IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 24, 2021, accepted January 31, 2022, date of publication February 7, 2022, date of current version February 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149897

Misclassification Bias in Computational Social Science: A Simulation Approach for Assessing the Impact of Classification Errors on Social Indicators Research

SERGEY SMETANIN AND MIKHAIL KOMAROV, (Senior Member, IEEE)
Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 119049 Moscow, Russia
Corresponding author: Sergey Smetanin (sismetanin@gmail.com)

This research was supported in part through computational resources of HPC facilities at HSE University [59].

ABSTRACT A growing body of literature has examined the potential of machine learning algorithms in constructing social indicators based on the automatic classification of digital traces. However, as long as the classification algorithms' predictions are not completely error-free, the estimate of the relative occurrence of a particular class may be affected by misclassification bias, thereby affecting the value of the calculated social indicator. Although a significant amount of studies have investigated misclassification bias correction techniques, they commonly rely on a set of assumptions that are likely to be violated in practice, which calls into question the effectiveness of these methods. Thus, there is a knowledge gap with respect to the assessment of misclassification bias's impact on a specific social indicator formula without strict reference to the number of classes. Moreover, given the erroneous nature of automatic classification algorithms, the quality of a predicted indicator can be assessed not only using regression quality metrics, as was done in existing literature, but also using correlation metrics. In this paper, we propose a simulation approach for assessing the impact of misclassification bias on the calculated social indicators in terms of regression and correlation metrics. The proposed approach focuses on indicators calculated based on the distribution of classes and can process any number of classes. The proposed approach allows selecting the most appropriate classification model for a particular social indicator, and vice versa. Moreover, it allows for assessment of the optimistic level of correlation between the indicator calculated based on the results of the classification algorithm and the true underlying indicator.

INDEX TERMS Misclassification bias, social indicators, classification, supervised machine learning, computational social science, sentiment analysis, digital traces.

I. INTRODUCTION

Many studies in the social sciences are presently examining the potential of machine learning (ML) algorithms [1], forming computational social science—the academic sub-discipline concerned with computational approaches to the social sciences. Digital trace data are of special interest in the context of ML-based analysis, as the huge volume of data makes it a significant challenge to analyze it manually. According to Howison *et al.* [2], digital trace data are found (rather than produced for research), event-based (rather than summary data), and longitudinal (since events occur over a period of time) data that are both produced through and stored by an information system. These characteristics make digital traces an ideal source for building social indicators defined by Ferriss [3] as statistical time series "used to monitor the social system, helping to identify changes and to guide intervention to alter the course of social change." A typical example of social indicators constructed using ML analysis is the estimation of subjective well-being (SWB) based on user-generated content from social media, by employing an ML model trained to classify the sentiment of posts [4], [5]. From a practical point of view, a classification algorithm is commonly used to classify digital traces to the classes of interest; then, based on the distribution of these classes, an indicator is calculated for the entire population [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Claudia Raibulet.

However, as long as the ML algorithm's predictions are not completely error-free (we will hereinafter refer to this phenomenon as **misclassification bias**), the estimate of the relative occurrence of a particular class can be biased [7]–[9]. The key issue here is that optimal individual digital trace classification can lead to biased estimates of the digital trace class proportions and, subsequently, biased estimation of a social indicator. Generally accepted success criteria for classification, such as accuracy and F-measure on a test dataset [10], are appropriate for individual-level classification but can be seriously misleading when characterizing document populations or dynamic within populations [11]. For example, Zunic *et al.* [4] conducted a survey on sentiment analysis in health and well-being studies and found that the average classification accuracy for sentiment analysis was around 80%. It can be considered an acceptable classification performance for sentiment analysis, but suppose that all the misclassified objects were in a particular direction for one or more class. In that case, the statistical bias in using this method to estimate the aggregate quantities of interest could be as high as 20 percentage points. Moreover, if we take into account that, furthermore, the target social indicator is somehow calculated on the basis of the obtained proportions, then the deviation of the calculated indicator from the true indicator value may remain unchanged or change both up and down, adding another degree of uncertainty. As was highlighted in the measurement error studies [12] by the US Department of Education, all data collections errors, including misclassification errors, affect the final value of the calculated indicator depending on the specific formula of interest. Researchers have repeatedly reported cases where, because of errors in the data, including incorrect classification, the research results contained data that did not fully correspond to the real state of affairs. For example, Wolff *et al.* [13] examined data error in health, education, and income statistics used to construct the Human Development Index and found that up to 34% of countries were misclassified. By replicating prior studies, the authors showed that key estimated parameters varied by up to 100% due to data errors. Other papers also indicated errors in aggregate statistical data for suicide [14], disability [15], mortality [16], and life satisfaction [17]. Thus, despite the fact that many studies have examined methods of correcting classification bias [9], [11], we can conclude based on previously mentioned cases that it is essential to analyze this bias together with a mathematical formula[1] for calculating the indicator for social indicators research. Also, even though the influence of the classification bias on the classification results has been indicated in the literature, to the best of our knowledge, the impact of the misclassification bias on the calculated social indicators has not yet been assessed.

This paper provides a simulation approach for estimation of the impact of misclassification bias on the calculated social indicators. In this paper, we considered only indicators calculated based on the distribution of classes without any restrictions on the number of classes. The contributions of this study is five-fold.

- We propose a simulation approach for assessing the impact of misclassification bias on the calculated social indicators, which can be used for the following purposes:
  - Selecting the most appropriate classification model for a particular social indicator and vice versa.
  - Assessing the level of correlation between the indicator calculated based on the results of classification algorithm and true underlying indicator being inherited.
- Within the proposed simulation approach, we also define a formal model of online social data for social indicators research, which can be further used by academics.
- Within the proposed simulation approach, we also propose a method for approximation of predicted indicator based on the algorithm's confusion matrix and true indicator.
- Within the proposed simulation approach, we also propose a method for aggregation and interpretation of multiple correlation coefficients with p-values.
- We provide illustrative application examples of the proposed simulation approach and making conclusions based on the simulation outcomes.

Considering that the proposed simulation approach relies on a series of assumptions—as defined further (see assumptions 1, 2, and 3)—that can be violated in practice, the outcomes of the approach for real-life studies should be considered as naive and optimistic. However, we believe that this study contributes to the body of knowledge on computational social sciences and lays the foundation for future research on the impact of misclassification bias on calculated social indicators.

This article is organized as follows. In Section II, we provide a brief overview of simulation modeling. In Section III, we describe the literature analysis and indicate the knowledge gap. In Section IV, we propose a model for social indicators research based on digital traces. In Section V, we propose a simulation approach for assessing the impact of misclassification bias on social indicators research. In Section VI, we provide an illustrative example of applying the proposed approach to synthetic and real-life classification algorithms. Finally, in Section VII, we draw conclusions and suggest future research directions.

II. BACKGROUND

Simulation modeling is a special kind of mathematical modeling in which the system under study is replaced by a model describing the real system with sufficient accuracy, with

[1] As an example of a mathematical formula, we can consider the formula for educational attainment of 30–34 year olds described in the ETF Manual on the Use of Indicators [18]. Educational attainment refers to the highest educational level achieved by individuals expressed as a percentage of all persons in that age group. Mathematically, the formula is defined as follows.

$$\frac{\text{population 30–34 years old with tertiary education}}{\text{total population 30–34 years old}} * 100\%$$

.

which further experiments are conducted to obtain information about this system. In other words, a simulation model attempts to approximate a system's behavior and development over time by running a model [19]. Simulation models tend to be simplified abstractions of the system being modeled, the purpose of which is to capture a certain level of detail necessary to achieve the objectives of the study [20]. Simulation modeling is commonly used in such cases when a real system cannot be engaged, the analytical description cannot be formulated, or creating an analytical model is fundamentally impossible. In a broader sense, computer simulation attempts to approximate the behavior of a system and its development over time by implementing and running a computer simulation model. By changing the conditions and variables in the implemented simulation model, researchers can make predictions about the behavior of the simulated system without having to implement the entire system. Computer simulations are commonly used when performing system emulation is challenging or when it is necessary to emulate a system as part of more complex environment [20].

