Journal of Computational Social Science (2023) 6:589–608
https://doi.org/10.1007/s42001-023-00205-y

RESEARCH ARTICLE

Predicting perceived ethnicity with data on personal names in Russia

Alexey Bessudnov[1] · Denis Tarasov[2] · Viacheslav Panasovets[3] · Veronica Kostenko[4] · Ivan Smirnov[5] · Vladimir Uspenskiy[6]

Received: 20 December 2022 / Accepted: 13 March 2023 / Published online: 4 April 2023

Abstract

In this paper, we develop a machine learning classifier that predicts perceived ethnicity from data on personal names for major ethnic groups populating Russia. We collect data from VK, the largest Russian social media website. Ethnicity was coded from languages spoken by users and their geographical location, with the data manually cleaned by crowd workers. The classifier shows the accuracy of 0.82 for a scheme with 24 ethnic groups and 0.92 for 15 aggregated ethnic groups. It can be used for research on ethnicity and ethnic relations in Russia, with the data sets that have personal names but not ethnicity.

Keywords Ethnicity · Russia · Machine learning · Prediction · Personal names

Introduction

Over the past decades, social scientists gained access to many large-scale data sets thanks to the proliferation of digital traces [1]. The explosive growth in new data even raised hopes that social science was entering its golden age [2]. However, digital traces are typically not collected with a research purpose in mind and are framed by the needs of data providers. As a result, they often lack information on individuals that is important to researchers. One potential solution is to

✉ Alexey Bessudnov
a.bessudnov@exeter.ac.uk

1 Social and Political Sciences, Philosophy and Anthropology, University of Exeter, Exeter, UK

2 Computer Science, Constructor University Bremen, Bremen, Germany

3 Applied Mathematics and Control Processes, St Petersburg State University, St Petersburg, Russia

4 Sociology, European University at St Petersburg, St Petersburg, Russia

5 Computational Social Sciences and Humanities, RWTH Aachen University, Aachen, Germany

6 Digital Transformation, ITMO University, St Petersburg, Russia

infer missing information using machine learning methods. For example, various socio-demographic characteristics were predicted from profile images [3], mobile phone metadata [4], Facebook likes [5], and images of street scenes [6].

One of important characteristics that is of great interest to social scientists but rarely present in digital traces is ethnicity. Taking ethnicity into account is important for analysing social inequalities in health [7], political participation [8], the labour market and housing [9], among other areas.

While lacking information on ethnicity, some large-scale data sets have not been anonymised and include personal names. Examples of such data sets include US voter registration data [10] or Twitter data [11]. Personal names can be used as a signal for ethnicity for many ethnic groups. Experimental studies of discrimination in the labour market and in housing have been using this feature [9]; it was also applied to historic studies of social mobility [12]. An ability to infer ethnicity from personal names allows social scientists to use new administrative and social media data.

While several ethnicity classifiers are already available to researchers, most of them are focusing on a few immigration destination countries and are limited to a small number of ethnic groups [13]. In this paper, we are addressing this gap by developing a machine learning approach to coding perceived ethnicity from personal names for ethnic groups populating Russia, using data from VK, the largest Russian social media website.

There are several approaches to classifying ethnicity with data on personal names. Early studies employ a dictionary-based method where names were matched to a reference list of names already classified by ethnicity. [14] is perhaps the first example of automatic name binary classification, developed to separate Chinese from non-Chinese names in Canada. [15] offers a review of 13 studies published up to 2007 that use similar methodology. In a more recent application, [16, 17] use both matching and supervised learning to classify ethnicity of Facebook users in the Netherlands on the basis of their fist names, using the Dutch census as the reference list (also see [18]).

A successful application of the matching approach requires a large reference list that covers most ethnic first names and/or surnames. However, for many ethnic groups, the reference lists of names do not exist or are incomplete. These classifiers also rely on the assumption that name/ethnicity distributions are similar for the reference list and the target population. As reference lists are often compiled from census data, it is not clear how well they will work with social media data.

Another approach to ethnic name classification is based on supervised learning algorithms. The main advantage of this approach is that it allows researchers to classify previously unseen names. [19] develop a multiclass name classifier for 13 ethnic groups with the data from Wikipedia using hidden Markov models and decision trees. [20] use recurrent neural networks to predict ethnicity from names from the Olympic records data. In other recent studies, [21] apply several machine learning algorithms to infer religion from personal names in South Asia, [22] develop a multiclass classifier for 39 nationalities, and [11] predict gender and ethnicity from Twitter usernames (see other examples in [23]).

None of the existing studies specifically focuses on Russian names. The most well-known reference list of Russian surnames [24] is incomplete and does not directly link surnames to ethnic groups. [25] develop a method for identifying ethnically Russian surnames that uses suffix-based morphological regularities. ML-based methods can achieve higher accuracy, and, besides, the method proposed in [25] cannot distinguish between various ethnic groups populating Russia that is home to over 100 ethnic groups with often characteristic personal names.

There are many papers already that use supervised learning and data matching to code ethnicity from personal names. The novelty of this paper is that for the first time, we create a publicly available tool that can be applied to Russian data (using the supervised learning approach that is consistent with but different from previously developed methods). This tool will be useful for researchers of ethnicity and ethnic inequalities in Russia. For an example of the application of the tool, see [26] that explores ethnic inequalities in the Russian military fatalities in the war in Ukraine.

Data and methods

Data collection and processing

We use data from VK (www.vk.com), a Russian social media website. VK was created in 2006 as a clone of Facebook and quickly became the most popular Russian social networking website. In December 2020, its user base in Russia consisted of 73 million people. According to the VK Terms of Service, users understand and accept that information that they publish on their page becomes publicly available on the Internet. VK provides a public application programming interface (API) that allows downloading this information systematically in the open JSON format. In particular, it is possible to download user profiles from a selected region or VK community and access information on personal names and languages spoken by users. VK has been shown to be a valuable source of data for social science research [27, 28].

VK does not directly collect information on user ethnicity. To infer ethnicity, we combine information on users' locations and the languages they speak, improving the quality of inference by manual checks via crowdsourcing. We use the resulting data on personal names (first name and surname) and inferred ethnicity as an input for machine learning (ML) algorithms. We apply the following protocol for collecting data and coding ethnicity.

