discrimination based on ethnicity is pervasive in Hungary as well as in other European countries, even though two EU directives[1] were adopted in 2000 designed to reduce discrimination. In Hungary, both direct and indirect forms of discrimination occurs regularly in various spheres of everyday life such as the labour market, housing, access to education and health care (FRA, 2018), as well as police stops and search practices (on racial profiling see Miller et al., 2008). Furthermore, unequal treatment of minorities in all sorts of public services and retail (e.g. banks, municipality services, restaurants, car dealers, bars, and shops) is presumably prevalent; however, our knowledge in this field is limited. Since 2006 several field experiments have been carried out in Hungary, mostly to explore the mechanisms of discrimination in the labour market against various vulnerable social groups, i.e. the Roma, overweight people, and people with disabilities (Pálosi et al., 2007; Sik & Simonovits, 2008). Most recently discrimination by local governments was studied using randomized field experiments (Csomor et al., 2021; Simonovits et al., 2021).

In Hungary there is a long tradition of research exploring the structure of attitudes towards various minorities including the Roma, Jewish people, and more recently immigrants (Enyedi et al., 2004; Örkény & Váradi, 2010; Sik et al., 2016; Simonovits & Szalai, 2013). Experimental researchers have also explored ways to tackle exclusionary attitudes through interventions conducted in educational settings (Kende et al., 2017; Simonovits & Surányi, 2020), or embedded in online games (Simonovits et al., 2018). When designing the present study, we sought to build on this body of research by investigating possible ways in which exclusionary or discriminatory behaviour can be measured and changed in Hungary. Using the theoretical framework of attention discrimination—originally developed by Matejka (2013) and Bartoš et al. (2016) linking the concepts of discrimination and scarce attention—our research group approached all Hungarian local governments via sending them requests purportedly by Roma and non-Roma citizens in multiple waves in 2020. Beyond using the concept of attention discrimination, we also relied on the idea of digital discrimination, in line with the interpretation offered by Edelman and Luca (2014).

After the pilot study was completed (Csomor et al, 2021), we broadened the research design with an intervention (in close cooperation with a leading Hungarian NGO) to test whether the behaviour of the local governments can be changed with such a stimulus. We carried out a series of online correspondence studies of

[1] Race Equality Directive (2000/43/EC) prohibits discrimination on grounds of race and ethnic origin, covering various fields of social life. Employment Equality Directive (2000/78/EC) prohibits discrimination on grounds of religion and belief, age, disability, and sexual orientation, covering the fields of employment and occupation, vocational training, membership of employer and employee organisations. Source: https://ec.europa.eu/commission/presscorner/detail/en/MEMO_07_257 and https://ec.europa.eu/commission/presscorner/detail/en/MEMO_08_69

the Hungarian local governments (N = 1260). We aimed to explore to what extent human rights NGOs can protect minorities against unequal treatment by Hungarian local governments. We studied this question by combining an audit experiment in Hungary with an intervention conducted in collaboration with a major Hungarian NGO—cuing Roma ethnicity by using a combination of stereotypically Roma sounding first and family names.

In the audit experiment we demonstrated (Simonovits et al., 2021) that high status Roma individuals were about 13% less likely to receive responses to information requests from local governments, and the responses they received were of substantially lower quality. The intervention that reminded a random subset of local governments of their legal responsibility of equal treatment lead to a short-term reduction in discrimination, but the effects of the intervention dissipated within a month. These findings are similar to those of Distelhorst and Hou (2014) as well as of Einstein and Glick (2017) both in terms of baseline responsiveness and discrimination. We found similar patterns analyzing the content of the emails.

Based on our empirical results summarised above, using the data of our study Simonovits et al., (2021) we devote the present paper to fulfilling a double—a methodological and a substantive—aim. The methodological novelty of our paper is that we treat discrimination as a classification problem and study to what extent emails written to Roma and non-Roma clients can be distinguished, which in turn serves as a metric of discrimination in general. By doing so we examine whether it is possible to detect subtle forms of discrimination in textual data in an automated way without human coding, and whether machine learning can detect features of subtle discrimination that coding guidelines do not cover. We examine the role of potential modifiers by assessing the classification for certain subgroups separately. We think our study shows how machine learning and the method of natural language processing (NLP) can expand the methodological toolkit of discrimination research.

To the best of our knowledge, our study is the first to propose an approach to assess discrimination that employs machine learning techniques. Hence, we apply a broader perspective, and discuss the conceptual and practical aspects of this procedure in comparison with the traditional alternatives. The substantive aspect of our study focuses on the level of discrimination measured in this computational way, and on the strength of modifier factors measured at both individual and settlement-level. However, it is important to note, that with the application of NLP technique we are only able to assess *subtle forms* of discrimination, as we are only able to analyse texts of the answers, which means that the reactions of those officials who did not answered to the fictious clients are not included into the analysis. We also try to identify linguistic features that the algorithm considers most important in making the distinction between responses written to Roma or non-Roma clients. In other words, in addition to measuring discrimination, we also aim to understand the linguistic distinctions we detect. Finally, we examine whether potential modifier factors (the gender of the client, and the size of settlement the local government is located in) affect the level of detectable discrimination.

The rest of the paper is organised as follows. First, the literature review is presented (Sect. 2), then the methodology on which we based this study is explained

and justified (Sect. 3). In Sect. 4 the findings are presented, and then discussed (Sect. 5).

2 Literature review and our own approach

2.1 Understanding discrimination in audit and correspondence studies

Over the last decades there has been growing empirical evidence on ethnic and other types of discrimination, mostly obtained by field experiments. A basic differentiation between audit and correspondence studies can be made. The general idea of audit studies is using pairs of trained testers (auditors), matched for relevant characteristics except for the experimental variable (i.e. those presumed to lead to discrimination, e.g. race, ethnicity, gender, and age). Discrimination is detected when "auditors in the protected class are systematically treated worse than their teammates" (Yinger, 1998). Recently, beyond audit studies there have also been correspondence studies (or email audit studies), in which only written materials (primarily emails, reference letters, and cover letters) are used to test discrimination in the labour market or in other spheres of daily lives, such as access to services. The largest advantages of correspondence studies are (i) that "the method is largely immune to criticisms of failure to control for important differences between, for example, black and white job applicants" (Neumark, 2012, p. 3), (ii) that it can be applied in large numbers, and (iii) that it is relatively cheap to implement. On the other hand, correspondence technique is only appropriate for assessing the first phase of the interaction (most likely the application process). To sum it up, using correspondence studies is an excellent and cost-effective way to examine discriminatory behavior in the real world (Verhaeghe, 2022).

To better understand the subtle ways in which discriminatory selection mechanisms take place in online communication—first we need to distinguish between overt (or direct) and more subtle forms of discrimination. Overt discrimination can be defined as "blatant antipathy, beliefs that [members of stereotyped groups] are inherently inferior, [and] endorsement of pejorative stereotypes" (Cortina, 2008, p. 59). Whereas subtle discrimination means a behaviour that are ambiguously intent to harm, low in intensity, and often unintentional, therefore difficult to detect. (Cortina, 2008; Dipboye & Halverson, 2004). In their meta-analysis of 90 studies, Jones et al. (2017b) pointed out that subtle and overt forms of discrimination are highly correlated. The authors argue, therefore that subtle discrimination, which is often normalized and overlooked in the workplaces, should be taken more seriously.

As a theoretical background we use the terms of digital discrimination as well as attention discrimination (Matejka, 2013; Bartoš et al., 2016) to better understand subtle ways of discriminatory behaviour in the online communication between municipalities and their clients. In line with Bartoš et al. (2016), Huang et al. (2021) empirically prove that there is discrimination arising from differential attention allocation (or inattention) to minority and majority applicants in online lending markets in China. Huang et al. (2021) argue that discrimination is often aggravated by the attention constraints of decision makers. In other words, in selection or screening

processes, decision makers tend to pay less if any attention to minority applicants, who often have lower levels of proper language skills or qualifications. It is not surprising that this kind of inattention may result in discriminatory selection processes.