The literature has already described many examples of simulation models for systems of varying complexity [21]–[25], with which experiments are carried out to obtain information about these systems. In each study, the performance metrics of a simulation model were deeply related to the system being simulated and the goals of the simulation. For example, Gunal and Pidd [21] simulated the Accident and Emergency (A&E) Department at UK Hospitals and defined a performance measure as the percentage of patients who stayed in A&E more than 4 hours. Memon *et al.* [23] simulated blockchain systems and defined performance measures as the number of transactions per block, mining time of each block, system throughput, memorypool count, waiting time in memorypool, number of unconfirmed transactions in the whole system, total number of transactions, and number of generated blocks. Chan and Zhang [24] simulated a supply chain and defined a performance measure as the retailer's total cost. Thus, when creating a simulation model, it is also necessary to determine the key objectives to be achieved and the metrics to be obtained.

III. RELATED WORK

Correct classification of individuals, values, and attributes is an essential element of any study. Misclassification occurs when an individual, a value, or an attribute is assigned to a category other than that to which it should be assigned. This erroneous classification can lead to incorrect associations being observed between the assigned categories and the outcomes of interest [26], thereby biasing inferences drawn from the data collected [27], often substantially [28], or decreasing the power of the study [29]. As highlighted by Kloos *et al.* [9], misclassification bias occurs in a broad range of applications, including epidemiology [30], political science [31], and official statistics [32]. The objective of these applications is to shift focus from minimizing loss functions at the level of individual predictions to the level of aggregated predictions. In the context of ML, this objective is studied under quantification learning. Quantification learning aims to provide an aggregate estimation for unseen data by applying a model trained using a training dataset with a different data distribution [33]. However, there are certain drawbacks associated with the use of quantification learning in real-life studies. Firstly, the researchers note that since quantification learning is at an early stage of development, a more comprehensive theoretical analysis is required to better formulate both behavior of these algorithms and the learning objective in general [33], [34]. Secondly, although most of the efforts have focused on tackling binary quantification, quantification for more than two classes remains under-explored [33]. This is crucial for real-life applications: in a significant amount of cases, there are more than two classes of interest. Lastly, there is a lack of proper benchmark datasets for quantification [33]. Quantification studies require relatively large training and test datasets to train the model and obtain meaningful results and conclusions [9], [33]. In the studies mentioned above, data annotation tends to be expensive, and therefore there are little ready-made annotated data. Thus, much work remains to be done by the scientific community to freely apply quantitative learning to real-life research.

At the same time, a growing body of literature has investigated methods to reduce misclassification bias when aggregating categorical data from the level of individual predictions. For example, Hopkins *et al.* [11] explored new methods of automated content analysis designed to estimate the primary quantity of interest in many social science applications. As a part of the research, they also highlighted that misclassification bias may significantly affect the distribution of predicted classes. The authors proposed a classify-and-count method that gives approximately unbiased estimates of category proportions based on misclassification probabilities for each class—adjusting the distribution of predicted classes by confusion matrix normalized over rows. However, as shown later by Kloos *et al.* [9], the classify-and-count estimator is still (strongly) biased, so its application does not provide an unbiased estimate. In their paper, the authors also studied five existing estimation techniques to reduce the misclassification bias of binary classification algorithms. These methods for misclassification bias correction are commonly based on the assumption that misclassifications are independent across objects and that their probabilities are the same for each object. This assumption is often violated in practice, as misclassification in ML models is not random but tied to some separate groups of objects that are difficult for the model to separate from each other. This violation was partially confirmed in the study by Soroka *et al.* [35], which identified that different sentiment lexicons capture different underlying phenomena and highlighted "the importance of tailoring lexicons to domains to improve construct validity." As a consequence, the use of such methods can lead to the bias caused by misclassification being replaced by another bias caused by the application of the correction method. As Armstrong [36] mentioned, these methods can be complicated to use, however, and should be used cautiously because ''correction'' can magnify confounding if it is present. Moreover, although the majority of these methods focus on misclassification bias correction for two classes, real-life studies are commonly interested in more than two classes.

However, there has thus far been little discussion about the impact of misclassification bias on the specific social indicator formula rather than class proportions. We argue that these objectives should be treated separately, since not all classes may be taken into account in the target formula, and misclassifications between certain classes may compensate for each other to some extent. An absolutely accurate assessment of the quality of a social indicator calculated based on the results of automatic classification (hereinafter referred to as **predicted indicator**) is possible only if there is access to the true underlying value of the indicator (hereinafter referred to as **true indicator**), obtained using a completely correct classification approach. Manual data labeling is now considered the benchmark in ML, and in almost all cases, researchers try to train their models to classify data just as accurately. Thus, classes obtained using manual annotation with high-quality guidelines[2] and a high inter-rater agreement can be considered as practically the only[3] source of completely correct classification. However, since in the computational social sciences, the automatic analysis of a huge amount of data is of great interest, it is not only extremely difficult and time-consuming to annotate all the data manually in practice, but in the first place it is also extremely expensive. But even with enough annotated data, researchers may experience model overfitting when training a model. Overfitting is a fundamental challenge in supervised ML that prevents researchers from perfectly generalizing the models to sufficiently fit observed data on training data and unseen data on the testing set [38]. Due to the presence of overfitting, the model tends to work perfectly on the training set but does not fit well on the test set. Core to model training are appropriate ways for data splitting or resampling: the final training parameters we must be chosen only after evaluating a number tuning parameters via data splitting or resampling [39]. In particular, cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters, thereby preventing model overfitting [40]. The cross-validation procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, this procedure is often referred to as k-fold cross-validation. Various other strategies for addressing overfitting can be found in the recent survey papers [38], [41], [42] on that topic. It should also be noted that existing misclassification bias correction methods use primarily regression quality metrics to assess their performance, such as mean absolute error (MAE) or mean squared error (MSE). However, given the erroneous nature of the existing algorithms for automatic classification, interest for research may not be so much the absolute correspondence of the index calculated based on the predictions to the true underlying one, but rather their correlation. In this case, based on the results of the study, it will not be entirely correct to draw conclusions about the absolute values of the indicator, but it will be possible to analyze its changes over time. Thus, there is a knowledge gap regarding the assessment of the impact of misclassification bias on a specific social indicator formula [4] without strict reference to the number of classes. Moreover, given the erroneous nature of automatic classification algorithms, the quality of a predicted indicator can be assessed not only using regression quality metrics, but also with correlation metrics. Based on these findings, we propose a simulation approach for assessing the impact of misclassification error on a particular social indicators formula, given the algorithm classification performance and the information about data available for analysis.

IV. MODEL

In this Section, we propose a model for social indicators research based on digital traces. We applied classical set theory to develop our model, as recent literature [43], [44] articulated a series of its advantages in the case of computational social sciences.

A. CONCEPTUAL MODEL

The **Online Social Data Model for Social Indicators Research** consists of three elements: **Digital Traces**, **Classification**, and **Indicators**. The **Digital Traces** represent the source data found for the analysis, which are event-based and longitudinal, thus suitable for construction social indicators. The **Classification** represent the automated approaches for digital trace object classification based on ML methods. The **Indicators** represent a methodology for calculating social indicators based on classification results and estimating its quality. When constructing the model, we assumed that the source digital traces for the analysis are representative of the general population, so no additional sampling methods should be applied. As a consequence, all information about individuals can be omitted.

B. FORMAL MODEL

The **Online Social Data Model for Social Indicators Research** is defined as a tuple $OSDM_{SIR} = (DT, C, I)$ where

- DT is the **Digital Traces** representing the source digital traces for the analysis,

[2] By high-quality annotation guidelines, we mean, at a high level, guidelines that allow performing an annotation in such a way that the resulting annotation completely matches the true underlying quantitative or quantitative parameter being annotated. The exact definition of high-quality annotation guidelines and criteria for assessing the quality of guidelines lies outside of the scope of this paper.

[3] In theory, a ML model with 100% accuracy can also serve as a source of completely correct classes, provided that it was trained on a representative, high-quality training dataset. However, in practice, it is not only extremely difficult to achieve 100% accuracy of the model but there is evidence [37] that predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy.

[4] Hereinafter, we mean some specific mathematical formula for calculating the social indicator.

- C is the **Classification** representing ML components, allowing mapping of digital traces to corresponding classes of scientific interest, and
- I is the **Indicators** representing social indicators of interest that should be computed within a particular social indicators research.