First, we compile a list of 40 ethnic groups that, according to the 2010 Russian census, count more than 100,000 people. We exclude 9 groups in cases where either personal names are almost indistinguishable from ethnic Russian (Chuvash, Mordvin, Udmurt, Mari, Komi) or where it is not possible to assign ethnicity using the combination of the language spoken and location (Germans, Koreans, Roma, Turks). For the remaining ethnic groups, we collect data on names and sex from user profiles in the cities where these groups are geographically concentrated and from thematic ethnic communities. This is facilitated by the fact that many ethnic groups in Russia have their "titular" regions where most of their members live (such as Chechnya for Chechens, Tatarstan for Tatars, etc.).

At the next stage, we filter the data by the language spoken, only keeping the profiles of people who indicate that they can speak the language of the ethnic group they are intended to represent. Thus, someone who lives in Kazan (the capital of the Republic of Tatarstan) and can speak Tatar is assumed to be ethnically Tatar. At this stage, we combine together the ethnic groups who share the same language (Kabardin and Adyghe; Karachay and Balkar) or the ethnic groups with similar personal names who share the same locations (the Avar, Dargin, Kumyk, Lezgian, Laki, Tabasaran, and Nogai into the Dagestani).

Then, we manually clean the data at Yandex.Toloka, a crowdsourcing platform similar to Amazon Mechanical Turk. For most ethnic groups, we employ data cleaners from locations where the group is geographically concentrated. We ask the data cleaners to select only the names that belong to required ethnic groups. To improve data quality, we implement several quality control checks. Our aim is to collect about 10,000 personal names for each ethnic group, although in some cases, this is not possible.

In the resulting data set, some of the names are spelled in Cyrillic and others in Latin alphabet. We transliterate all the names to Cyrillic using the transliterate package in Python. We make some manual adjustments, such as replacing the Ukrainian letter 'i' with the Russian 'и'. Then, we remove all the names containing non-Cyrillic characters other than '-' and concatenate first names and surnames with the '#' delimiter. The final sample consists of 172,280 names for 24 ethnic groups.

Table 1 shows the list of ethnic groups and their population and sample sizes.

Machine learning pipeline

To apply ML algorithms, text must be transformed into numerical vectors. We use three different vectorisation methods (Bag of Words, TFIDF, and fastText) and compare their performance with different ML algorithms.

The Bag of Words (BoW) converts text into a vector with dimensionality equal to the size of the vocabulary formed with unique tokens (extracted n-grams in our case) from a corpus. n-gram is a sequence of n characters from a name: for example, 3-grams of the name ≪Alice≫ are ≪Ali≫, ≪lic≫, ≪ice≫. Vectorisation is then performed according to the token (n-gram) frequency. As an example, if a corpus contains only tokens 'A', 'T', 'G', 'C', then ≪AATGA≫ would be converted to $< 3, 1, 1, 0 >$.

TFIDF (term frequency-inverse document frequency) shares the same idea, but it uses the tf-idf function of a token, i.e., normalise the token frequency by the share of all words that contain the token. The motivation for this transformation is to decrease the impact of frequent tokens that often provide little information and to increase the impact of rare tokens that are more informative [29].

Finally, the fastText model (FT) [30] is a method that transforms words into a vector of real values (so-called word embeddings) using n-grams. It was trained on a large corpus to efficiently represent words as vectors. We pass the first names and surnames independently through the model that was trained on Russian texts

Table 1 Ethnic groups and their population and sample sizes

Ethnic group	Data source (cities)	Population size (2010 census, thousand)	Sample size
Ethnic Russian	Tambov, Vladimir, Vologda	111,000	11,879
Tatar	Kazan	5300	9862
Dagestani*	Makhachkala, Khasavyurt Derbent, Kaspiysk	2900	9555
Ukrainian	VK Ukraine	1900	8377
Bashkir	Ufa	1600	13,462
~~Chuvash~~	Not selected	1400	
Chechen	Grozny, Urus-Martan, Gudermes	1400	5257
Armenian	Yerevan	1200	9269
~~Mordvin~~	Not selected	740	
Kazakh	Nur-Sultan	650	9733
Adyghe / Kabardin	Nalchik, Baksan, Nartkala, Terek Chegem, Maykop, Adygeysk	640	1240
Azerbaijani	Baku	600	7922
~~Udmurt~~	Not selected	550	
~~Mari~~	Not selected	550	
Ossetian	Vladikavkaz, Mozdok, Beslan	530	4834
Belarusian	Minsk	520	13,393
Yakut	Yakutsk	480	1604
Buryat	Ulan-Ude	460	7691
Ingush	Nazran, Sunzha, Karabulak	440	1315
~~German~~	Not selected	390	
Balkar / Karachay	Cherkessk, Ust-Dzheguta Karachaevsk, Nalchik, Tyrnauz	331	1264
Uzbek	Tashkent	290	8709
Tuvan	Kyzyl	260	3556
~~Komi~~	Not selected	230	
~~Roma~~	Not selected	200	
Tajik	Dushanbe	200	10,636
Kalmyk	Elista	180	1745
Georgian	Tbilisi	160	9306
Jewish	Tel-Aviv, Jerusalem, Haifa	160	4054
Moldovan	Kishinev	160	8059
~~Korean~~	Not selected	150	
~~Turkish~~	Not selected	105	
Kyrgyz	Bishkek	100	9558

Notes: The Dagestani include the Avar, Dargin, Kumyk, Lezgian, Laki, Tabasaran, and Nogai ethnic groups. For Ukrainians, we only use the data from the largest Ukrainian VK community. Data from VK communities are used for some other ethnic groups as well. Some of the cities in the table are outside of Russia

(https://fasttext.cc/docs/en/crawl-vectors.html) and then concatenate the pairs of vectors resulting in the vectors of dimensionality 600.

We apply several ML algorithms and compare their performance. These are complement Naive Bayes (CNB) and several versions of the Stochastic Gradient Descent (SGD) classifier: with the Log loss (LR, equivalent to logistic regression), with the Hinge loss (SVM, equivalent to linear Support Vector Machine), and with the modified Huber loss (MH, equivalent to quadratically smoothed Support Vector Machine) [31], as well as the Gradient Tree Boosting (GB) [32]. The CNB and SGD models are implemented with scikit-learn [33], and GB with XGBoost [34]. We also implement several other approaches such as the Random Forest (RF), bidirectional Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), and the one-dimensional Convolutional neural network (CNN). They perform worse with our data, and we only report the results from the five best algorithms. The results for other algorithms can be found in Appendix B (Table 8). All the hyperparameters can be found in Appendix A.