We use the term digital discrimination based on the work of Edelman and Luca (2014), who earlier applied this concept to assess the prevalence of discrimination in online marketplaces, namely on the Airbnb platform. They argue that in contrast to face-to-face interaction, where it is impossible not to disclose information about the applicants' identity (e.g. a job interview), in digital transactions the flow of undesirable or unnecessary information can easily be reduced. There is contradictory research evidence whether the use of digital communication (in contrast to offline experiences) can successfully reduce racial discrimination. While Morton et al. (2003) pointed out that using online markets for buying a car is disproportionately beneficial to minorities (i.e. for African-American and Hispanic clients, who are least likely to use the Internet) as compared to offline bargaining situations. On the other hand, there is a growing amount of research evidence suggesting that discrimination seems to remain an important policy concern in online marketplaces (Edelman et al., 2017; Cui et al., 2020), as most of these platforms encourage users to provide personal profiles and even photos of themselves in order to build trust. We may conclude that the benefits of online communication depend on the design of the online platforms. In our research design we used simple emails as the means of online communication, and intentionally disclosed the racial background of the clients.

2.2 Using textual data in discrimination research

Many studies focus on discrimination as a behavioral act by a decision maker (in employment see the landmark study of Bertrand & Mullainathan, 2004; for housing Massey & Lundy, 2001; for access to private or public services see e.g. Zussman, 2013; etc.). However, subtle forms of discrimination may also appear as a linguistic element used when engaging in communication with members of the target group. The aforementioned audit and correspondence studies also examine these linguistic phenomena. Crabtree (2018) provides a good overview of the implementation of correspondence studies, and also offers recommendations on analyzing emails. According to Crabtree the most commonly used indicators in correspondence studies are quantitative indicators, i.e. whether the sender received a reply, and the time elapsed between the request and reply. In other cases, qualitative aspects of the texts are also assessed, i.e. the content (helpfulness or sentiment) of the reply. Crabtree offers different assessment methods: texts can be coded manually, or quasi-automatically using a predefined dictionary, compiled by researchers. Crabtree also mentions natural language processing (NLP) as a fully automated method, but does not describe the specific way in which it can be used.

When it comes to unequal access to public services, there is a large and growing body of literature on discrimination by local governments (Distelhorst & Hou, 2014; Hemker & Rink, 2017). However, there has been little work done exploring ways in which such bias may be ameliorated. To our knowledge the only such effort is by

Butler and Crabtree (2017), who find no effect of an information treatment delivered by researchers. One crucial issue is that interventions implemented by researchers might not be taken seriously by local governments (Butler & Crabtree, 2017; Kalla & Porter, 2019), and governments themselves might not have sufficient incentives to intervene when they observe discriminatory behaviour.

In our research our starting point was that a binary outcome variable alone—whether an email is answered or not—is a rather coarse measure of discrimination (Banerjee & Duflo, 2017). We believe that numerical outcome measures measuring the quality and sentiment of responses are more appropriate tools to tackle subtle forms of discrimination. As subtle forms of discrimination (as opposed to more overt manifestations) are more prevalent in countries where discrimination is legally banned, in our view it is worth exploring these subtle mechanisms with complex questions, which can be better implemented through emails. We also agree with Jones et al. (2017a) arguing that due to the rising pressure of egalitarianism, the prevalence of subtle and unintentional forms of discrimination could be understood as a vicious cycle at the workplace, but in our view, this is the case also beyond the labour market, in various areas of our social lives (e.g. housing market, access to public and private services).

In the first phase of our study (Simonovits et al., 2021) we applied human coding to measure the helpfulness and politeness of the replies. In the present paper we have relied on automated processing of the emails by using NLP. Our approach was not to evaluate the content of the email without taking into account the ethnicity of the sender, but on the contrary, we tried to identify patterns of the text that are connected to the ethnicity of the sender. Technically, we did this by using machine learning techniques. We tried to find an algorithm that predicts the ethnicity of the sender well from a limited number of features derived from the text of the response they received. We limited the number of features used because one of our goals was to be able to interpret the built model, despite machine learning methods being often described as black boxes. In order to find the most optimal algorithm, we tried different statistical models and different text features.

In addition to prediction, an analysis of the linguistic features most relevant to the prediction is also presented. For more details on the method as an operationalisation of discrimination measurement and its advantages/disadvantages compared to other alternatives, see Sect. 3.3.

2.3 Combining experiments and textual analysis to assess discrimination

Based on our extensive literature review (Csomor et al., 2021) we were only able to identify one study that combined the techniques of randomised field experiment and textual analysis aiming to assess discrimination at public services, one based on a large-scale online correspondence study—however, this study (Giulietti et al., 2019) used simple content analysis, relying on some predefined features of the text. The researchers found that requests coming from a person with a distinctively black name were less likely to receive a reply compared to those from a person with a distinctly white name (72 percent vs. 68 percent). The qualitative analysis of the

answers pointed out that replies to requests coming from black names were less likely to have a cordial tone.

In certain respects the study by Giulietti et al. (2019) was a model for the research design of our own study. However, Giulietti et al. (2019) used a simple form of text analysis, and did not use natural language processing. Although not in the context of local government research, there is an example (if only one, to the best of our knowledge) of experimental design and NLP being used together in other areas of discrimination research, namely that by Bohren et al. (2018). They carried out a partly similar research design to that of our own present research in an online setting in the United States. The research context was a popular online mathematics Q&A forum, where users may post, reply, and comment. The researchers randomly assigned male or female usernames to the posted questions (140 original mathematics questions at college-level) and tested whether people respond differently to questions posed by women versus men. The study is limited to very simple solutions when using NLP methods. To find differences between responses to male versus female questions, two types of statistical natural language processing methods were used. They tested the difference between the probability distributions over the two (female vs. male) sets of words. Next, they applied a dictionary-based sentiment analysis approach, with a binary classification (positive or negative) of words, and calculated the difference between the two subcorpora in the average positive and negative sentiment score. It is worth mentioning that they did not use machine learning methods. According to the results, both methods showed a significant difference to the disadvantage of women.

The last study to describe is beyond the field of discrimination research (Boulis & Ostendorf, 2005). They studied linguistic differences between genders in telephone conversations. The study corpus consisted of telephone conversations between pairs of people, randomly assigned to speak to each other about a randomly selected topic. They tried to classify the transcript of each speaker into the appropriate gender category. Additionally, they tried to classify the gender of speaker B given only the transcript of speaker A, the purpose of which is directly parallel with our own research question. Similarly to our design, they aimed at not only detecting differences but also explaining them: they tried to reveal the most characteristic features for each gender.

3 Data and methods

3.1 Research questions

As detailed in the previous section, our research questions were as follows:

RQ1 (methodological). How to expand the toolkit of discrimination research with computational text analysis, and how to define a measure of discrimination.

RQ2 (substantive). Whether the computational text analysis system can detect differential treatment against the Roma minority where human coding can as well. (Aim: measuring.) What are the linguistic features the algorithm detects behind the discrimination? (Aim: understanding.)

We describe below in detail the database and the methods we used. We report all measures, manipulations, and exclusions made on the data. Due to space limitations, technical details have been moved to the Supplementary Material accompanying our paper.