The **Digital Traces** of the Online Social Data Model for Social Indicators Research is defined as a tuple $DT = (TI, X, \rightarrow_{interval})$, where

- $TI = \{ti_0, ti_1, \ldots, ti_K\}$ is an ordered set of $K \in \mathbb{N}$ non-overlapping time intervals such as $ti_i < ti_{i+1}$,
- $X = \{x_1, x_2, \ldots, x_N\}$ is a finite set of $N \in \mathbb{N}$ digital trace objects, and
- $\rightarrow_{interval}: TI \rightarrow P_{disj}(X)$ is a partial function mapping time intervals to mutually disjoint non-empty subsets of digital traces created in that time interval.

A subset of digital trace objects created in the time interval ti_i will be hereinafter referred to as X_{ti_i}, i.e. $\rightarrow_{interval}(ti_i) = X_{ti_i}$. The number of items in a subset X_{ti_i} will be hereinafter referred to as $N_{ti_i} \in \mathbb{N}$.

The **Classification** of the Online Social Data Model for Social Indicators Research is defined as a tuple $C = (Y, f_T, f_P, CM)$, where

- $Y = \{y_1, y_2, \ldots, y_M\}$ is a finite set of $M \in \mathbb{N}$ classes,
- $f_T : X \rightarrow Y$ is a true mapping function,
- $f_P : X \rightarrow Y$ is an algorithm approximating the mapping function f (i.e., classification model), and
- $CM \in \mathbb{N}_{\nvdash}{}^{M \times M}$ is a confusion matrix for the algorithm f_P.

The **Indicators** of the Online Social Data Model for Social Indicators Research is defined as a tuple $I = (TSI, QM, AQM)$, where

- $TSI = \{I_{ti_1}, I_{ti_2}, \ldots, I_{ti_K}\} \in \mathbb{R}^K$ is a vector representing time series indicator, where $I_{ti_i} : Y^{N_{ti_i}} \rightarrow \mathbb{R}$ is an indicator function mapping a set of $N_{ti_i} \in \mathbb{N}_{\nvdash}$ classified digital traces created in time interval ti_i to an indicator value,
- $QM = \{qm_1, qm_2, \ldots, qm_U\}$ is a set of U target quality measures where each item represents a function $qm_i : (\mathbb{R}^K \times \mathbb{R}^K) \rightarrow \mathbb{R}^l$ returning a vector of $K \in \mathbb{N}_{\nvdash}$ real numbers, and
- $AQM = \{aqm_1, aqm_2, \ldots, aqm_U\}$ is a set of $U \in \mathbb{N}$ aggregated target quality measures where each i-th item represents an aggregation function suitable for qm_i and defined as $aqm_i : (\mathbb{R}^L)^V \rightarrow \mathbb{R}^H$, where $L \in \mathbb{N}$ is the number calculated target quality measures to be aggregated and $H \in \mathbb{N}$ is the size of the vector representing the aggregated target quality measure.

A particular target quality measure function qm_i can be represented in a variety of ways depending on the needs of the particular social indicators research—for example, MAE for identifying deviation of the approximated indicator from the true indicator, or the Pearson correlation coefficient (Pearson's r) for identifying correlation between those indicators. As a consequence, it returns a vector of real numbers since difference quality measure may return different number of values (e.g., it may contain one value in the case of MSE and two values representing the confidence interval in the case of Pearson's r). The aggregated target quality measure aqm_i returns a vector of H real numbers because the aggregation approach may vary depending on the target quality measure qm_i and specifics of the research (e.g., macro-averaging can be applied for MAE, resulting in a one-dimensional vector, and the confidence interval can be calculated based on Fisher z-transformation for Pearson's r, resulting in a two-dimensional vector).

TSI calculated based on mapping function f_T will be further referred to as TSI_T. TSI calculated based on the algorithm f_P will be referred to as TSI_P.

C. PROBLEM STATEMENT

In terms of defined notations, the problem statement for the estimation of the impact of misclassification bias on the calculated social indicators can be defined in the following way. Given a trained classification model f_P and its error matrix on a test dataset CM, data for analysis X, an indicator calculation formula I, and formulas for the target quality metric qm_i and aggregated target quality metric aqm_i, it is necessary to estimate the classification bias AQ_m.

V. SIMULATION APPROACH FOR ASSESSING THE IMPACT OF MISCLASSIFICATION BIAS ON SOCIAL INDICATORS RESEARCH

As mentioned earlier, the assessment of the impact of misclassification bias on calculated social indicators is possible only if there is access to the true value of the indicator, obtained using a completely correct classification approach. In our approach, we propose to simulate the true indicator, then, on its basis, approximate the results of the classification algorithm, and then calculate the quality metrics. Formally, the proposed approach consists of three steps.

1) Simulate the true indicator TSI_T by simulating true mapping function f_T.
2) Approximate the predicted indicator TSI_P by approximating an algorithm f_p based on the true mapping function f_T.
3) Calculate the quality qm_i of the predicted indicator TSI_P for multiple simulations, and then calculate the aggregated quality score aqm_i.

The proposed approach is based on the following assumptions.

Assumption 1: The training data for the classification model was labeled manually using high-quality guidelines, and the annotators demonstrated a high inter-rater agreement score.

Consequently, we can consider that all digital traces in the training dataset were assigned with class labels that completely match the true underlying parameter being annotated.

Assumption 2: The classification model was trained on the training data representative of the digital traces available for analysis.

This assumption in combination with assumption 1 allows us to consider that class distribution in the digital traces available for analysis is equal to class distribution in the training dataset.

Assumption 3: (Mis)classifications are independent across objects, and the (mis)classification probabilities are the same for each object, conditional on their true class label.
Consequently, we could use a confusion matrix for approximating the predicted index based on true index using the inverse classify-and-count approach [11] for misclassification bias correction.

Taking into account the simulation and approximation nature of the proposed approach, as well a set of applied assumptions that can be violated in practice, the outcomes of the approach for real-life studies should be considered as naive and optimistic. In other words, we recommend considering the outcomes as an optimistic assessment representing the best case in real-life studies, provided it is not possible to prove the fulfillment of all assumptions.

A. TRUE INDICATOR SIMULATION

Since both true and predicted indicators are calculated based on the number of objects mapped to specific classes for each time interval, the simulated data for each time interval are a vector with a dimension equal to the number of classes, and it is defined as follows:

$$SCD_{ti_i} = (scd_{ti_i,y_1}, scd_{ti_i,y_2}, \dots, scd_{ti_i,y_M}) \in \mathbb{N}\nvdash^{M},$$
$$\sum_{j=1}^{M} scd_{ti_i,j} = N_{ti_i}. \tag{1}$$

Also, the simulated data can be presented as a time series

$$STS_{y_i,TI} = (scd_{ti_1,y_i}, scd_{ti_2,y_i}, \dots, scd_{ti_K,y_i}) \in \mathbb{N}\nvdash^{K}, \tag{2}$$

where each element scd_{ti_i,y_j} represents the number of digital traces contained in time interval ti_i and labeled as a class y_j. Since the true indicator is unknown, we propose to synthetically generate the number of objects of each class for each analyzed time interval and calculate the true indicator TSI_T based on the generated data. Considering that the distribution in the digital traces available for analysis is equal to class distribution in the training dataset (see assumption 2), we can expect the simulated data to satisfy the following condition:

$$\frac{\sum_{j=1}^{K} scd_{ti_j,y_i}}{\sum_{o=1}^{M}\sum_{j=1}^{K} scd_{ti_j,y_o}} = \frac{\sum_{j=1}^{M} cm_{y_i,y_j}}{\sum_{o=1}^{M}\sum_{j=1}^{M} cm_{y_o,y_j}}, \tag{3}$$

where cm_{y_i,y_j} is the number of objects with true class y_i classified as y_j, as further defined in Eq. (4). At the same time, we do not expect class distribution for a specified time interval to be equal to the class distribution in the training dataset, since according to assumption 2, the training dataset is representative of the whole set of data available for the analysis but not necessarily of a particular slice of these data.

In essence, this means that we simulate behavior of the mapping function f_T. For each time interval, the total number of generated objects should not exceed the number of objects contained in the data for analysis during the same time interval. SCD_{ti_i} generated for the calculation of true indicator will be hereinafter referred to as SCD_{T,ti_i}.