We optimise the model hyperparameters with 3-fold cross-validation on the train data set (75% of the data), with F1 as the target metric. The model with the best hyperparameters is then evaluated on the test set (25% of the data). To prevent data leakage, we remove from the test set the names that are also present in the train set.

Results

Detailed ethnic classification

Table 2 reports the results for the five ML algorithms implemented with different vectorisation techniques, compared with a baseline random classifier that predicts ethnicity with a probability proportional to its frequency in the training set. We report four metrics: accuracy, precision, recall, and F1. Accuracy is the proportion of correctly predicted names. Precision is the fraction of true positives among all positives (i.e., out of all names predicted to be ethnically Russian how many are actually ethnically Russian?). Recall is the fraction of true positives among true positives and false negatives (i.e., out of all ethnically Russian names how many did we label correctly?). F1 is a weighted average of precision and recall that provides a single measure of prediction accuracy (Table 2).

With our data, the modified Huber (MH) model with TFIDF vectorisation shows the best fit and correctly classifies 82% of the names in the test set (see Table 2).

Figure 1 shows the confusion matrix for 24 ethnic groups based on the MH model. Table 3 shows the prediction metrics for each ethnic group. Prediction accuracy varies by group, from precision as high as 0.99 for Armenians and Georgians (two groups with very characteristic names that follow a simple pattern) to 0.68 for Tatars, 0.69 for Dagestani, and 0.71 for ethnic Russians. Recall is lowest for Karachay/Balkar (0.61) and Kabardin / Adyghe (0.63), often classified as other North Caucasian groups, and Belarusians (0.65), often classified as ethnic Russians.

Can further improvements in prediction accuracy be made if we increase the sample size? Figure 2 demonstrates how the prediction metrics change depending on

the number of names in the training set. The steepest increase in accuracy occurs up to the point where we have approximately 50,000 – 60,000 names (i.e., about 2500 names per group on average). Beyond this sample size, the improvements are modest. We conclude that the training set with about 130,000 names is sufficient for classification purposes.

Aggregated ethnic classification

Figure 1 shows that there is a pattern in the classification errors for several ethnic groups. For example, Tatar and Bashkir names often get confused by the algorithm. This can be explained by the characteristics of our data rather than by deficiencies of the classification tool. Indeed, Tatars and Bashkirs both are Turkic groups populating the Volga region who share common origins and culture. Historically, the boundaries between the Tatar and Bashkir identities were not always clear [35]. Some other ethnic groups in the data set also often share many common names.

For many social research questions, the classification we developed is too detailed and a schema with a smaller number of aggregated ethnic groups would be preferable. At the next step of the analysis, we merge several ethnic groups together. We take into account the confusion matrix (Fig. 1), as well as the historical and cultural factors and likely applications of the classification tool in social science research. We combine together the following groups: (1) Tatars and Bashkirs (two Turkic groups populating the Volga region), (2) ethnic Russians, Belarusians, and Ukrainians (eastern Slavic groups), (3) Chechens, Dagestanis, and Ingushes (ethnic groups populating the Eastern part of the North Caucasus), (4) Kabardins, Adyghe, Karachays, Balkars, and Ossetians (ethnic groups populating the Western part of the North Caucasus), (5) Kazakhs and Kyrgyz (two Central Asian groups with nomadic origins), (6) Tajiks and Uzbeks (two Central

Table 2 Model performance on the test set

Algorithm	Vectorisation	Accuracy	Precision	Recall	F1
Random		0.05	0.04	0.04	0.04
CNB	BoW	0.79	0.83	0.76	0.78
	TFIDF	0.78	0.82	0.74	0.76
SVM	BoW	0.80	0.82	0.80	0.81
	TFIDF	0.81	0.83	0.80	0.81
	fastText	0.71	0.70	0.67	0.68
LR	BoW	0.80	0.83	0.79	0.80
MH	BoW	0.80	0.82	0.80	0.81
	TFIDF	**0.82**	**0.84**	**0.81**	**0.82**
GB	fastText	0.78	0.80	0.75	0.77

The best model shown in bold

Notes: CNB complement Naive Bayes; *SVM* linear Support Vector Machine; *LR* logistic regression; *MH* modified Huber, quadratically smoothed SVM; *GB* Gradient Tree Boosting; *BoW* Bag of Words; *TFIDF* term frequency-inverse document frequency

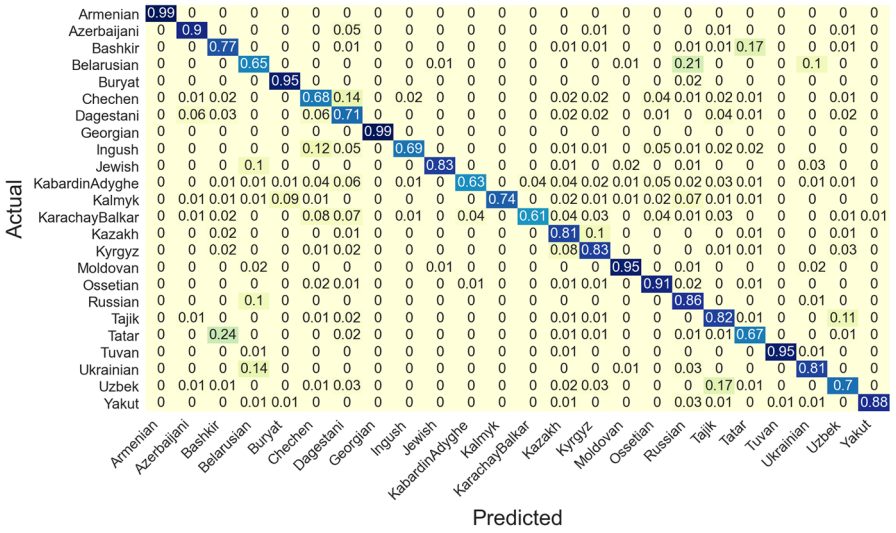

Fig. 1 Confusion matrix for 24 ethnic groups based on the MH model. Prediction accuracy is high (0.99) for groups with names that follow a simple pattern (Armenians and Georgians) and low for groups such as Karachay/Balkar (0.61) and Kabardin/Adyghe (0.63), often classified as other North Caucasian groups, and Belarusians (0.65), often classified as ethnic Russians

Asian groups with settled agricultural origins). The aggregated classification has 15 groups.

