Method for analysis and formation of representative text datasets*

Olena Sobko[1,*,†], Olexander Mazurets[1,†], Maryna Molchanova[1,†], Iurii Krak[2,3,†] and Olexander Barmak[1,†]

[1] Khmelnytskyi National University, Khmelnytskyi, 11, Institutes str., 29016, Ukraine
[2] Taras Shevchenko National University of Kyiv, Kyiv, 64/13, Volodymyrska str., 01601, Ukraine
[3] Glushkov Cybernetics Institute, Kyiv, 40, Glushkov ave., 03187, Ukraine

Abstract

The paper is devoted to the creation and approbation of method for analysis and formation of representative text datasets according to FATE fairness principle for subject areas. The method performs an analysis of dataset representativeness according to ethical aspects, as result of which a representative adjustment of the dataset according to ethical aspects is performed. When adjusting the dataset, optimization problem is solved both for the selection of redundant elements for removal, and for the formation of requirements for ethical aspects of belonging to each element for data augmentation. To investigate the effectiveness of the method, software was created that uses machine learning models to classify texts according to various ethical aspects – age, gender, religion, ethnicity, etc. The obtained deviations of the sample distributions by ethical aspects classes of dataset, transformed according to the created method, compared to the ideal representative distribution were: minimum 0.00%, maximum 0.04%, average 0.02%. The obtained results contribute to improvement of representativeness of text datasets and fair and unbiased representation of demographic groups in them, which increases trust in decisions made by artificial intelligence.

Keywords

NLP, data ethical correctness, ethical principles, non-discrimination, text datasets representative

1. Introduction

In today's world, numerous solutions using artificial intelligence are being actively developed to solve various tasks that people face every day. Accordingly, the results generated by artificial intelligence depend on the training datasets on which they were trained, in other words, the content of these datasets directly affects the final result. Lack of transparency regarding the sources and characteristics of the data used to train AI algorithms reduces confidence in the results obtained. In this case, users are often unable to appreciate the potential biases or discriminatory elements built into these algorithms. Insufficient awareness of the content of educational datasets increases the risk of spreading unfair or inaccurate decisions, which can have serious consequences for individuals and society as a whole [1].

Means for evaluating the representativeness of a textual data set in accordance with the principles of ethical non-discrimination are currently lacking. This is especially relevant for

AdvAIT-2024: 1st International Workshop on Advanced Applied Information Technologies, December 5, 2024, Khmelnytskyi, Ukraine - Zilina, Slovakia

[*] Corresponding author.

[†] These authors contributed equally.

olenasobko.ua@gmail.com(O. Sobko); exe.chong@gmail.com (O. Mazurets); m.o.molchanova@gmail.com (M. Molchanova); yuri.krak@gmail.com(I. Krak); barmak@gmail.com (O. Barmak)

0000-0001-5371-5788 (O. Sobko); 0000-0002-8900-0650 (O. Mazurets); 0000-0001-9810-936X (M. Molchanova); 0000-0002-8043-0785 (I. Krak); 0000-0003-0739-9678 (O. Barmak)

socially important and sensitive tasks according to SDG3 (good health and well-being), SDG4 (quality education) and SDG16 (peace, justice, and strong institutions), for example, detecting cyberbullying, determining the emotional state of people based on text messages, etc. The lack of attention to ethical components when creating and using datasets leads to bias in algorithms, which negatively affects the fairness and reliability of the decisions made [2].

Well-known datasets for training neural networks, for example [3, 4], are actively used by researchers, because they have a large amount of data, but they were not validated by the authors regarding representativeness according to the principle of fairness, and therefore, the use of such datasets for training artificial intelligence algorithms may potentially violate ethical principles and, hence, have a low reliability of the decisions made.

The representativeness of the data in the datasets not only affects the accuracy of the results and models, but is also closely related to the principles of FATE (Fairness, Accountability, Transparency, Ethics) in the use of data and development of artificial intelligence technologies. If dataset does not include adequate representation of all social, demographic, or cultural groups, it can lead to discriminatory patterns that prioritize one group over another, so are not fair. The representativeness of datasets according to ethical principle of FATE is achieved by correct balancing according to various ethical aspects: gender, religious, age, etc. [5].