B. PREDICTED INDICATOR APPROXIMATION

Once the true mapping function is defined and the true indicator is calculated, we must define an algorithm approximating true mapping function (i.e., classification model) f_P. In other words, we need to correct misclassification bias. To begin with, we need estimates of the algorithm's (mis)classification probabilities. Following [45], we assume that misclassifications are independent across objects and that the (mis)classification probabilities are the same for each object, conditional on their true class label. The (mis)classification probabilities for each class are estimated via confusion matrix normalized over true classes, which is calculated based on a confusion matrix CM. After that, we must adjust the true classes distribution SCD_{T,ti_i} by (mis)classification probabilities to acquire the approximate predicted classes distribution. A similar approach has been widely used in the literature [9], [11] but only for the inverse problem—to correct the classification bias from the already predicted classes.

The confusion matrix can be presented as

$$CM = \begin{pmatrix} cm_{y_1,y_1} & cm_{y_1,y_2} & \cdots & cm_{y_1,y_M} \\ cm_{y_2,y_1} & cm_{y_2,y_2} & \cdots & cm_{y_2,y_M} \\ \vdots & \vdots & \ddots & \vdots \\ cm_{y_M,y_1} & cm_{y_M,y_2} & \cdots & cm_{y_M,y_M} \end{pmatrix} \in \mathbb{N}\nvdash^{M\times M}, \tag{4}$$

where each row of the matrix represents the instances in an actual class, and each column represents the instances in a predicted class. An asterisk refers to whole rows or columns in a matrix. For example, $cm_{i,*}$ refers to the i-th row of CM, and $cm_{*,j}$ refers to the j-th column of CM.

$$cm_{y_i,y_*} = \begin{pmatrix} cm_{y_i,y_1} & cm_{y_i,y_2} & \cdots & cm_{y_i,y_M} \end{pmatrix}. \tag{5}$$
$$cm_{y_*,y_j} = \begin{pmatrix} cm_{y_1,y_j} & cm_{y_2,y_j} & \cdots & cm_{y_M,y_j} \end{pmatrix}^{T}. \tag{6}$$

A confusion matrix normalized over true classes can be further calculated as follows:

$$CM^{ntc} = \begin{pmatrix} \frac{1}{\sum cm_{y_1,y_*}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sum cm_{y_2,y_*}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sum cm_{y_M,y_*}} \end{pmatrix} \times CM$$
$$= \begin{pmatrix} cm^{ntc}_{y_1,y_1} & cm^{ntc}_{y_1,y_2} & \cdots & cm^{ntc}_{y_1,y_M} \\ cm^{ntc}_{y_2,y_1} & cm^{ntc}_{y_2,y_2} & \cdots & cm^{ntc}_{y_2,y_M} \\ \vdots & \vdots & \ddots & \vdots \\ cm^{ntc}_{y_M,y_1} & cm^{ntc}_{y_M,y_2} & \cdots & cm^{ntc}_{y_M,y_M} \end{pmatrix} \in \mathbb{R}^{N\times N} \tag{7}$$

Assuming that our model is unbiased toward a specific type of errors (i.e., the probability of a model to make a error is distributed randomly) and always follows a given confusion matrix CM, we can approximate the non-normalized confusion matrix of our model for a simulated data as follows:

$$
\begin{aligned}
&CM' \\
&= \begin{pmatrix} scd_{T,ti_i,y_1} & 0 & \cdots & 0 \\ 0 & scd_{T,ti_i,y_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & scd_{T,ti_i,y_M} \end{pmatrix} \times CM^{ntc} \\
&= \begin{pmatrix} cm'_{1,1} & cm'_{1,2} & \cdots & cm^{ntc}_{1,M} \\ cm'_{2,1} & cm'_{2,2} & \cdots & cm'_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ cm'_{M,1} & cm'_{M,2} & \cdots & cm'_{M,M} \end{pmatrix} \in {\mathbb{N}_{\nvdash}}^{N \times N}, \\
&\sum_{o=1}^{M} cm'_{y_j,y_o} = scd_{T,ti_i,y_j} \qquad (8)
\end{aligned}
$$

Note that the normalized confusion matrix CM^{nte} operates with $\mathbb{R}$, whereas the non-normalized confusion matrix operates with $\mathbb{N}_{\nvdash}$, so it is necessary to round the results of matrix multiplication and randomly adjust them to meet the target class distribution if necessary.

The simulated distribution of predicted classes based on simulated true classes distribution SCD_{T,ti_i} for a given time interval ti_i is as follows:

$$
\begin{aligned}
SCD_{P,ti_i} &= (\sum_{j=1}^{M} cm'_{y_j,y_1}, \sum_{j=1}^{M} cm'_{y_j,y_2}, \ldots, \sum_{j=1}^{M} cm'_{y_j,y_M}) \\
&\in {\mathbb{N}_{\nvdash}}^{M} \\
\sum_{j=1}^{M} & scd_{P,ti_i,y_j} = N_{ti_i} \qquad (9)
\end{aligned}
$$

C. QUALITY ASSESSMENT

Finally, we can calculate $Y^{N_{ti_i}}$ (i.e., a set of N_{ti_i} classified objects created in time interval ti_i to an indicator value) based on obtained class distributions SCD_{T,ti_i} and SCD_{P,ti_i} for further calculation of the true indicator and predicted indicator, respectively. Since the order of the items in $Y^{N_{ti_i}}$ is not important, we can define the order of items in any way following our class distributions. After that, we can calculate TSI_T, TSI_P, and qm_i. By repeating the entire procedure multiple times,[5] we can obtain multiple qm_i and calculate aqm_i.

However, if for such metrics as MAE and MSE the aggregation methods are well defined (for example, it can be a simple average value), then the correlation aggregation tends to be a more challenging task to accomplish. Note that in the case of correlation analysis of time series, it is important to check that these time series are stationary, and if they are not, then apply some technique to make them stationary (e.g., differencing). Moreover, in this section we assumed that the analyzed time series are stationary.

Furthermore, the method of aggregating Pearson's or Spearman's correlation coefficients is presented. This method consists of two parts: aggregation of correlation coefficients and aggregation of p-values.

1) AGGREGATION OF CORRELATION COEFFICIENTS

For combining Pearson or Spearman correlations, we propose to use the aggregation method based on Fisher z-transformation described in [47]. Let p_i denote an estimate of the Pearson or Spearman correlation in study i and n_i denote the number of observations in study i. An estimate of the average study population correlation is defined as follows.

$$p = m^{-1} \sum_{i=1}^{n} p_i \qquad (10)$$

An estimate of the variance is defined as follows.

$$var(p) = m^{-2} \sum_{i=1}^{n} var(p_i), \qquad (11)$$

where $var(p_i)$ is the variance for a particular correlation. Variance for Pearson correlation is defined as follows.

$$var(p_i) = \frac{(1 - p_i^2)^2}{n_i - 3}. \qquad (12)$$

Variance for Spearman correlation is defined as follows.

$$var(p_i) = \frac{(1 - p_i^2)^2 (1 + \frac{p_i}{2})}{n_i - 3}. \qquad (13)$$

An approximate two-sided $(1 - \alpha)\%$ confidence interval for the average study population correlation is

$$CI_{corr} = \tanh(\arctan(p) \pm z \sqrt{\frac{var(p)}{(1 - p^2)^2}}), \qquad (14)$$

where z is the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution (i.e., the probit) corresponding to the target error rate α. This method should use all Spearman correlations or all Pearson correlations because these correlations are not comparable [47].