Table 4 shows prediction accuracy metrics for several ML models fitted to the aggregated data. The MH algorithm with the TFIDF vectorisation again provides the best model fit, with the overall accuracy increasing from 0.82 in the original classification with 24 ethnic groups to 0.92 in the aggregated classification with 15 groups.

Table 5 shows prediction accuracy for each ethnic group with the aggregated classification. Figure 3 shows the confusion matrix. Prediction accuracy for several ethnic groups has improved, in particular for eastern Slavic names (ethnic Russians, Belarusians, and Ukrainians, now predicted with the precision of 0.94), Bashkirs/Tatars (0.92), Kazakhs / Kyrgyz (0.91) and Uzbeks/Tajiks (0.91). Precision is the lowest for Chechens/Dagestani/Ingush (0.81), due to the confusion with similar names of neighbouring Caucasian ethnic groups (mainly Western North Caucasian and Azerbaijani).

Recall is the lowest for Kalmyks (0.74) where a fraction of names gets classified as either Slavic or Buryat. Kalmyks and Buryats are two Buddhist groups with common Mongolian origins, although populating the opposite ends of Russia. Recall is also lower for Jewish names (0.79) where many names get classified as Slavic. Note that many Ashkenazi Jews have surnames with German and Slavic origins.

To further validate the classifier, we apply it to two random samples of names collected on VK in Moscow and Kazan, two Russian cities with different ethnic structure of the populations (2000 names in each city), and compare the ethnic

Table 3 Prediction accuracy for ethnic groups

Ethnic group	Precision	Recall	F1
Armenian	0.99	0.99	0.99
Azerbaijani	0.87	0.90	0.89
Bashkir	0.76	0.77	0.76
Belarusian	0.74	0.65	0.69
Buryat	0.94	0.95	0.95
Chechen	0.74	0.68	0.71
Dagestani	0.69	0.71	0.70
Georgian	0.99	0.99	0.99
Ingush	0.86	0.69	0.76
Jewish	0.92	0.83	0.87
Kabardin / Adyghe	0.82	0.63	0.71
Kalmyk	0.94	0.74	0.83
Karachay / Balkar	0.85	0.61	0.71
Kazakh	0.82	0.81	0.82
Kyrgyz	0.81	0.83	0.82
Moldovan	0.94	0.95	0.94
Ossetian	0.86	0.91	0.88
Russian	0.71	0.86	0.78
Tajik	0.78	0.82	0.80
Tatar	0.68	0.67	0.68
Tuvan	0.97	0.95	0.96
Ukrainian	0.79	0.81	0.80
Uzbek	0.76	0.70	0.73
Yakut	0.97	0.88	0.93

Notes: Prediction accuracy estimated on the test set

distributions with the data from the 2010 Russian census. There are many limitations to this approach. The census and VK represent different populations (VK's being considerably younger). VK data were collected in 2021, and the census was conducted in 2010. The census likely under counts many immigrant groups, especially from Central Asia, as well as internal immigrants from the North Caucasus. However, while we should not expect the VK and census data to have the same ethnic distributions, we would still hope to see some consistency.

Table 6 presents the results of the comparison between the VK and census data, with the aggregated ethnic classification. The distributions are generally consistent both in Moscow and Kazan. In Moscow, 86% of the names in the VK sample get classified as ethnic Russian (or Belarusian and Ukrainian), compared to 93% in the census data. The proportions of non-Slavic ethnic groups are higher in the VK sample than in the census data. This does not necessarily represent a bias in the classifier and may reflect the undercounting of non-ethnically Russian groups in the census data for Moscow.

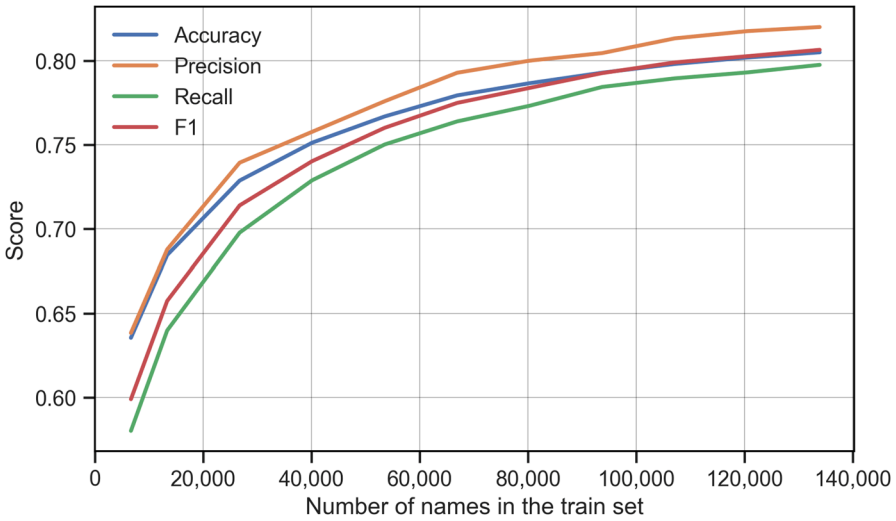

Fig. 2 Training set size and prediction accuracy. The steepest increase in accuracy occurs up to the point of approximately 50,000 to 60,000 names (i.e. about 2,500 names per group on average). Further increase in sample size leads to only marginal improvements

Table 4 Model performance with the aggregated classification

Algorithm	Vectorisation	Accuracy	Precision	Recall	F1
Random		0.11	0.07	0.07	0.07
CNB	BoW	0.89	0.92	0.85	0.88
SVM	TFIDF	0.91	0.92	0.90	0.91
LR	BoW	0.92	0.93	0.89	0.91
MH	TFIDF	**0.92**	**0.94**	**0.90**	**0.92**

The best model shown in bold

Notes: CNB complement Naive Bayes; *SVM* linear Support Vector Machine; *LR* logistic regression; *MH* modified Huber, quadratically smoothed SVM; *BoW* Bag of Words; *TFIDF* term frequency-inverse document frequency. Model accuracy evaluated on the test set

In Kazan, the VK classifier returns more ethnically Russian names compared to the census (60% compared to 49%) and fewer Tatar names (31% vs 48%). It may be the case that ethnic Russians are more likely to be VK users. It is also possible that some people who self-identify as Tatars may have ethnically Russian surnames, for example, as a result of ethnic intermarriage [36].