3.2 Data

As discussed above, in the present paper we carry out a secondary analysis of our audit study (Simonovits et al., 2021) that took place in Hungary in 2020.[2]We contacted local governments from fake email accounts signaling the sender's Roma or non-Roma ethnicity. We sent emails from 9 different accounts with one of four different requests. We compiled questions that do not require extra effort to answer by the clerks, based on previous international studies (mentioned above), i.e. about a biking trip the requester was planning to make, about nurseries in the area of the municipality, about the local cemetery, and about possible wedding venues in the area. The key manipulation in the request was the purported ethnicity of the requester. Beyond ethnicity, we also varied the gender of the requester through both the fake email address itself, and through the signature of the requester. We used relatively educated language in order to increase response rates.

We cued Roma ethnicity using stereotypically Roma sounding names (both first and family names). In line with previous landmark experimental research (Bertrand & Mullainathan, 2004), our primary aim was to identify distinctive Roma and non-Roma names. However, as opposed to the relevant US research tradition,[3] we not only used first names to express Roma identity, but also family names, as many Roma people living in Hungary have distinctive Roma family names. To select appropriate Roma and non-Roma names, we used multiple sources. Based on previous results of Hungarian surveys (e.g. Simonovits et al., 2018; Váradi, 2012) and experimental research completed in the Hungarian field (Sik & Simonovits, 2008), we carefully selected distinctively Roma and non-Roma names (both family and first names) for our testers.

Sample size was determined before any data analysis: in our within subject design each municipality received two emails. The order of gender and ethnicity was independently randomized so that each of the 1260 municipalities received a Roma and a non-Roma request. A last treatment arm randomly assigned whether municipalities were followed-up if a response was not received within a week. For further methodological details of the study see Simonovits et al. (2021). The original design was broadened with an intervention to test whether the behaviour of the local

[2] The audit study received IRB clearance, and was compliant with relevant Hungarian laws. We debriefed the research subjects soon after the data collection phase was completed. Our experiment was pre-registered. The pre-analysis plan is available at https://osf.io/38gax/, and the data is available at https://doi.org/10.7910/DVN/KPSCLK. The codes used in our secondary analysis are available at this Github repository: https://github.com/jakabbuda/langdiscr. We report all data exclusions, all manipulations, and all measures in our study.

[3] Most of which took the work of Bertrand and Mullainathan (2004) as a starting point.

governments could be changed. In the present paper we do not investigate the effect of the intervention, as the sample is too small to conduct a supervised classification (only 147 recipients opened the intervention email). However, with a larger sample it would be possible to test the effectiveness of the intervention by comparing the performance of the predictive models for the intervention with the control samples.

We included every response received in this study (only automatic responses were excluded), so the valid sample size is 1330. The response rate was 52.8% overall, but this was not the same across the two ethnic groups: purportedly Roma requesters received a response only 47.2% of the time as opposed to purportedly non-Roma clients, who had a response rate of 58.3%. This difference is statistically significant (Simonovits et al., 2021), and it affects our models through the unbalanced training set. A balanced set of 200 was randomly selected from the responses to be used as test data. (Although non-Roma requesters had a significantly higher chance of receiving an answer, we used a balanced test set because there is no reliable information on the Roma/non-Roma ratio in the population considered. The estimation of the model's performance is therefore expected to be rather conservative.) The remaining 1130 responses were used as a training and development set (44% of these responses were written to purportedly Roma requesters). Since this is not a large dataset (consisting of 66 561 words), we had to limit the size of the models that we built.

3.3 Expanding the methodological toolkit: a machine-learning-based discrimination measure defined by textual data

In traditional discrimination research, measures of discrimination have been defined in terms of the average difference in certain outcomes between majority and minority groups. In contrast, the presence of discrimination in our computational approach is due to the existence of a model that predicts ethnicity with some efficiency. What does "some efficiency" mean? We do not need a model that classifies almost every observation into the right category since that would imply that all officials are discriminatory. Rather, it is sufficient to achieve a predictive accuracy that is significantly better than the accuracy of random classification. Obviously, if the ethnicity of the clients did not influence the responses, the accuracy of the classification model would not exceed that of a random classification. This approach allows to define a polarization metric: the greater the classification model's ability to identify ethnicity, the higher the level of polarization.

To the best of our knowledge, our study is the first to propose predictive accuracy as a discrimination measure. However, similar approaches can be found in other areas, such as language polarisation research in political science. Classification in this context is used to identify the ideological position of an author based on the words she used (see eg. Green et al., 2020; Gentzkow et al., 2019; Bayram et al, 2019). The greater the classification model's ability to identify the position of the author, the greater the polarization.

How to interpret the measure? Officials writing differently to Roma clients does not necessarily constitute negative discrimination: they might simply be

overcompensating. This is an important but easily overlooked aspect of studying discrimination by comparing textual data. Eg. Bohren et al. (2018) detect "discrimination" when observing a significant difference in word distribution in replies written to female or male users. This uncertainty of interpretation is the reason why our goal is not only to compute the value of the measure, but also to explain it by determining which text patterns are strongly correlated with responses to Roma versus non-Roma clients. Human coding or dictionary-based approaches allow for a more unambiguous interpretation, but also have drawbacks from a hermeneutical point of view: the coding guideline, or the dictionary, reflects the horizon of the researchers, limiting the range of detectable linguistic differences, whereas NLP can potentially find any kind of difference.

Alternatives to NLP (human coding/dictionary-based assessment) give results that can be used as numerical input for any classical statistical method (see for example Bohren et al., 2018, who assigned numerical sentiment scores to the texts), so that discrimination can be tested statistically by comparing treated and control groups. The logic of predictive modelling used by us, however, is fundamentally different. The differences between these two approaches (statistical modeling vs. predictive modeling; explanation vs. prediction) are important both from an epistemological and a statistical point of view. On the statistical side, there are at least two decades of reflection on this opposition: see Breiman (2001) and Shmueli (2010) for two highly cited examples. Predictive models focus on how well a trained model can estimate an outcome that was not used in training. The models are assessed on their predictive performance so that researchers may discover the best ones. These models perform well on out-of-sample data, but often produce hard-to-interpret results as the link between the prediction and the input cannot be easily described (Molnar, 2019). Predictive modelling is very typical in natural language processing. In our research we have chosen a predictive approach as it fits better (1) with the logic of our research question ("can we guess the ethnicity of the sender from the reply alone?") and (2) with the nature of our data (a large dataset with potentially complex relationships between the ethnicity of the sender and the linguistic features of the reply).

In the case of statistical modeling, classical statistical quality criteria exist. Most of the equivalents to these criteria in the case of predictive modeling are not interpretable, poorly quantifiable, or not yet available. However, some important considerations can be stated. The size and composition of the test set strongly influence the estimated model performance. A smaller test set will obviously give less reliable results. Therefore, in what follows, when we create a new subset and examine predictive performance on on it, a reduction in the sample size itself introduces uncertainty into the results. Additionally, if the Roma versus non-Roma ratio in the training set is unbalanced, this in itself introduces some performance degradation.

3.4 Modelling

The emails in their raw form need further preprocessing to be usable for analytics. We removed all parts of the emails that were not within the body of the answer,

including automatic signatures and contact information of the responders. We changed to a specific token all municipality names, personal names, addresses, dates, times, telephone numbers, email addresses, URLs, and any other numbers, and standardized the formatting of the messages (e.g. we removed double whitespaces, newlines, etc.). For the n-gram models (see later), we experimented with different levels of preprocessing. In the simplest one, the only additional change we made was to lowercase the messages. For our second model, we also removed all very common words (e.g. *stop words*, like the articles "the," "a," and "an"), and for the third one, words were unified by reducing them to their lemmas (*lemmatization*). For the latter as well as for POS tagging we used Python's spaCy NLP toolkit, developed for Hungarian by György Orosz[4].