2) AGGREGATION OF P-VALUES

According to Heard and Rubin-Delanchy [48], there are six most fundamental or commonly used statistics for combining p-values: Fisher's method ($S_F = \sum_{i=1}^{n} \log p_i$) [49], Pearson's method ($SP = -\sum_{i=1}^{n} \log(1 - p_i)$) [50], George's method ($S_G = S_F + S_P = \sum_{i=1}^{n} \log(p_i/(1 - p_i))$) [51], Edgington's method ($S_E = \sum_{i=1}^{n} p_i$) [52], Stouffer's method ($S_S = \sum_{i=1}^{n} \Phi^{-1}(p_i)$) [53], where Φ is the standard normal cumulative distribution function, and Tippett's method ($ST = min(p_1, p_2, \ldots, p_n)$) [54]. Although each method is optimal in some setting, all of them can be considered as strict methods tending to reject a hypothesis if even a very small part of the tests did not show the specified level of statistical

[5] Determining the number of required simulation runs lies outside the scope of this paper. For more information on this topic, please refer to [46].

significance. Given that the approximation algorithm is not error-free and simulations can be repeated an infinite number of times, it is extremely likely that in certain cases the confidence interval for a particular simulation iteration may not be statistically significant. Thus, to interpret the results of this work, it is necessary to formulate a softer approach to the aggregation of n-values, which will take into account the not error-free nature of the classification algorithm. As a softer method of aggregation, we propose to use a probabilistic approach and focus not on the absolute values of the calculated p-values but on whether they satisfy a predefined level of significance—for example, less than 0.05. In this case, the list of absolute p-values can be converted into a list of 0 (not satisfying) and 1 (satisfying) and then perceived as a binomial distribution.

$$\hat{P} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n\} = \{p_1 < \alpha, p_2 < \alpha, \ldots, p_n < \alpha\}. \quad (15)$$

For the obtained binomial distribution $\hat{P}$, we can calculate a binomial proportion confidence interval and consider it as our soft approximation. We can further interpret it as a confidence interval of obtaining statistically significant results with a predefined confidence level at least as extreme as the observed results of a statistical hypothesis test. The resulting formula is defined as follows:

$$CI_p = \frac{n_S}{n} \pm \frac{z}{n\sqrt{n}}\sqrt{n_S n_F}, \quad (16)$$

where $n = n(\hat{P})$ is the total number of experiments, $n_S = n(\{\hat{p}_i = 1 | \forall \hat{p}_i \in \hat{P}\})$ is the number of successes, $n_F = n(\{\hat{p}_i = 0 | \forall \hat{p}_i \in \hat{P}\}) = n - n_S$ is the number of failures, and z is the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution (i.e., the probit) corresponding to the target error rate α. Since $\hat{p}_i = 0$ is considered as a failure and $\hat{p}_i = 1$ is considered as a success, the bound of the confidence interval of aggregated p-values CI_p will be 1 in case $p_i < \alpha$ and 0 in case $p_i \geq \alpha$. For example, if all bound of CI_p are greater than 0.95 (considering $\alpha = 0.05$), then the aggregated correlation CI_{corr} is statistically significant.

3) INTERPRETATION OF RESULTS

If upper and lower bounds of aggregated p-values CI_p are high (generally more than 0.95), then the correlation CI_{corr} is statistically significant, so we can use the calculated Pearson's or Spearman's coefficient. Several authors have offered guidelines for the interpretation of a correlation coefficient, so we could use the most appropriate for a particular study—for example, guidelines by Zou *et al.* [55]. Strength of the correlation should be defined by the lower bound for positive correlation and by the upper bound for negative correlation. Depending on the strength of the correlation, we could make the following conclusions.

- If CI_{corr} is perfect, then we can confirm that there is **no impact** of the misclassification bias on the calculation of the indicator, allowing us to achieve the **perfect level** of correlation between predicted and true indicators.
- If CI_{corr} is strong, then we can confirm that there is a **weak impact** of the misclassification bias on the calculation of the indicator, allowing us to achieve a **strong level of correlation** between predicted and true indicators.
- If CI_{corr} is moderate, then we can confirm that there is a **moderate impact** of the misclassification bias on the calculation of the indicator, allowing us to achieve a **moderate level of correlation** between predicted and true indicators.
- If CI_{corr} is weak, then we can confirm that there is a **strong impact** of the misclassification bias on the calculation of the indicator, allowing us to achieve the **weak level of correlation** between predicted and true indicators.
- If CI_{corr} is absent, then we can confirm that there is a **perfect impact** of the misclassification bias on the calculation of the indicator, allowing us to achieve **no correlation** between predicted and true indicators.

However, considering that assumption 3 is commonly not satisfied in practice (or it is extremely difficult to prove that it is satisfied for a certain case), in real-life studies it would be more correct to use the obtained conclusion as the best case—that is, the lower estimation of the potential impact of the classification bias on the social indicators research.

If the upper or the lower bound of aggregated p-values is not high (generally less than 0.95), then the correlation is not statistically significant (it might have happened just by chance) and we should not rely upon the Pearson's or Spearman's coefficient. In other words, we cannot confirm that there is a correlation between predicted and true indicators, and, consequently, we cannot recommend this algorithm for calculating the indicator based on available data.

VI. ILLUSTRATIVE EXAMPLE: SUBJECTIVE WELL-BEING

Let us say that the research domain is SWB and the research aim is to construct a SWB indicator based on sentiment analysis of posts from social networks. Then, the proposed simulation approach can be employed for the assessment of the impact of misclassification bias on social indicators calculations. Let us also assume that we are interested in three SWB indicators, calculated based on sentiment classification of 1,000,000 posts distributed between 36 time intervals into three classes: *negative*, *neutral*, and *positive*. The class distribution is drawn from a real-life dataset of social media posts, RuSentiment [56]: 11.65% negative posts, 64.13% neutral posts, and 24.22% positive posts. The indicators of interest are defined as follows.

1) SWB_{P2E} represents the share of expressed positive emotions relative to all emotions and is defined as

$$SWB_{P2E} = \frac{POS}{POS + NEG + NEU}, \quad (17)$$

where POS, NEG, and NEU are numbers of posts classified as *positive*, *negative*, and *neutral*, respectively. We assume this SWB indicator as an approximation of

TABLE 1. Results of simulation runs.

Algorithm	Indicator	Aggregated Quality Measures			
		Pearson's Correlation		macro-avg MSE	macro-avg MAE
		$\mathbf{CI_{corr}}$	$\mathbf{CI_p}$		
Random	SWB_{P2E}	-	-	0.0125	0.0972
	SWB_{P2PN}	-	-	0.0365	0.1715
	SWB_{P-N2E}	-	-	0.0224	0.1286
Poor	SWB_{P2E}	(0.4298, 0.4323)	(0.7521, 0.7596)	0.0097	0.0784
	SWB_{P2PN}	(-0.3982, -0.3957)	(0.6700, 0.6782)	0.1799	0.4124
	SWB_{P-N2E}	(-0.0717, -0.0687)	(0.0933, 0.0985)	0.2128	0.4538
Basic	SWB_{P2E}	(0.9978, 0.9978)	(1.0, 1.0)	0.0124	0.1013
	SWB_{P2PN}	(0.6868, 0.6885)	(0.9970, 0.9979)	0.2536	0.4984
	SWB_{P-N2E}	(0.9408, 0.9412)	(1.0, 1.0)	0.4778	0.6903
Advanced	SWB_{P2E}	(0.9988, 0.9988)	(1.0, 1.0)	0.0007	0.0236
	SWB_{P2PN}	(0.8396, 0.8406)	(1.0, 1.0)	0.0205	0.1351
	SWB_{P-N2E}	(0.9790, 0.9791)	(1.0, 1.0)	0.0080	0.0881
Perfect	SWB_{P2E}	(1.0, 1.0)	(1.0, 1.0)	0.0	0.0
	SWB_{P2PN}	(1.0, 1.0)	(1.0, 1.0)	0.0	0.0
	SWB_{P-N2E}	(1.0, 1.0)	(1.0, 1.0)	0.0	0.0

positive affect, which is defined by "the extent to which a person feels enthusiastic, active, and alert" [57].

2) SWB_{P2PN} represents the share of expressed positive emotions relative to the sum of positive and negative emotions and is defined as

$$SWB_{P2PN} = \frac{POS}{POS + NEG}. \quad (18)$$

We also assume this SWB indicator as one of the possible approximations of positive affect, where the influence of neutral sentiment is not taken into account.

3) SWB_{P-N2E} represents the difference between positive and negative emotions divided by the number of all emotions.

$$SWB_{P-N2E} = \frac{POS - NEG}{POS + NEG + NEU}. \quad (19)$$

We assume this SWB indicator as an approximation of life satisfaction, which takes into account the difference between positive and negative emotions in relation to the total number of emotions.

For the generation of synthetic time series, we applied the nonlinear autoregressive moving average model from the TimeSynth [58] library with random hyperparameters for each simulation run. We selected five different approaches with different classification algorithms. For each algorithm, we provided confusion matrices for classes $Y = \{negative, neutral, positive\}$ as required by the proposed simulation approach. We chose Pearson's correlation coefficient as the main metric and MAE and MSE as secondary metrics. Since the generated synthetic time series are not stationary, we differentiated the series before calculating the correlation coefficient. For each calculated indicator, we ran 50,000[6] simulation iterations. We performed all the calculations on the supercomputer facilities at HSE University. The university HPC cluster occupies seventh place in rating the most powerful computers of the CIS TOP50 and helps to solve ML problems, population genomics, hydrodynamics, atomistic and continuous modeling in physics, generative probabilistic models, financial row forecasting algorithms, and other actual problems [59]. However, having access to a supercomputer is not a prerequisite for applying this simulation approach. Calculations can be carried out both on a personal computer or in cloud services—for example, Google Colab. We used CPU nodes of the supercomputer for faster development iterations.