Validation with external historical data

So far, we have only used VK data for designing and validating the classifier. One may wonder how it performs with external data where the data generating process is

Table 5 Prediction accuracy for ethnic groups in the aggregate classification

Ethnic group	Precision	Recall	F1
Armenian	0.99	0.99	0.99
Azerbaijani	0.90	0.88	0.89
Bashkir/Tatar	0.92	0.94	0.93
Russian/Belarusian/Ukrainian	0.94	0.98	0.96
Buryat	0.96	0.94	0.95
Chechen/Dagestani/Ingush	0.81	0.83	0.82
Georgian	0.99	0.99	0.99
Jewish	0.94	0.79	0.86
Adyghe/Balkar/Kabardin/ Karachay/Ossetian	0.90	0.82	0.86
Kalmyk	0.95	0.74	0.84
Kazakh/Kyrgyz	0.91	0.91	0.91
Moldovan	0.96	0.93	0.95
Tajik/Uzbek	0.91	0.90	0.91
Tuvan	0.97	0.95	0.96
Yakut	0.98	0.87	0.92

Notes: Prediction accuracy was estimated with the test set

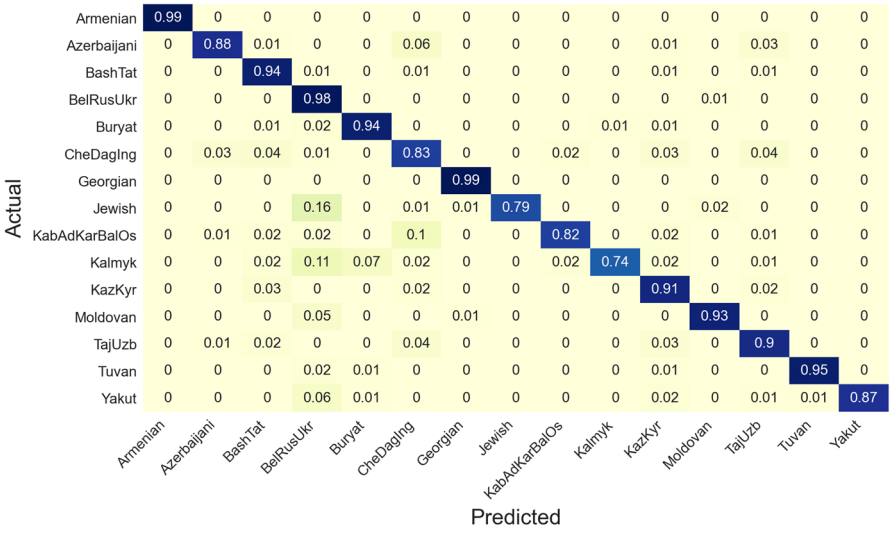

Fig. 3 Confusion matrix for 15 aggregated ethnic groups based on MH model Aggregation improves prediction accuracy for several ethnic groups, in particular for eastern Slavic names (0.94), Bashkirs/Tatars (0.92), Kazakhs/Kyrgyz (0.91) and Uzbeks/Tajiks (0.91). Precision is the lowest for Chechens/Dagestani/Ingush (0.81), due to the confusion with similar names of neighbouring Caucasian ethnic groups

different. It is, however, difficult to find a data set that has both personal names and recorded ethnicity for Russian ethnic groups. The only data source that we identified is of historical nature. These are data on the victims of political repression campaigns in the USSR in the 1920–30s collected by the Memorial society from various published source (see http://lists.memo.ru and https://github.com/nextgis/memorial_data). The data contain over 2.7 million individual records, often with assigned ethnicity. We remove records with missing ethnicity and from ethnic groups that are not part of our classification scheme (Poles, Germans, Latvians, etc.). The final analytic sample consists of 909,012 names with recorded ethnicity.

This data set is not ideal for our purposes. The data are about 100 years old; since then, the naming conventions for some ethnic groups (as well as the boundaries between groups) have evolved. Data on ethnicity were mostly recorded by the Soviet secret police, with many possible sources of bias, and were not necessarily based on self-identification of individuals. Soviet political terror affected some ethnic groups stronger than others. However, the results from the application of the classifier are still informative.

As expected, the classifier performs worse with an external data set, compared to VK data (see Table 7 and Fig 4). Both precision and recall are very high for ethnic Russians (combined with Ukrainians and Belarusians) who represent about 78% of the data set. For some ethnic groups (Azerbaijanis, Moldovans, Tajiks/Uzbeks, Yakuts), both precision and recall are low. Other groups (Armenians, Bashkirs/

Table 6 Validation of the classifier with the census data for Moscow and Kazan

City	Ethnic group	VK (%)	Census (2010, %)
Moscow			
	Russians/Belarusians/Ukrainians	86.0	93.4
	Jews	2.9	0.5
	Tatars/Bashkirs	2.6	1.4
	Chechens/Dagestanis/Ingushes	1.6	0.4
	Tajiks/Uzbeks	1.4	0.6
	Armenians	1.2	1.0
	Kazakhs/Kyrgyz	1.2	0.3
	Moldovans	1.1	0.2
	Other	2.0	2.2
Kazan			
	Russians/Belarusians/Ukrainians	60.0	49.2
	Tatars/Bashkirs	30.6	47.7
	Jews	2.0	0.2
	Chechens/Dagestanis/Ingushes	1.5	<0.5
	Kazakhs/Kyrgyz	1.4	<0.5
	Tajiks/Uzbeks	1.3	0.4
	Moldovans	1.0	<0.1
	Other	1.2	<1.5

Notes: VK data include samples of 2000 names in Moscow and Kazan each

Tatars, Georgians, Jews, Kalmyks) show high precision and recall even with historical and arguably not very reliable data.