After the preprocessing, we built two different models based on different variables. The first model was based on descriptive statistical features characterizing the text of the emails. Since our goal was to build a model which is well-performing but at the same time interpretable, we selected features that are commonly used and effective for many binary classification applications. By selecting these features we aimed to understand their role in discriminating against Roma people. The second model was based on the *n-grams* (consecutive words of size n) of the emails. We can say that while the first model took into account properties of the texts, the second model worked with the texts themselves. We used XGBoost for both models, as it is a generally well-performing method for prediction.[5]

As we already mentioned in Sect. 3.3, the objective of predictive modelling is to have a model that performs well both on the data that we used to train it on, and on new data the model will be used on to make predictions. We used the method of *train/test split* to estimate the ability of the model to generalize to new data.

To better understand the role of each descriptive statistical feature, we performed an *ablation* study, i.e. dropping each feature from the feature set one by one. We also compared the performance of the two models to better understand their strengths and weaknesses. Finally, we built a stacking model to see how much the performance can be improved by combining the information gained from the two separate feature sets.

3.5 Descriptive statistical model (model 1)

At this point we worked with the original, unprocessed emails. First, we generated 29 descriptive statistical features which are often used for binary classification tasks. These features can be divided roughly into three categories to ease the model's interpretability. The first category consists of six variables that describe how elaborate or complex the response is: number of words, variability of punctuation, moving average type-token ratio (MATTR, see Covington & McFall, 2010), average length of sentences, average length of words, and frequency of punctuation. The eleven

[4] https://github.com/spacy-hu/spacy-hungarian-models.

[5] See the official website at xgboost.readthedocs.io, and Python's scikit-learn package, Pedregosa et al., 2011.

features in the second category describe the information density and structure of the responses. They are the ratio of *stop words* (we expect that the more general and uninformative an email is, the bigger is this ratio), and the ratio of *special tokens* (numbers, names, etc.) added to the responses (they describe the type and quantity of hard information mentioned in the emails). The third category consists of 12 features describing the linguistic structure of the responses. These are the ratios of different parts of speech (particles, proper nouns, pronouns, determiners, verbs, adpositions, conjunctions, nouns, auxiliaries, adverbs, adjectives and puntuation.). Since we used XGBoost models, we did not have to normalize our features.

In predictive modelling, it is desirable to reduce the number of predictors in order to improve the model's predictive power. Hence we performed a feature selection procedure; for more details, see the Supplementary Material. As a result, the most important nine features left in the feature set are:

Features describing the complexity and information density of the responses:

1. Number of words
2. Variation of punctuation
3. Ratio of numbers to all tokens

Features describing the linguistic structure of the responses:

4. Ratio of punctuation
5. Ratio of verbs
6. Ratio of proper nouns
7. Ratio of pronouns
8. Ratio of determiners
9. Ratio of adpositions

3.6 N-gram features based model (model 2)

Our second model used the text of the emails directly, or more precisely, it started with the n-grams that make up the text, and then assigned importance values to them, thus defining features that can be used in prediction. For this purpose, we used TF-IDF weighted (*term frequency—inverse document frequency*: high values are given to terms that are present in only a few documents) vectorization. We fitted separate vectorizers on the whole corpus and on the Roma and non-Roma *subcorpora*, i.e. responses written only to purportedly Roma or non-Roma clients separately. This way, we provided the model not just with the most frequent tokens in the corpus but also with the most frequent distinctive ones (i.e. the ones that are most distinctive between emails written to Roma or non-Roma). This allows us to build a relatively effective model despite the small data size. To construct the dictionary, we first selected the most frequent n-grams of each subcorpora (the number of the most frequent features that we selected was tuned during the hyperparameter search). Then we discarded those that are also amongst the most frequent n-grams in the other subcorpus (the number of the most frequent n-grams considered here is provided as a ratio of the feature selection number, and was tuned during the

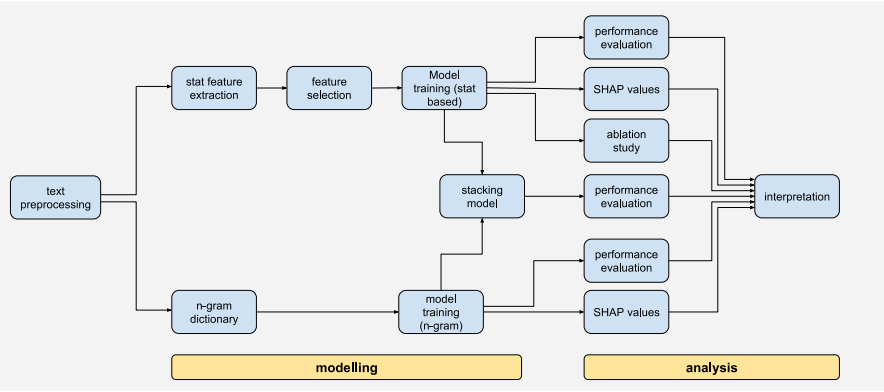

Fig. 1 The flow of the analytic process

hyperparameter search). Finally, we combined the two separate dictionaries with the most frequent n-grams in the whole corpus.

3.7 Stacking model (model 3)

To obtain an estimate for how much the combination of the two feature sets could improve the performance of the classifier, we also built a *stacking ensemble model* with a logistic regression algorithm (see 'stacking model', Fig. 1). This model is a logistic regression classifier that predicts the probability of an answer being written to a purportedly Roma requester based on the predicted probability of the two previously built models (hence the name 'ensemble'). The hyperparameters were set based on a cross-validation grid search carried out on the same training data on which the two base models were trained. It would have been preferable to train the ensemble model on a separate training set, but this was not feasible due to our sample size limit.

3.8 Model explanation

As mentioned in Sect. 3.3, predictive models often produce hard-to-interpret results, i.e. it is difficult to understand why the model makes a certain prediction ('black box models'), which is due to the fact that the models were developed to identify high dimensional underlying structure in the data, and, contrary to social science applications, the aim of a typical industrial research project is not to provide an interpretation but simply to find the optimal predictive model. We considered interpretability to be important as it can help us learn more about why and how a model is working, i.e. better understand the phenomenon at hand. New methods to interpret black-box machine learning models are being developed and published (for a summary, see Molnar, 2019). These are indirect methods in the sense that they do not (and cannot) give a direct meaningful relationship between the input and the output.

Table 1 Performance of the three models on the whole test set (200 responses)

	Model 1 (descriptive statistics)				Model 2 (n-gram)				Model 3 (stacking)			
	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc
Non-roma	0.56	0.71	0.63	58%	0.56	0.78	0.65	58.5%	0.59	0.72	0.65	61%
Roma	0.61	0.45	0.52		0.64	0.39	0.48		0.64	0.50	0.56	

To see how important the features individually are, we examined their SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017). SHAP assigns each feature an importance value for a particular prediction, which can be used to rank the features in order of importance, and to infer the nature of the effect of the feature based on the sign of the SHAP value.

To gain additional information on the importance of the features we also conducted a depletion study on the descriptive statistics model: during the study we left out each feature one-by-one to see how the performance of the models built on the remaining features would change. In addition to testing the importance of individual features, we also tested sets of features (ones describing the complexity and information density of the responses, and ones describing the linguistic structure of the responses) separately.

4 Results

4.1 Model performances

Table 1 presents the performance measures for the three models: precision, recall, F1-score, and accuracy (for the exact definition of these measures and more details see Eisenstein, 2019).

Precision for "Roma emails" is better than for "non-Roma emails" in the case of each model, i.e. there are fewer false positives among those predicted to be "Roma emails." On the contrary, *recall* is consistently greater in the case of the "non-Roma emails", which shows that the models are more capable of finding the relevant "non-Roma emails" in the corpus. By *accuracy*, both the descriptive statistics and the n-gram feature based models performed significantly better than a random classifier would have, with no significant difference between the accuracy score of the two models: 58% ($p=0.024$) and 58.5% ($p=0.016$) respectively. Only in 64.5% of the cases were the predictions of the two models the same. This suggests that the two feature sets grasp different sides of the differential treatment present in the emails.