A. RANDOM CLASSIFICATION ALGORITHM

The random algorithm randomly assigns sentiment classes to posts, thereby representing the worst classification quality. The normalized confusion matrix for this algorithm is defined as follows.

$$CM^{ntc} = \begin{pmatrix} 0.3(3) & 0.3(3) & 0.3(3) \\ 0.3(3) & 0.3(3) & 0.3(3) \\ 0.3(3) & 0.3(3) & 0.3(3) \end{pmatrix}. \quad (20)$$

According to the results of the simulation, the predicted index is a constant (see Fig. 1), so the Pearson's correlation coefficient (see Table 1) is undefined for all indicators because it has variance equal to zero. Thus, because the correlation is undefined, we cannot confirm that there is a correlation between predicted and true indicators. Consequently, we cannot recommend using this algorithm for calculating the indicator based on available data.

B. POOR CLASSIFICATION ALGORITHM

The poor algorithm classifies all objects with a high level of errors. The normalized confusion matrix for this algorithm is defined as follows.

$$CM^{ntc} = \begin{pmatrix} 0.5 & 0.0 & 0.5 \\ 0.5 & 0.4 & 0.1 \\ 0.5 & 0.3 & 0.2 \end{pmatrix}. \quad (21)$$

[6] This value was obtained empirically and should not serve as a recommendation for further research. An analytical calculation of the exact number of required iterations was not carried out, since the main goal of this section is to show an example of the application of the simulation approach, as well as ways of interpreting the results.

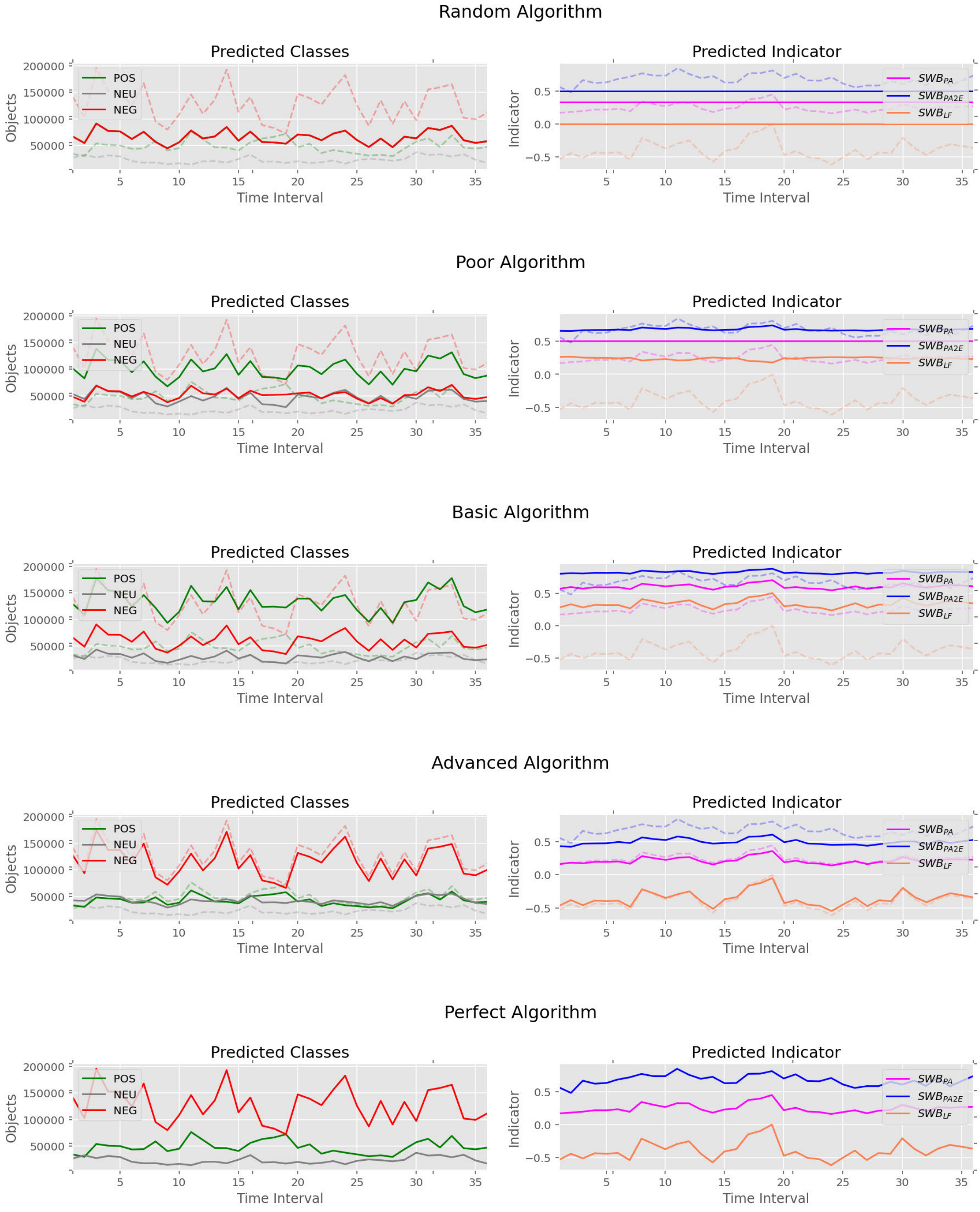

FIGURE 1. Example of a simulation run. Regular lines represent predicted classes and predicted indicators. Dotted lines represent true classes distribution and true indicators.

According to simulation results, the aggregated p-values are lower than 0.95 (see Table 1), so the correlation is not statistically significant and we should not rely upon the correlation coefficient. In other words, we cannot confirm that there is a correlation between predicted and true indicators. Consequently, we cannot recommend using this algorithm for calculating the indicator based on available data (see Fig. 1).

C. BASIC CLASSIFICATION ALGORITHM

The basic classification algorithm is a multinomial logistic regression (MLR) method, a common baseline approach for sentiment analysis task. As a real-life example, we selected an MLR model presented by Ismail *et al.* [60]. According to their paper, the normalized confusion matrix for their classification model is defined as follows.

$$CM^{ntc} = \begin{pmatrix} 0.98 & 0 & 0.02 \\ 0.77 & 0.16 & 0.06 \\ 0.53 & 0.2 & 0.45 \end{pmatrix}. \tag{22}$$

According to the simulation results, the aggregated p-values are higher than 0.95 (see Table 1), so we can make the following conclusions from the given assumptions, data, algorithm, and conditions under consideration.

- We can confirm that there is a negligible impact of the misclassification bias on the calculation of SWB_{P2E}, allowing us to achieve an almost perfect level of correlation between the predicted and true indicators.
- We can confirm that there is a moderate impact of the misclassification bias on the calculation of SWB_{P2PN}, allowing us to achieve a moderate level of correlation between the predicted and true indicators.
- We can confirm that there is a weak impact of the misclassification bias on the calculation of SWB_{P-N2E}, allowing us to achieve a strong level of correlation between the predicted and true indicators.

However, considering the particulars of assumption 3 mentioned above, in real-life studies it would be more correct to use obtained conclusions as estimations for the best cases.

D. ADVANCED CLASSIFICATION ALGORITHM

The third classification algorithm is based on one of the most recent advances in natural language processing, a pre-trained language model. As a real-life example, we selected one of our previously developed models, ruBert-FiT-RuReviews [61], which achieved state-of-the-art results on the RuReviews dataset [62]. According to the paper, the normalized confusion matrix for this classification model is defined as follows.

$$CM^{ntc} = \begin{pmatrix} 0.74 & 0.24 & 0.01 \\ 0.22 & 0.70 & 0.08 \\ 0.01 & 0.11 & 0.88 \end{pmatrix}. \tag{23}$$

According to the simulation results, the aggregated p-values are higher than 0.95 (see Table 1), so we can make the following conclusions for the given assumptions, data, algorithm, and conditions under consideration.