Discussion

In this paper, we develop a classifier that predicts perceived ethnicity from data on personal names for major ethnic groups populating Russia. The multiclass classifier achieves the overall accuracy of 0.82 with 24 ethnic groups and 0.92 with 15 ethnic groups. It can be used in further studies of ethnic groups and relations in Russia, especially with VK and other social media data. We make the data and Python code for the classifier available in a Github repository at https://github.com/abessudnov/ruEthnicNamesPublic.

Ethnicity is a complex concept in the social sciences and it definitely cannot be reduced to patrilineal descent as reflected in surnames. We should be careful with defining the limits of what the classifier can and cannot do. We cannot and do not aim to predict ethnic self-identification. It is of course possible to have a surname originating in one ethnic group and identify with another. It is also possible to have mixed or double ethnic identities, and change and activate these depending on the context. Married women who take their husbands' surnames do not necessarily adopt their ethnic identity, in cases when it is different from their own. For some people, the sense of ethnic belonging may be more important than for others. The consensus in modern social science is that ethnicity ``is best thought of as an ongoing *process* of ethnic identification'' [37] rather than as a constant characteristic that is inherited across generations.

However, ethnicity as self-identity is different from *perceived* ethnicity that is an outcome of the process of ethnic categorisation. People use various signals to draw symbolic boundaries between themselves and members of other ethnic groups, such as language or accent, appearance, cultural norms, and habits [38, 39]. Surname can be one of such signals, especially in cases where other information is missing, as in job or housing applications. Perceived ethnicity can lead to differences in treatment by others, even in cases when it does not align with self-identity, and has real consequences in many areas of social life.

Therefore, we should emphasise again that in this paper, we predict perceived ethnicity (as reflected in surnames) that may or may not be the same as the sense of ethnic belonging. However, while in some cases, perceived ethnicity will be different from ethnic identity, often they will match. In the absence of better data, perceived ethnicity can be used as a rough proxy for ethnic identity. For some research questions, tracing ethnic origins may be more important than ethnic self-identification. For example, we may be interested in socio-economic outcomes of second generation immigrants (i.e., children of immigrants born in Russia), irrespective of whether and to what extent they identify with the ethnicity of their parents.

We should also acknowledge other limitations of the classifier.

First, while we include most major ethnic groups populating Russia, some groups are missing, as personal names of most members of these groups are

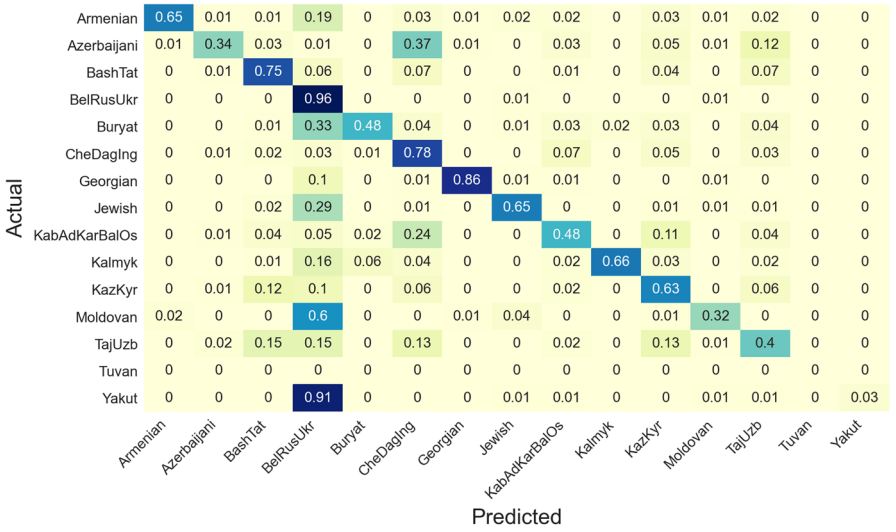

Fig. 4 Confusion matrix for 15 ethnic groups based on the MH model applied to the Memorial data. The classifier performs worse with an external data set, compared to VK data, with low precision for some ethnic groups (Azerbaijanis, Moldovans, Tajiks/Uzbeks, Yakuts). Precision is high for ethnic Russians combined with Ukrainians and Belarusians (0.97) and for some other groups (Armenians, Bashkirs/Tatars, Georgians, Jews, Kalmyks)

Table 7 Prediction accuracy with the Memorial data

Ethnic group	Precision	Recall	F1	*n*
Armenian	0.85	0.65	0.74	3,376
Azerbaijani	0.05	0.34	0.09	325
Bashkir/Tatar	0.78	0.75	0.76	41,282
Russian/Belarusian/Ukrainian	0.97	0.96	0.96	704,636
Buryat	0.38	0.48	0.42	5,690
Chechen/Dagestani/Ingush	0.06	0.78	0.10	1,986
Georgian	0.76	0.86	0.81	1,714
Jewish	0.74	0.65	0.69	43,883
Adyghe/Balkar/Kabardin/ Karachay/Ossetian	0.88	0.48	0.62	59,753
Kalmyk	0.81	0.66	0.73	4,406
Kazakh/Kyrgyz	0.11	0.63	0.19	37,182
Moldovan	0.09	0.32	0.14	1,495
Tajik/Uzbek	0.01	0.40	0.01	1,568
Yakut	0.17	0.03	0.05	1,716

indistinguishable from ethnic Russian. These groups are geographically concentrated in several Russian regions (Chuvashiya, Mordoviya, Udmurtiya, Mari El, and Komi) and the classifier is of limited used when applied to the data from these

regions. It is not possible to differentiate between ethnic Russians and members of indigenous ethnic groups in these regions on the basis of personal names only.

Second, to increase the reliability of the classifier, we create an aggregated ethnic classification scheme, combining some ethnic groups together, even when they are culturally and historically different. This applies, for example, to groups combining ethnic Russians, Ukrainians and Belarusians, or Chechens, Ingushes and Dagestanis (the latter is also an aggregation of several ethnic groups with their own languages). Many personal names in these groups are of common origin and more nuanced analysis would have lower reliability. Moreover, the idea of separating ethnic Russians and Ukrainians, or Tatars and Bashkirs, on the basis of personal names is of little conceptual validity. The boundaries between these groups have been fluid and changing over time, and personal names are not a strong marker of perceived ethnicity in these cases.