4.2 The most important linguistic features

Figures 2 and 3 show the features' SHAP values for the descriptive statistics based model. Figure 2 describes the magnitude of the feature's effect (mean absolute

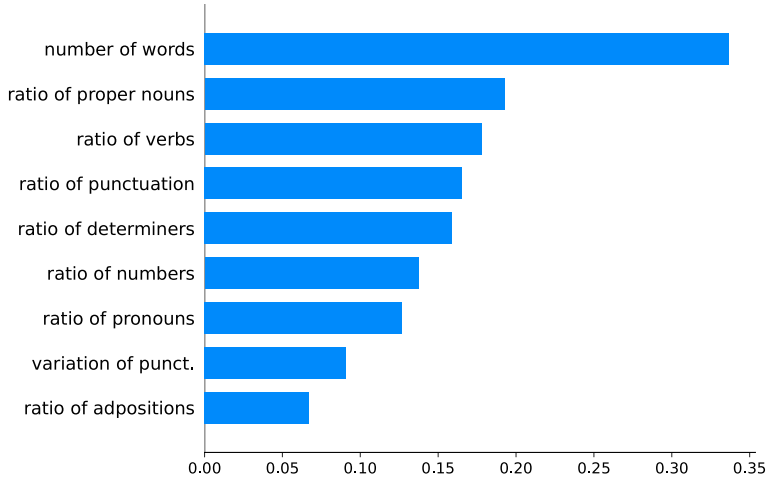

Fig. 2 Mean absolute SHAP values for each feature in the descriptive statistics model (feature importance)

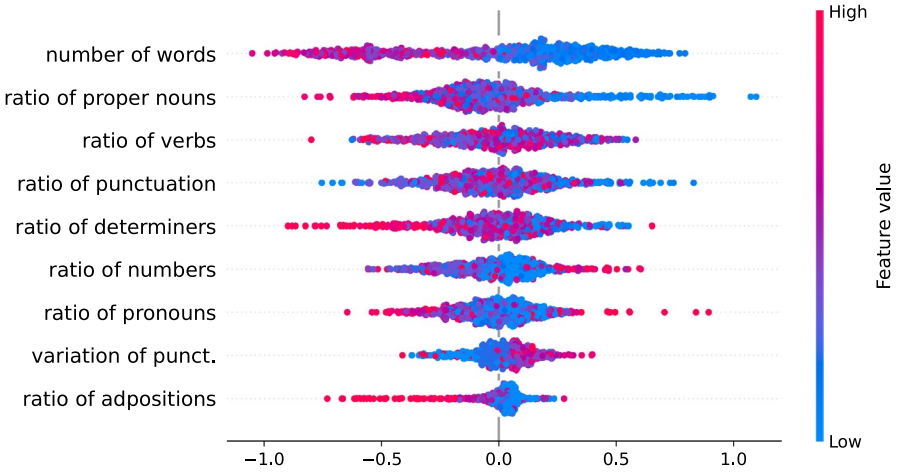

Fig. 3 SHAP values of each feature for each individual prediction. (Sign of the relationship of the features with the target variable)

values for each feature). The figure lists the most important variables in descending order, the top variables having the highest predictive power.

In addition to their importance, the sign of the relationship of the features with the target variable is also informative. Figure 3 presents signed SHAP values for each individual prediction when predicting the probability of an email being written to a purportedly Roma client. A dot in this graph represents the SHAP value for one feature in one individual prediction. The color scale indicates the feature values,

which range from low (blue) to high (red). A positive SHAP value (x axis) contributes to classifying the email as "Roma", while a negative one contributes to classifying it as "non-Roma". For example, if a feature's positive values are colored blue or its negative valued are colored red, then the feature's smaller values contribute to classifying the email as "Roma," and its larger values contribute to classifying the email as "not Roma", that is, the feature shows a negative correlation with "being Roma".

Based on these values, the length of a response is the most important feature (Fig. 2). According to Fig. 3, this feature has a negative effect, as its larger values are coloured blue: longer emails are less likely to be written to purportedly Roma clients. The ratio of proper nouns and the ratio of determiners have similar but smaller effects on the predictions, as does the ratio of adpositions. The ratio of number tokens and the variability of punctuation seems to have the opposite effect. These results suggest that when a purportedly Roma client receives an answer, it is often briefer, or even concise, with the information provided in fewer words.

The results of the depletion study can be seen in Table 2 in the Supplementary Material. We searched for the best hyperparameters for each feature set with separate cross-validation on the train set (see mean CV values averaged over folds in Table 2 in the Supplementary Material), and tested the performance of the best model on the test set (Test accuracy and Test F1 in the table).

According to these results, the model's performance is practically not affected by the dropping of the ratio of pronouns or the ratio of verbs. This suggests that these features mostly encode information that are present in other features. In accordance with our previous findings, without a feature describing the length of the reply, the model's performance is not much better than that of a random classifier (accuracy of 0.525). This is also true for the ratio of number tokens and the ratio of adpositions, which is slightly surprising since the ratio of adpositions has a relatively small mean absolute SHAP value. As we can see in Fig. 3, although most of the time this feature has a small impact on the predictions, there are several instances when it has a large effect on them. Therefore we can say that although this feature is only important in a few cases, in those few cases it contains information that is not present in other features. This is also backed up by its low weight value (the number of times a feature is used to split the data across all trees) in the full model. In the case of the other features (ratio of proper nouns, ratio of punctuation, ratio of determiners, and variability of punctuations), although they have a relevant impact on the model's performance, the model performs better even without them than a random classifier would.

Turning to the n-gram model: a relevant proportion of the most important tokens in terms of their SHAP-value can be sorted into three groups. (1) Greetings and salutations (for example.: 'respected' ('tisztelt', a more formal salutation than 'dear'), 'sir' ('uram'), 'madam' ('hölgyem'), 'inquirer' (' érdeklődő'), 'addressee' ('címzett'), and 'dear' ('kedves'); (2) official titles, which are mostly part of the signature, such as 'public notary' ('jegyző'), 'office' ('hivatal'), 'registrar' ('anyakönyvvezető'); and (3) reference to other information or contact information, such as 'on the internet' ('interneten'), 'at this phone number' ('telefonszámon'), 'contact information' ('elérhetőséget'), and 'on the phone' ('telefonon').

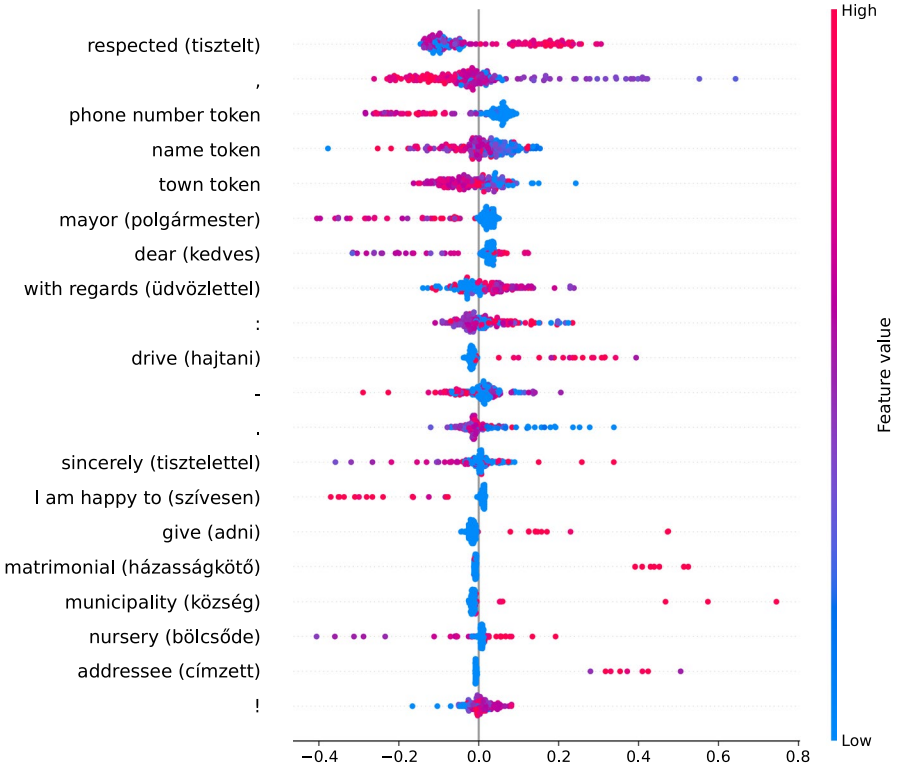

Fig. 4 SHAP values of the most important tokens in the n-gram model. The colours indicate the values of the features