- We can confirm that the there is a negligible impact of the misclassification bias on the calculation of SWB_{P2E} and SWB_{P-N2E}, allowing us to achieve an almost perfect level of correlation between the predicted and true indicators.
- We can confirm that there is a weak impact of the misclassification bias on the calculation of SWB_{P2PN}, allowing us to achieve a strong level of correlation between the predicted and true indicators.

However, considering the particulars of assumption 3 mentioned above, in real-life studies it would be more correct to use obtained conclusions as estimations for the best cases.

E. PERFECT CLASSIFICATION ALGORITHM

The fourth classification algorithm is the perfect algorithm, which correctly classifies all objects. The normalized confusion matrix for this algorithm is defined as follows.

$$CM^{ntc} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{24}$$

According to the simulation results, the aggregated p-values are higher than 0.95 (see Table 1), so we can make the following conclusions for the given assumptions, data, algorithm, and conditions under consideration. We can confirm that there is no impact of the misclassification bias on the calculation of the SWB_{P2E}, SWB_{P2PN} and SWB_{P-N2E}, allowing us to achieve the perfect level of correlation between predicted and true indicators. However, considering the particulars of the assumption 3 mentioned above, in real-life studies it would be more correct to use obtained conclusions as estimations for the best cases.

VII. CONCLUSION

In this paper we propose a simulation approach for assessing the impact of misclassification bias on the calculated social indicators. We considered only (1) indicators calculated based on the distribution of classes and (2) the case of multiclass classification. As mentioned in earlier, the contributions of this study are five-fold. Firstly, we proposed a simulation approach for assessing the impact of classification bias on the calculated social indicators. This approach can be used for selecting the most appropriate classification model for a particular social indicator and vise versa, as well as assessing the level of correlation between the true and predicted indicators. Secondly, we defined a formal model of online social data for social indicators research, which can be used further by academics. Thirdly, we proposed a method for approximation of predicted indicator based on the algorithm's confusion matrix and true indicator. Fourthly, we proposed a method for aggregation and interpretation of multiple correlation coefficients with p-values. Lastly, we provided illustrative examples of applying the proposed simulation approach and making conclusions based on the simulation outcomes. Considering that the assumptions used in our model can be violated in practice, the outcomes of the approach for real-life studies should be considered as naive and optimistic. However, we believe that this study contributes to the body of knowledge of computational social sciences and lays the foundation for future research on the impact of misclassification bias on calculated social indicators.

Future research directions on the current topic are therefore recommended.

- Based on the assumption 2, we expected in Eq. (3) that class distribution in a representative training dataset is equal to the distribution in the data for analysis. For a more accurate assessment, it is possible to consider the class distribution in the training dataset not as a fixed ratios but as an interval of possible values. For example, we can consider calculating binomial proportion confidence interval for each class distribution and simulate true index based on these intervals.
- Depending on the particular formula of a social indicator, different types of errors may affect the resulting value of an indicator in a different way. For example, mutually correcting errors (i.e., misclassification errors that correct each other during the indicator calculation) can negatively affect individual-level classification quality and at the same time have no impact on the calculated indicator. A more detailed study of different types of errors and their influence on the calculated indicator may allow researchers to develop a more comprehensive strategy for training classification models.
- Based on assumption 1, we expect that all objects in the training dataset were assigned with class labels that completely match the true underlying parameter being annotated. However, given that quite often annotators may disagree with each other, especially when working with subjective concepts, it is logical to suppose that for a model to have the same markup quality as a human, it is not necessary to have 100% accuracy on the test subset. Thus, further research can focus on what quality of classification on a given dataset can be considered equivalent to that of a human, given the metrics of inter-rater agreement on a given dataset.

REFERENCES

[1] J. Grimmer, M. E. Roberts, and B. M. Stewart, "Machine learning for social science: An agnostic approach," *Annu. Rev. Political Sci.*, vol. 24, no. 1, pp. 395–419, May 2021.

[2] J. Howison, A. Wiggins, and K. Crowston, "Validity issues in the use of social network analysis with digital trace data," *J. Assoc. Inf. Syst.*, vol. 12, no. 12, pp. 767–797, Dec. 2011.

[3] A. L. Ferriss, "The uses of social indicators," *Social Forces*, vol. 66, no. 3, pp. 601–617, 1988.

[4] A. Zunic, P. Corcoran, and I. Spasic, "Sentiment analysis in health and well-being: Systematic review," *JMIR Med. Informat.*, vol. 8, no. 1, Jan. 2020, Art. no. e16023.

[5] S. Smetanin, "The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives," *IEEE Access*, vol. 8, pp. 110693–110719, 2020.

[6] V. Voukelatou, L. Gabrielli, I. Miliou, S. Cresci, R. Sharma, M. Tesconi, and L. Pappalardo, "Measuring objective and subjective well-being: Dimensions and data sources," *Int. J. Data Sci. Anal.*, vol. 11, no. 4, pp. 1–31, 2020.

[7] J. E. Schwartz, "The neglected problem of measurement error in categorical data," *Sociol. Methods Res.*, vol. 13, no. 4, pp. 435–466, May 1985.

[8] S. Scholtus and A. van Delden, *On Accuracy Estimators Based a Binary Classifier*. The Hague, The Netherlands: Statistics Netherlands, 2020.

[9] K. Kloos, Q. Meertens, S. Scholtus, and J. Karch, "Comparing correction methods to reduce misclassification bias," in *Proc. Benelux Conf. Artif. Intell.* Leiden, The Netherlands: Springer, 2020, pp. 64–90.

[10] J. D. Novakovic, A. Veljovic, S. S. Ilic, Z. Papic, and T. Milica, "Evaluation of classification models in machine learning," *Theory Appl. Math. Comput. Sci.*, vol. 7, no. 1, pp. 39–46, Apr. 2017.

[11] D. J. Hopkins and G. King, "A method of automated nonparametric content analysis for social science," *Amer. J. Political Sci.*, vol. 54, no. 1, pp. 229–247, Jan. 2010.

[12] S. Salvucci, E. Walter, V. Conley, S. Fink, and M. Saba, "Measurement error studies at the national center for education statistics," U.S. Dept. Educ. Office Educ. Res. Improvement, Washington, DC, USA, Tech. Rep. NCES-97-464, 1997.

[13] H. Wolff, H. Chong, and M. Auffhammer, "Classification, detection and consequences of data error: Evidence from the human development index," *Econ. J.*, vol. 121, no. 553, pp. 843–870, Jun. 2011.

[14] I. R. H. Rockett, J. B. Samora, and J. H. Coben, "The black–white suicide paradox: Possible effects of misclassification," *Social Sci. Med.*, vol. 63, no. 8, pp. 2165–2175, Oct. 2006.

[15] A. Gosling and E.-C. Saloniki, "Correction of misclassification error in disability rates," *Health Econ.*, vol. 23, no. 9, pp. 1084–1097, Sep. 2014.

[16] A. Helkin, S. V. Jain, A. Gruessner, M. Fleming, L. Kohman, M. Costanza, and R. N. Cooney, "Impact of ASA score misclassification on NSQIP predicted mortality: A retrospective analysis," *Perioperative Med.*, vol. 6, no. 1, pp. 1–6, Dec. 2017.

[17] E. Oparina and S. Srisuma, "Analyzing subjective well-being data with misclassification," *J. Bus. Econ. Statist.*, pp. 1–14, Feb. 2021.

[18] *ETF Manual Use Indicators*, ETF, Turin Process, 2013.

[19] J. Banks, *Discrete Event System Simulation*. London, U.K.: Pearson, 2005.

[20] S. Smetanin, A. Ometov, M. Komarov, P. Masek, and Y. Koucheryavy, "Blockchain evaluation approaches: State-of-the-Art and future perspective," *Sensors*, vol. 20, no. 12, p. 3358, Jun. 2020.

[21] M. Gunal and M. Pidd, "Understanding accident and emergency department performance using simulation," in *Proc. Winter Simul. Conf.*, Dec. 2006, pp. 446–452.

[22] H. S. Na and A. Banerjee, "Agent-based discrete-event simulation model for no-notice natural disaster evacuation planning," *Comput. Ind. Eng.*, vol. 129, pp. 44–55, Mar. 2019.

[23] R. A. Memon, J. P. Li, and J. Ahmed, "Simulation model for blockchain systems using queuing theory," *Electronics*, vol. 8, no. 2, p. 234, Feb. 2019.

[24] F. T. S. Chan and T. Zhang, "The impact of collaborative transportation management on supply chain performance: A simulation approach," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2319–2329, Mar. 2011.