For most practical research purposes, the detailed classification would be of little interest and we advise using the aggregated classification. In some cases, even further aggregation would be possible, such as in the analysis of labour market discrimination where the main difference is between the groups of Eastern European and 'Southern' origin [40].

Third, the data cleaning procedure we use could introduce some bias in the data. Crowd workers on Yandex.Toloka could filter out names that do not look ``ethnic enough'', for example, by excluding names that look similar to ethnic Russian. This could affect ethnic groups with a large proportion of Russified personal names, such as Yakuts. Although the bias in crowdsourced data mining is usually recognized as an undesirable effect [41, 42], in our case, it can have a mixed effect on the reliability of the classifier. Keeping in the data the names of the ethnic Russian origin for non-ethnically Russian groups would result in a higher proportion of false negatives for ethnic Russian individuals. At the same time, excluding these names leads to more false negatives for the members of non-ethnically Russian ethnic groups with ethnic Russian names. While the name can be a strong marker of ethnicity, it cannot guarantee complete reliability.

Finally, we should emphasise the ethical aspect of this study. A tool that classifies ethnicity from personal names can be potentially misused by various actors ranging from state authorities to nationalist political movements [43]. The issue of ethnicity in Russia, very sensitive in the Soviet times, remains significant today in interpersonal relations, as well as in the labour market and housing. Its significance increased after the Russian invasion of Ukraine in 2022 (please note that this study was started and largely concluded before February 2022). It is important to recognize that our classifier cannot, and is not intended to identify ethnicity at the individual level. While this tool can produce reliable distributions for ethnicity for data sets with hundreds and thousands names, for each individual name, there remains a margin of error that does not let the classifier to be used for individual profiling.

A Hyperparameters' choice

In this section, we provide the hyperparameters search grid for algorithms and highlight the best parameters with **bold**. For fastText vectorisation, we used the default checkpoint from https://fasttext.cc/docs/en/crawl-vectors.html.

A.1 Complement Naive Bayes (CNB)

For CNB, we conduct search only through vectorisation methods parameters. N-grams ranges: {(1, 1), (1, 3), (1, 5), **(1, 7)**}, maximal vocabulary elements frequencies: {**5%**, 10%, 20%}, minimal vocabulary elements frequencies: {**1**, 10, 1%, 5%}, lowercasing: {True, **False**}.

A.2 Logistic regression

A.2.1 Vectorisation parameters

N-grams ranges: {(1, 3), **(1, 5)**, (1, 7)}, maximal vocabulary elements frequencies: {60%, **65%**, 70%, 75%, 80%}, minimal vocabulary elements frequencies: {**1**, 5, 1%, 10%}, lowercasing: {True, **False**}.

A.2.2 Model parameters

Regularization method: elasticnet, regularization term: {0.000005, 0.00001, **0.000025**, 0.00005, 0.0001}.

A.3 Support vector machine

A.3.1 Vectorisation parameters

N-grams ranges: {(1, 3), **(1, 5)**, (1, 7)}, maximal vocabulary elements frequencies: {35%, 45%, **55%**, 65%, 75%}, minimal vocabulary elements frequencies: {**1**, 5, 10, 50, 1%, 10%}, lowercasing: {True, **False**}.

A.3.2 Model parameters

Regularization method: elasticnet, regularization term: {**0.000005**, 0.00001, 0.000025, 0.00005, 0.0001}.

A.4 Quadratically smoothed support vector machine

A.4.1 Vectorisation parameters

N-grams ranges: {(1, 3), **(1, 5)**, (1, 7)}, maximal vocabulary elements frequencies: {35%, **45%**, 55%, 65%, 75%}, minimal vocabulary elements frequencies: {**1**, 5, 10, 50, 1%, 10%}, lowercasing: {True, **False**}.

A.4.2 Model parameters

Regularization method: elasticnet, regularization term: {0.000005, **0.00001**, 0.000025, 0.00005, 0.0001}.

A.5 Gradient tree boosting

A.5.1 Model parameters

Number of estimators: {100, **500**, 1000}.

A.6 Random Forest

A.6.1 Vectorisation parameters

N-grams ranges: {(1, 3), **(1, 5)**, (1, 7)}, maximal vocabulary elements frequencies: {1%, 5%, 10%, **20%**, 30%, 40%, 50%, 55%, 65%, 70%, 75%}, minimal vocabulary elements frequencies: {**1**, 5, 10, 10%}, lowercasing: {True, **False**}.

A.6.2 Model parameters

Number of estimators: {25, 50, 100, 200, **300**}, maximal depth: {50, **100**, 500, 1000, ∞}.

A.7 Multilayer perceptron

A.7.1 Model parameters

Hidden size: {100, **300**}, number of hidden layers: {1, **2**}, dropout probability: {0.1, **0.4**, 0.5}, activation function: {**ReLU**, ELU}, learning rate: {**0.0002**, 0.0005}, number of epoches: {**100**}.

A.8 Long short-term memory

A.8.1 Model parameters

Hidden size: {100, **150**, 200}, number of hidden layers: {2, **3**}, dropout probability: {0, 0.1, **0.5**}, bidirectional: {False, **True**} activation function: {**ReLU**}, learning rate: {**0.0005**, 0.0007}, number of epoches: {30, **60**}.

A.9 Convolutional neural network

A.9.1 Model parameters

Number of channels: {50, 100, 200}, number of hidden layers: {**3**}, dropout probability: {**0.5**}, activation function: {**ELU**}, learning rate: {**0.0001**}, number of epoches: {**20**}.

B Additional models' results

(See Table 8)

Table 8 Models performance on the test set

Algorithm	Vectorisation	Accuracy	Precision	Recall	F1
RF	BoW	0.73	0.78	0.67	0.69
MLP	fastText	0.71	–	–	–
LSTM	OHE	0.43	–	–	–
CNN	OHE	0.43	–	–	–

Notes: OHE: one-hot encoding, RF: random forest, MLP: multilayer perceptron, CNN: convolutional neural network, LSTM: Long Short-Term Memory

Acknowledgements We are grateful to Ivan Bibilov (European University at St Petersburg; Yandex) and Alexey Shpilman (HSE University) for their advice on developing this paper.