Examining the sign of the SHAP values (Fig. 4), it seems that purportedly Roma clients received more formal and rather reserved replies: The presence of more 'respected' and 'addressee', and fewer name tokens, lead to a higher predicted probability of classifying the email as one written to a purportedly Roma client. (Whereas 'Tisztelt' + the name of the person being addressed is a frequently used formal salutation in Hungarian, 'Tisztelt címzett' translates something like 'respected addressee': It is more friendly to address somebody by using their name, or even with 'kedves'—Hungarian for 'dear'—plus their name.) This is also backed up by the fact that the token 'szívesen', which translates as 'I am happy to…', has a negative effect on the predicted probability of the email being classified as one addressed to a purportedly Roma client. Furthermore, some forms of addressing with which the writer of the email can avoid using the name of the client—'addressee' ('címzett'); 'madam' ('hölgyem'); 'inquirer' ('érdeklődő'))—come from the part of the dictionary that is distinctive to the emails written to purportedly Roma clients. The fact that many phone number tokens and many town name tokens affect a smaller predicted probability

also suggests that purportedly Roma clients received emails that contained less information.

4.3 Potential effect modifiers

To better understand where the differential treatment identified by the models is stronger, we examined the performance of the models by the gender of the client and the size of the town addressed by the client. The role of these potential effect modifiers can be examined by assessing the prediction for the subgroups separately. For this analysis we divided the towns into two groups, with a cut point defined by the median.

Both models perform better among men than women (7 and 18 percentage points better accuracy for descriptive statistical and n-gram models respectively) and in smaller towns than larger ones (4 and 15 percentage points better accuracy). This means that the differential treatment found by these models prevails stronger against men and in smaller municipalities. Table 3 in the Supplementary Material presents details of the models' performance by these background variables.

The models' better performance in smaller settlements and among men is even more obvious if we plot the population against the absolute error of each model. (See Figs. 5 and 6 in the Supplementary Material).

This result was also confirmed when we compared the two models according to their individual predictions on the test set. We found that in the test set of 200 emails there were 48 in which the two models gave the same wrong predictions: these may be considered as hard cases, where the discrimination was not present (or at least it was very subtle). The ratio of women among these was high (60%) as was the ratio of bigger settlements (20% is from the top decile). There were 81 cases where both models gave the correct predictions: these are probably cases where the differential treatment is the least subtle. Here the ratio of women was relatively low (42%) as was the ratio of large towns (5%).

5 Discussion

In our study we assessed the responsiveness of Hungarian local governments to requests for information by Roma and non-Roma clients relying on a nationwide correspondence study. The methodological novelty of our paper is that we showed that it is possible to detect discrimination in textual data in an automated way without human coding, and that machine learning may detect features of discrimination that coding guidelines do not cover. We employed natural language processing and machine learning techniques to automatically predict the ethnicity of the clients, based only on the text of the responses sent to them by public officials, and operationalized the measure of discrimination as the accuracy of this prediction.

Our first model took some of the descriptive statistical properties of the emails'into account, the second model worked with the texts themselves, and the third (ensemble) model combined the former two approaches. We detected a certain

level of *attention discrimination*: our models worked significantly better compared to the random classification, confirming the differential treatment of Roma clients. We found that even our first model, with only a few variables, was able to capture discrimination.

The accuracy of the best model was only 61%, implying that unequal treatment of Roma clients is not rampant in Hungary. In our previous paper (Simonovits et al., 2021), we used a Heckman selection model with randomly assigned follow-up as an excluded predictor of the selection equation when comparing "Roma" and "non-Roma emails" based on human coding. At that time we found a significant and moderate size of discrimination both regarding politeness and information content: a 6.4 [3.3–9.4] and a 5.2 [2.5–7.8] point difference on a 100-point scale, respectively. When comparing our present results with those of our previous study, a similar conclusion arises: there is empirical evidence for the differential treatment of Roma clients, but we cannot conclude that there is rampant discrimination.

Our models consistently have shown that *precision* for Roma clients is better than for their non-Roma counterparts, while *recall* is greater in the case of the non-Roma emails. This can be interpreted as such that the models detected linguistic features that—if used by officials—are used mainly in emails to Roma clients but by no means in all emails written to Roma clients. However, the better *recall* for "non-Roma" as opposed to "Roma emails" may also be partly due to the higher a priori probability of a response being given to a non-Roma than to a Roma client, which is reflected in the unbalanced training set. The role of potential modifiers was also examined by assessing the prediction for certain subgroups separately. Higher levels of discrimination were detectable against male senders and in smaller settlements. The plausible explanation for this might be that (1) social norms prescribing behaviour towards women may override ethnic discrimination, and that (2) administration is more standardized in larger settlements.

In addition to *measuring* subtle forms discrimination, we also aimed at *understanding* the linguistic distinctions we detected. Hence we tried to identify linguistic features that the algorithm considers most important when distinguishing between responses written to Roma or non-Roma clients. These features may include a number of features that human coding would either not take into account, or that are outside the researchers' horizon. In other words: our algorithmic procedure to find distinctive linguistic patterns that distinguish emails written to Roma or non-Roma clients can be considered an inductive one compared to the deductive method based on human coding, which makes its objectivity more justified from this point of view. The analysis of the linguistic features showed that the answers sent to purportedly Roma clients are not only shorter, but their tone is rather reserved and less polite. It is important to note that the discrimination we found is not overtly negative: an email with a rather reserved and distant wording is perhaps showing signs of overcompensation that the official is using to avoid the appearance of discrimination.

From a conceptual point of view we conclude that subtle forms of discrimination emerging from the texts of replies may be best interpreted in the framework of *attention discrimination* (Bartos et al., 2016), meaning that officials at local governments pay less attention (with shorter and less polite answers) to purportedly Roma clients. As the unequal treatment of clients was detected on the digital platforms

of administration, we may also call this kind of behaviour as *digital discrimination,* following Edelman and Luca (2014). Our results support findings by Edelman et al. (2017) and Cui et al. (2020), according to whom discrimination remains an important policy concern not only in online marketplaces, but in online administration processes. When it comes to policy relevance, we may assume that public awareness of the EU directives have an effect on the behaviour of public bodies. Our recent research (see the results of Simonovits et al., 2021) showed that cooperation between NGOs and scholars is an important and effective tool to reduce discriminatory behaviour by local governments. In the present paper we have shown that, based on the applied NLP methods, differences in linguistic style and pattern imply that subtle forms of digital discrimination (as opposed to more overt manifestations) still exist in Hungary, similarly to other European countries (see Adman & Jansson, 2017; Ahmed & Hammarstedt, 2019 and Hemker & Rink, 2017) and to the US (Einstein, 2017; Giulietti, 2015), where discrimination is legally banned. In fostering the work of anti-discrimination bodies (e.g. governmental bodies devoted to promote equal opportunities, human rights NGOs, and private sector actors) to uncover the more covert forms of discrimination, introducing experimental methods may prove to be useful in uncovering discriminatory selection mechanisms. Rorive (2009) pointed out that after the adoption of the above mentioned EU Directives, there were certain member states—primarily France, Belgium, and Hungary—where situation-testing (based on field experimental methods) has become a recognised tool by the courts as a potential means of evidence. The report highlights that proving discrimination in court as well as in various spheres of social life remains a major challenge, as in many cases there is no clear evidence for unequal treatment.