[25] M. Saidani, H. Kim, and J. Kim, "Designing optimal COVID-19 testing stations locally: A discrete event simulation model applied on a university campus," *PLoS ONE*, vol. 16, no. 6, Jun. 2021, Art. no. e0253869.

[26] E. A. Spencer, K. R. Mahtani, J. Brassey, and C. Heneghan, "Misclassification bias," *CatalogueOf Bias*, 2018. [Online]. Available: http://www.catalogueofbiases.org/biases/misclassificationbias

[27] D. Haine, I. Dohoo, and S. Dufour, "Selection and misclassification biases in longitudinal studies," *Frontiers Veterinary Sci.*, vol. 5, p. 99, May 2018.

[28] D. Quade, P. A. Lachenbruch, F. S. Whaley, D. K. McCLISH, and R. W. Haley, "Effects of misclassifications on statistical inferences in epidemiology," *Amer. J. Epidemiol.*, vol. 111, no. 5, pp. 503–515, May 1980.

[29] E. White, "The effect of misclassification of disease status in follow-up studies: Implications for selecting disease classification criteria," *Amer. J. Epidemiol.*, vol. 124, no. 5, pp. 816–825, Nov. 1986.

[30] T. L. Lash, M. P. Fox, R. F. MacLehose, G. Maldonado, L. C. McCandless, and S. Greenland, "Good practices for quantitative bias analysis," *Int. J. Epidemiol.*, vol. 43, no. 6, pp. 1969–1985, Dec. 2014.

[31] G. Wiedemann, "Proportional classification revisited: Automatic content analysis of political manifestos using active learning," *Social Sci. Comput. Rev.*, vol. 37, no. 2, pp. 135–159, Apr. 2019.

[32] Q. A. Meertens, C. G. H. Diks, H. J. van den Herik, and F. W. Takes, "A data driven supply-side approach for estimating cross-border internet purchases within the European union," *J. Roy. Stat. Soc., A, Statist. Soc.*, vol. 183, no. 1, pp. 61–90, Jan. 2020.

[33] P. González, A. Castaño, N. V. Chawla, and J. J. D. Coz, "A review on quantification learning," *ACM Comput. Surv.*, vol. 50, no. 5, pp. 1–40, 2017.

[34] P. González, J. Díez, N. Chawla, and J. J. del Coz, "Why is quantification an interesting learning problem?" *Prog. Artif. Intell.*, vol. 6, no. 1, pp. 53–58, Mar. 2017.

[35] S. Soroka, L. Young, and M. Balmas, "Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content," *Ann. Amer. Acad. Political Social Sci.*, vol. 659, no. 1, pp. 108–121, May 2015.

[36] B. G. Armstrong, "Effect of measurement error on epidemiological studies of environmental and occupational exposures," *Occupational Environ. Med.*, vol. 55, no. 10, pp. 651–656, Oct. 1998.

[37] F. J. Valverde-Albacete and C. Peláez-Moreno, "100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e84217.

[38] X. Ying, "An overview of overfitting and its solutions," *J. Phys., Conf.*, vol. 1168, Feb. 2019, Art. no. 022022.

[39] M. Kuhn and K. Johnson, "Over-fitting and model tuning," in *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013, pp. 61–92.

[40] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford, U.K.: Academic, 2019, pp. 542–545.

[41] C. Gambella, B. Ghaddar, and J. Naoum-Sawaya, "Optimization problems for machine learning: A survey," *Eur. J. Oper. Res.*, vol. 290, no. 3, pp. 807–828, May 2021.

[42] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Trans. Softw. Eng.*, vol. 48, no. 1, pp. 1–36, Jan. 2022.

[43] R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Towards a set theoretical approach to big data analytics," in *Proc. IEEE Int. Congr. Big Data*, Jun. 2014, pp. 629–636.

[44] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: A set theoretical approach to big data analytics," *IEEE Access*, vol. 4, pp. 2542–2571, 2016.

[45] A. van Delden, S. Scholtus, and J. Burger, "Accuracy of mixed-source statistics as affected by classification errors," *J. Off. Statist.*, vol. 32, no. 3, p. 619, 2016.

[46] F. E. Ritter, M. J. Schoelles, K. S. Quigley, and L. C. Klein, "Determining the number of simulation runs: Treating simulations as theories by not sampling their behavior," in *Human-in-the-Loop Simulations*. London, U.K.: Springer, 2011, pp. 97–116.

[47] D. G. Bonett, *An Introduction to Meta-Analysis*. Santa Cruz, CA, USA: Univ. of California, Santa Cruz, 2017.

[48] N. A. Heard and P. Rubin-Delanchy, "Choosing between methods of combining-values," *Biometrika*, vol. 105, no. 1, pp. 239–246, 2018.

[49] R. A. Fisher, "Statistical methods for research workers," *Statistical Methods for Research Workers*, 5th ed. Edinburgh, U.K.: Oliver & Boyd, 1934.

[50] K. Pearson, "On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random," *Biometrika*, vol. 25, nos. 3–4, pp. 379–410, Dec. 1933.

[51] G. S. Mudholkar and E. O. George, "The logit method for combining probabilities," in *Symposium on Optimizing Methods in Statistics*. New York, NY, USA: Academic, 1979, pp. 345–366.

[52] E. S. Edgington, "An additive method for combining probability values from independent experiments," *J. Psychol.*, vol. 80, no. 2, pp. 351–363, Mar. 1972.

[53] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams, *The American Soldier: Adjustment During Army Life*, vol. 1. Princeton, NJ, USA: Princeton Univ. Press, 1949.

[54] L. H. C. Tippett, *The Methods of Statistics*. London, U.K.: Williams & Nargate, 1934.

[55] K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and simple linear regression," *Radiology*, vol. 227, no. 3, pp. 617–628, 2003.

[56] A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, and A. Gribov, "RuSentiment: An enriched sentiment analysis dataset for social media in Russian," in *Proc. 27th Int. Conf. Comput. Linguistics*, Aug. 2018, pp. 755–763.

[57] M. Brondino, D. Raccanello, R. Burro, and M. Pasini, "Positive affect over time and emotion regulation strategies: Exploring trajectories with latent growth mixture model analysis," *Frontiers Psychol.*, vol. 11, p. 1575, Jul. 2020.

[58] J. R. Maat, A. Malali, and P. Protopapas. (2017). *TimeSynth: A Multipurpose Library for Synthetic Time Series in Python*. [Online]. Available: http://github.com/TimeSynth/TimeSynth

[59] P. S. Kostenetskiy, R. A. Chulkevich, and V. I. Kozyrev, "HPC resources of the higher school of economics," *J. Phys., Conf.*, vol. 1740, no. 1, Jan. 2021, Art. no. 012050.

[60] R. Ismail, M. Omer, M. Tabir, N. Mahadi, and I. Amin, "Sentiment analysis for Arabic dialect using supervised learning," in *Proc. Int. Conf. Comput., Control, Electr., Electron. Eng. (ICCCEEE)*, Aug. 2018, pp. 1–6.

[61] S. Smetanin and M. Komarov, "Deep transfer learning baselines for sentiment analysis in Russian," *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102484.

[62] S. Smetanin and M. Komarov, "Sentiment analysis of product reviews in Russian using convolutional neural networks," in *Proc. IEEE 21st Conf. Bus. Informat. (CBI)*, Jul. 2019, pp. 482–486.

SERGEY SMETANIN received the B.S. degree in software engineering and the M.S. degree in business informatics from the National Research University Higher School of Economics, Moscow, Russia, in 2016 and 2018, respectively. His research interests include computational social science, computational linguistics, sentiment analysis, mobile applications development, and engineering management.

MIKHAIL KOMAROV (Senior Member, IEEE) received the bachelor's degree in IT, in 2008, the bachelor's degree in management, in 2009, the degree in IT, in 2010, the Ph.D. degree in Russia, in 2012, and the Ph.D. degree in Finland, in 2016. Since 2012, he has been an Invited Expert and a Speaker with the UN Internet Governance Forum. He is currently a Full Professor with the Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics. His research interests include e-business, mobile commerce, distributed ledger technology, and new business models. He has a membership of the Steering Committee of the IEEE Conference on Business Informatics, Academy of Management, Association of Information Systems, and Technical Committee on Business Informatics and Systems IEEE. He is also a Founding Member of the Special Interest Group on Big Data Applications at the Association of Information Systems.