Data availability statement The research data supporting this publication and the Python code are openly available from Github at: https://github.com/abessudnov/ruEthnicNamesPublic.

Declarations

Conflict of interest The authors do not have competing financial or non-financial interests related to this work.

References

1. Lazer, D., & Radford, J. (2017). Data ex machina: introduction to big data. *Annual Review of Sociology, 43*, 19–39.
2. Buyalskaya, A., Gallo, M., & Camerer, C. F. (2021). The Golden Age of Social Science. *Proceedings of the National Academy of Sciences., 118*(5), e2002923118.
3. An J, Weber I (2016). # Greysanatomy vs.# Yankees: Demographics and Hashtag Use on Twitter. In: Proceedings of the tenth international AAAI conference on web and social media. vol. 10; . p. 523-6.
4. Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science., 350*(6264), 1073–6.
5. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences., 110*(15), 5802–5.
6. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., et al. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences, 114*(50), 13108–13.
7. Khunti, K., Routen, A., Banerjee, A., & Pareek, M. (2021). The need for improved collection and coding of ethnicity in health research. *Journal of Public Health, 43*(2), e270-2.
8. Flesken, A., & Hartl, J. (2020). Ethnicity, inequality, and perceived electoral fairness. *Social Science Research, 85*, 102363.
9. Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In A. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 1, pp. 309–93). Elsevier.
10. Imai, K., & Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis, 24*(2), 263–72.
11. Wood-Doughty Z, Andrews N, Marvin R, Dredze M (2018). Predicting Twitter User Demographics from Names Alone. In: Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media; . p. 105-11.
12. Clark, G. (2014). *The son also rises*. Princeton: Princeton University Press.
13. Mateos, P. (2014). Classifying ethnicity through people's names. *Names, ethnicity and populations* (pp. 117–144). Berlin: Springer.
14. Coldman, A. J., Braun, T., & Gallagher, R. P. (1988). The classification of ethnic status using name information. *Journal of Epidemiology & Community Health., 42*(4), 390–5.
15. Mateos, P. (2007). A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place., 13*(4), 243–63.
16. Hofstra, B., Corten, R., Van Tubergen, F., & Ellison, N. B. (2017). Sources of segregation in social Networks: a novel approach using facebook. *American Sociological Review., 82*(3), 625–56.
17. Hofstra, B., & de Schipper, N. C. (2018). Predicting ethnicity with first names in online social media networks. *Big Data & Society, 5*(1), 1–14.
18. Chang J, Rosenn I, Backstrom L, Marlow C (2010). ePluribus: Ethnicity on Social Networks. In: Proceedings of the international AAAI conference on web and social media; , vol.4, p.18-25.
19. Ambekar A, Ward C, Mohammed J, Male S, Skiena S (2009). Name-ethnicity classification from open sources. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining; . p. 49-58.
20. Lee J, Kim H, Ko M, Choi D, Choi J, Kang J (2017). Name nationality classification with recurrent neural networks. Proceedings of the twenty-sixth international joint conference on artificial intelligence. ;p. 2081-7.

21. Chaturvedi R, Chaturvedi S (2020). It's All in the Name: A Character Based Approach To Infer Religion. arXiv:2010.14479. . Available from: https://arxiv.org/abs/2010.14479.
22. Ye J, Han S, Hu Y, Coskun B, Liu M, Qin H, et al (2017). Nationality classification using name embeddings. In: Proceedings of the 2017 ACM on conference on information and knowledge management; . p. 1897-1906.
23. Cesare N, Grant C, Nguyen Q, Lee H, Nsoesie EO. How Well Can Machine Learning Predict Demographics of Social Media Users? arXiv:1702.01807v2. 2017. Available from: .
24. Unbegaun, B. O. (1972). *Russian surnames*. Oxford: Clarendon Press.
25. Karaulova, M., Gök, A., & Shapira, P. (2019). Identifying author heritage using surname data: an application for Russian surnames. *Journal of the Association for Information Science and Technology., 70*(5), 488–98.
26. Bessudnov A(2022). Ethnic and regional inequalities in the Russian military fatalities in the 2022 war in Ukraine SocArXiv. . Available from: https://osf.io/preprints/socarxiv/s43yf.
27. Sivak, E., & Smirnov, I. (2019). Parents mention sons more often than daughters on social media. *Proceedings of the National Academy of Sciences., 116*(6), 2039–41.
28. Smirnov, I. (2020). Estimating educational outcomes from students' short texts on social media. *EPJ Data Science., 9*(1), 27.
29. Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An introduction to information retrieval*. Cambridge: Cambridge University Press.
30. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient text classification. arXiv:1607.01759. 2016. Available from: https://arxiv.org/abs/1607.01759.
31. Zhang T (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the twenty-first international conference on machine learning. ICML . New York; 2004. p. 116.
32. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics., 29*(5), 1189–1232.
33. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research, 12*(85), 2825–30.
34. Chen T, Guestrin C (2016). XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16. New York; . p. 785-94.
35. Gorenburg, D. (1999). Identity change in Bashkortostan: Tatars into Bashkirs and back. *Ethnic and Racial Studies., 22*(3), 554–80.
36. Bessudnov, A., & Monden, C. (2021). Ethnic intermarriage in Russia: the tale of four cities. *Post-Soviet Affairs., 37*(4), 383–403.
37. Jenkins, R. (2008). *Rethinking Ethnicity* (2nd ed.). London: Sage.
38. Lamont, M., & Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology., 28*(1), 167–95.
39. Wimmer, A. (2013). *Ethnic boundary making: institutions, power, networks*. New York: Oxford University Press.
40. Bessudnov, A., & Shcherbak, A. (2020). Ethnic discrimination in multi-ethnic societies: evidence from Russia. *European Sociological Review., 36*(1), 104–20.
41. Ghai B, Liao QV, Zhang Y, Mueller K. Measuring social biases of crowd workers using counterfactual queries. arXiv:2004.02028. 2020. Available from:
42. La Barbera, D., Roitero, K., Demartini, G., Mizzaro, S., & Spina, D. (2020). Crowdsourcing truthfulness: the impact of judgment scale and assessor bias. *Advances in Information Retrieval., 12036*, 207–14.
43. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society., 3*(2), 1–21.