Our analysis also has its limitations. (1) Any correspondence study is negatively impacted by a certain level of nonresponse rate. Comparisons of response quality conditionals on observing a response may lead to post-treatment bias (Coppock et al., 2019). We found a higher non-response rate towards Roma clients (Simonovits et al., 2021). Based on this, we may assume that most officials harboring anti-Roma attitudes tend not to write a response, so an under-estimation rather than an over-estimation of discrimination is more likely in our data. (2) What a computational text analytic procedure can detect is differential treatment, but not necessarily negative discrimination. We tried to approach this issue by detecting which text patterns play the most important role in this differential treatment, but this limitation has to be taken into account when interpreting the results. (3) Since the analysis of potential effect modifiers is based on a relatively small test set, and are not separately tested on an independent set, our findings should be considered exploratory, with further research needed to provide conclusive results. In this vein, it is important to note, that with the application of NLP technique we are only able to assess *subtle forms of discrimination* (in contrast *to direct forms* of discrimination which can be mostly identified by ignorance of clients).

To the best of our knowledge, our study is the first attempt to assess discrimination using ML techniques. According to our findings, ML is a suitable tool to detect subtle forms and ways of digital discrimination. As textual data also occur in many other areas, our methods can be generalized to domains other than public services, e.g. to the labour market (e.g. in the selection of future employees), to the housing

market (e.g. in the provision of loans), in education (in the admissions process), or to the academy (e.g. in the scholarly reviewing processes). We believe that our study will contribute to this new research direction.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10579-022-09612-5.

Acknowledgements The authors would like to thank Adam Vig, Peter Hobot and Gabor Csomor for their contribution to data collection and Endre Sik for his valuable comments.

Author contributions JB Methodology, Software, Visualization, Formal Analysis, Writing—Original Draft. RN Conceptualization, Methodology, Formal Analysis, Writing—Original Draft. BS Conceptualization, Investigation, Writing—Original Draft. GS Conceptualization, Investigation, Writing—Original Draft.

Funding Open access funding provided by Eötvös Loránd University. JB's work was supported by the Higher Education Excellence Program of the Ministry of Human Capacities (ELTE–FKIP). RN's work was supported by NKFIH (National Research, Development and Innovation Office, Hungary) grant K-134428. BS's work was supprted NKFIH (National Research, Development and Innovation Office, Hungary) grant FK-127978. Our funding sources had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Data availability In the present paper we carry out a secondary analysis of our audit study (Simonovits, G., Simonovits, B., Vig, A., Hobot, P., Nemeth, R., & Csomor, G.2021. Back to "normal": The short-lived impact of an online NGO campaign of government discrimination in Hungary. Political Science Research and Methods, 1–9. https://doi.org/10.1017/psrm.2021.55). Data is available at https://doi.org/10.7910/DVN/KPSCLK. The original audit study was based on qualitative coding of the e-mails, the present secondary analysis processes the e-mails in an automated way without using human coding. This is the essence of the current research: we show that it is possible to detect discrimination in textual data in an automated way without human coding, and that machine learning may detect features of discrimination that human coders may not recognize.

Code availability The codes used in the present paper are available at this (temporary) anonymous Github repository: https://github.com/langdiscr/langdiscr.

Declarations

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval The audit study received IRB clearance and was compliant with relevant Hungarian law. We debriefed the research subjects soon after the data collection phase was completed (October, 2020).

References

Adman, P., & Jansson, H. (2017). A field experiment on ethnic discrimination among local Swedish public officials. *Local Government Studies, 43*(1), 44–63. https://doi.org/10.1080/03003930.2016.1244052

Ahmed, A., & Hammarstedt, M. (2019). Ethnic discrimination in contacts with public authorities: A correspondence test among Swedish municipalities. *Applied Economics Letters, 27*(17), 1391–1394. https://doi.org/10.1080/13504851.2019.1683141

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review, 106*(6), 1437–1475. https://doi.org/10.1257/aer.20140571

Bayram, U., Pestian, J., Santel, D., & Minai, A. A. (2019). What's in a word? Detecting partisan affiliation from word use in congressional speeches. In *2019 International Joint Conference on Neural Networks* (IJCNN) (pp. 1–8). IEEE.

Bertrand, M., & Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? A field experiment on labor market discrimination. *American Economic Review, 94*, 991–1013. https://doi.org/10.1257/0002828042002561.

Bohren, A., Imas, A., & Rosenberg, M. (2018). The language of discrimination: Using experimental versus observational data. *AEA Papers and Proceedings, 108*, 169–174. https://doi.org/10.1257/pandp.20181099

Boulis, C., & Ostendorf, M. (2005) A quantitative analysis of lexical differences between genders in telephone conversations. In Din R (eds). *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL'05). Association for Computational Linguistics.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–215.

Butler, D. M., & Crabtree, C. (2017). Moving beyond measurement: Adapting audit studies to test bias-reducing interventions. *Journal of Experimental Political Science, 4*, 57–67.

Coppock, A. (2019). Avoiding post-treatment bias in audit experiments. *Journal of Experimental Political Science, 6*, 1–4.

Cortina, L. M. (2008). Unseen injustice: Incivility as modern discrimination in organizations. *The Academy of Management Review, 33*, 55–75. https://doi.org/10.2307/20159376

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17*, 94–100. https://doi.org/10.1080/09296171003643098

Crabtree, C. (2018). An introduction to conducting email audit studies. In S. Gaddis (Ed.), *Audit studies: Behind the scenes with theory method and nuance*. Springer. https://doi.org/10.1007/978-3-319-71153-9_5

Csomor, G., Simonovits, B., & Németh, R. (2021). Hivatali diszkrimináció?: Egy online terepkísérlet eredményei (discrimination at local governments? Results of an online field experiment. *Szociológiai Szemle, 31*(1), 4–28. https://doi.org/10.51624/szocszemle.2021.1.1

Cui, R., Li, J., & Zhang, D. J. (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Science, 66*(3), 1071–1094. https://doi.org/10.1287/mnsc.2018.3273

Dipboye, R. L., & Halverson, S. K. (2004). Subtle (and not so subtle) discrimination in organizations. *The Dark Side of Organizational Behavior, 16*, 131–158.

Distelhorst, G., & Hou, Y. (2014). Ingroup bias in official behavior: A national field experiment in China. *Quarterly Journal of Political Science, 9*(2), 203–230. https://doi.org/10.1561/100.00013110

Duflo, E., & Banerjee, A. V. (2017). *Handbook of economic field experiments*. North-Holland.

Edelman, B. G. & Luca, M. (2014). Digital discrimination: The case of Airbnb.com. *Harvard business school NOM unit working paper*. http://dx.doi.org/https://doi.org/10.2139/ssrn.2377353

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics, 9*(2), 1–22. https://doi.org/10.1257/app.20160213

Einstein, K. L., & Glick, D. M. (2017). Does race affect access to government services? An experiment exploring street-level bureaucrats and access to public housing. *American Journal of Political Science, 61*(1), 100–116. https://doi.org/10.1111/ajps.12252

Eisenstein, J. (2019). *Introduction to natural language processing (adaptive computation and machine learning series)*. The MIT Press.

Enyedi, Z., Fábián, Z., & Sik, E. (2004). Nőttek-e az előítéletek Magyarországon? Antiszemitizmus, cigányellenesség és xenofóbia változása az elmúlt évtizedben. In T. Kolosi, I. . Gy. . Tóth, & Gy. Vukovich (Eds.), *Társadalmi riport*. TÁRKI.

Fundamental Rights Agency (FRA). (2018). A persisting concern: Anti-Gypsyism as a barrier to Roma inclusion. Report. Luxembourg: Publications Office of the European Union. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-antigypsyism-barrier-roma-inclusion_en.pdf

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica, 87*(4), 1307–1340.

Giulietti, C., Tonin, M., & Vlassopoulos, M. (2015). Racial discrimination in local public services: A field experiment in the US. *CESifo working paper*, (No. 5537). Center for Economic Studies and Ifo Institute (CESifo). https://ssrn.com/abstract=2681054.

Giulietti, C., Tonin, M., & Vlassopoulos, M. (2019). Racial discrimination in local public services: A field experiment in the United States. *Journal of the European Economic Association, 17*(1), 165–204. https://doi.org/10.1093/jeea/jvx045

Green, J., Edgerton, J., Naftel, D., Shoub, K., & Cranmer, S. J. (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances, 6*(28), eabc2717.

Hemker, J., & Rink, A. (2017). Multiple dimensions of bureaucratic discrimination: Evidence from German welfare offices. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12312

Huang, B., Li, J., Lin, T. C., Tai, M. & Zhou, Y. (2021). *Attention discrimination under time constraints: Evidence from retail lending*. https://doi.org/10.2139/ssrn.3865478

Jones, K., Arena, D., Nittrouer, C., Alonso, N., & Lindsey, A. (2017a). Subtle discrimination in the workplace: A vicious cycle. *Industrial and Organizational Psychology, 10*(1), 51–76. https://doi.org/10.1017/iop.2016.91

Jones, K. P., Sabat, I. E., King, E. B., Ahmad, A., McCausland, T. C., & Chen, T. (2017b). Isms and schisms: A meta-analysis of the prejudice-discrimination relationship across racism, sexism, and ageism. *Journal of Organizational Behavior, 38*(7), 1076–1110.

Kalla, J. L., & Porter, E. (2019). Correcting bias in perceptions of public opinion among American elected officials: Results from two field experiments. *British Journal of Political Science, 51*(4), 1792–1800.

Kende, A., Tropp, L., & Lantos, N. A. (2017). Testing a contact intervention based on intergroup friendship between Roma and non-Roma Hungarians: Reducing bias through institutional support in a non-supportive societal context. *Journal of Applied Social Psychology, 47*(1), 47–55. https://doi.org/10.1111/jasp.12422

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Luxemburg, U. V., Guyon, I., Bengio, S., Wallach, H., & Fergus, R. (eds.) *Proceedings of the 31st international conference on neural information processing systems (NIPS'17)*. (pp. 4768–4777) Curran Associates Inc.

Massey, D. S., & Lundy, G. (2001). Use of black English and racial discrimination in urban housing markets. New methods and findings. *Urban Affairs Review, 36*(4), 452–469.

Matejka, F. (2013). Attention discrimination: Theory and field experiments. In *2013 meeting papers,* (No. 798). Society for Economic Dynamics. https://ideas.repec.org/p/red/sed013/798.html

Miller, J., Gounev, P., Pap, A. L., Wagman, D., Balogi, A., Bezlov, T., Simonovits, B., & Vargha, L. (2008). Racism and police stops: adapting US and British debates to continental Europe. *European Journal of Criminology, 5*(2), 161–191.

Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable*. Leanpub (eBook). https://christophm.github.io/interpretable-ml-book/

Morton, S., Zettelmeyer, F. F., & Silva-Risso, J. (2003). Consumer information and discrimination: Does the internet affect the pricing of new cars to women and minorities? *Quantitative Marketing and Economics, 1*(1), 65–92. https://doi.org/10.1023/A:1023529910567

Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *The Journal of Human Resources, 47*(4), 1128–1157. https://doi.org/10.3368/jhr.47.4.1128

Örkény, A., & Váradi, L. (2010). Az előítéletes gondolkodás társadalmi beágyazottsága, nemzetközi összehasonlításban. *Alkalmazott Pszichológia, 12*(1–2), 29–46.

Pálosi, E., Sik, E., & Simonovits, B. (2007). Discrimination in shopping centers. *Szociológiai Szemle, 3*(17), 135–148.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., & Perrot, M. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*(85), 2825–2830.

Rorive, I. (2009). *Proving discrimination cases-the role of situation testing.* MPG and the Centre for Equal Rights. https://www.migpolgroup.com/_old/public/docs/153.ProvingDiscriminationCases_theroleofSituationTesting_EN_03.09.pdf

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Sik, E., & Simonovits, B. (2008). Egyenlő bánásmód és diszkrimináció. In T. Kolosi & I. . Gy. . Tóth (Eds.), *Társadalmi riport 2008.* TÁRKI.

Sik, E., Simonovits, B., & Szeitl, B. (2016). Az idegenellenesség alakulása és a bevándorlással kapcsolatos félelmek Magyarországon és a visegrádi országokban. *REGIO, Kisebbség Kultúra Politika Társadalom, 24*(2), 81–108.

Simonovits, B., & Surányi, R. (2020). The Jews are just like any other human being. *Intersections*. https://doi.org/10.17356/ieejsp.v5i4.575

Simonovits, B., & Szalai, B. (2013). Idegenellenesség és diszkrimináció a mai Magyarországon. *Magyar Tudomány, 3*(March), 251–262.

Simonovits, G., Kezdi, G., & Kardos, P. (2018). Seeing the world through the other's eye: An online intervention reducing ethnic prejudice. *The American Political Science Review, 112*(1), 186–193. https://doi.org/10.1017/S0003055417000478

Simonovits, G., Simonovits, B., Víg, Á., Hobot, P., Csomor, G., & Németh, R. (2021). Back to normal: The short-lived impact of an online NGO campaign of government discrimination in Hungary. *Political Science Research and Methods*. https://doi.org/10.1017/psrm.2021.55

Váradi, L. (2012). Preliminary study on the selection of Roma surnames for discrimination testing Előtanulmány a roma családnevek diszkriminációteszteléséhez való kiválasztásához. Sík E. & Simonovits B.(Szerk.) (Measuring Discrimination) A diszkrimináció mérése, 236–244. https://www.tarki.hu/hu/about/staff/sb/Diszkriminacio_merese.pdf

Verhaeghe, P. P. (2022). Correspondence studies. In K. F. Zimmermann (Ed.), *Handbook of labor, human resources and population economics.* Springer.

Yinger, J. (1998). Evidence on discrimination in consumer markets. *Journal of Economic Perspectives, 12*(2), 23–40. https://doi.org/10.1257/jep.12.2.23

Zussman, A. (2013). Ethnic discrimination: lessons from the Israeli online market for used cars. *The Economic Journal, 123*(572), F433–F468. https://doi.org/10.1111/ecoj.12059

Authors and Affiliations

Jakab Buda[1] · Renáta Németh[1] · Bori Simonovits[2] · Gábor Simonovits[3,4,5]

Jakab Buda
bakajb@gmail.com

Bori Simonovits
simonovits.borbala@ppk.elte.hu

Gábor Simonovits
simonovitsg@ceu.edu

[1] Research Center for Computational Social Science, Faculty of Social Sciences, Eötvös Loránd University, Pázmány Péter sétány 1/a, Budapest 1117, Hungary

[2] Faculty of Education and Psychology, Eötvös Loránd University, Izabella utca 46, Budapest 1064, Hungary

[3] Associate Professor, Department of Political Science, Central European University, Budapest, Hungary

[4] Co-director for Academic Affairs, Rajk Laszlo College for Advanced Studies, Budapest, Hungary

[5] Senior Researcher, Institute for Political Science, Budapest, Hungary