The main contribution of the paper is the development and validation of an approach to the analysis and formation of representative text samples of data according to the principle of fairness of FATE for subject areas.

Further, in chapter 2, a review of works related to the topic of the study, namely the formation of representative text samples and the issue of impartial representation of demographic groups according to the principle of justice, is carried out. Chapter 3 offers a description of the method of analysis and formation of representative samples of text data, the datasets used for further experimental studies of the effectiveness of the given method are given and described. Chapter 4 contains an experimental study. Section 5 presents the results and discussion. Chapter 6 concludes the work.

2. Related work

Many works have been devoted to the study of the representativeness of text samples and the fair and unbiased representation of demographic groups in them, since the concepts of representativeness, fairness and impartiality are important in the creation of ethical and fair machine learning models [6]. Natural language processing tools are widely used for this purpose [7]. Recently, authors have increasingly paid attention to the issue of representativeness of data in samples, but the current state suggests that data sets have gaps in the representation of gender and race, and the complex nature of demographic variables makes classification difficult and inconsistent. Thus, the question of representativeness of data in sets that include people with disabilities and the elderly is considered. The authors recommend increasing representativeness by adding samples for underrepresented groups, including by collecting additional data or using synthetic data methods to improve representation of minorities and people with disabilities.

In the article [8], the authors raise the important problem of sample representativeness in the context of machine learning and artificial intelligence, emphasizing the need for accurate representation of population data. The main strategy that the authors propose to achieve high quality models is the use of stratified samples, which allow to reduce the variability between subgroups and accurately reflect the proportions between different categories in the population.

The authors of the study [9] consider biases arising both from class imbalances in the data and from sensitive (protected) characteristics such as race or gender. The approach increases model accuracy by balancing classes and reduces dependence on sensitive features, which improves group fairness.

IBM researchers have developed an open-source AI Fairness 360 toolkit for evaluating and reducing discrimination in machine learning models [10]. The main purpose of the toolkit is to detect bias based on attributes such as race, gender or age, and to provide methods for representation of all given social groups at different stages of model development.

The article [11] highlights the problem of intersectional biases in natural language processing (NLP) models, namely the unrepresentative and biased representation of different groups of people in textual datasets. The results showed that although existing debiasing methods (for example, for BERT or RoBERTa) preserve the predictive accuracy of the models well, their ability to reduce intersectional biases is limited.

The authors of [12] propose a specialized model of machine learning to detect and minimize bias in textual data, in particular, in news articles. The authors claim that their approach is effective because of deep models and transformative architectures that are able to detect and correct biases at different stages of machine learning.

The article [13] presents the problem of gender bias in natural language processing models, solving it using two main approaches: statistical and causal fairness. Researchers use techniques such as counterfactual data augmentation for causal debiasing, as well as resampling and revaging methods for statistical debiasing. The results showed that the combination of these techniques allows for significant bias in the models by both statistical and causal metrics.

Article [14] is devoted to solving the problem of intersectional bias in the predictions of machine learning models, in particular deep neural networks. Researchers propose a new method based on the Apriori algorithm for automatically detecting biased subgroups in data. It allows efficient generation of frequent subgroups and calculation of fairness metrics for them.

In [15], the authors identify and classify bias in natural language processing using transformer models such as BERT. The authors explore different ways to identify bias, including identifying social characteristics such as gender, race, religion, and sexual orientation.

The study [16] examines the problem of cyberbullying, which is a threat to people based on different characteristics, such as religion, age, ethnicity, and gender. The data set used by the authors has been modified with ethical considerations in mind, which ensures responsible AI.

The cited works show that the formation of representative and unbiased samples is a relevant research area, however, most of the works are devoted either to the detection of unbiasedness or to the analysis of the representativeness or unbiasedness of data samples, however, data samples must be modified to achieve compliance with FATE principles.

So, summarizing, it is possible to highlight the features of the modern approach, which is applied to the development of AI models (Fig. 1). However, this approach does not take into account existing ethical principles and non-discriminatory, representative presentation of existing population subgroups, which should be applied to obtain AI models.

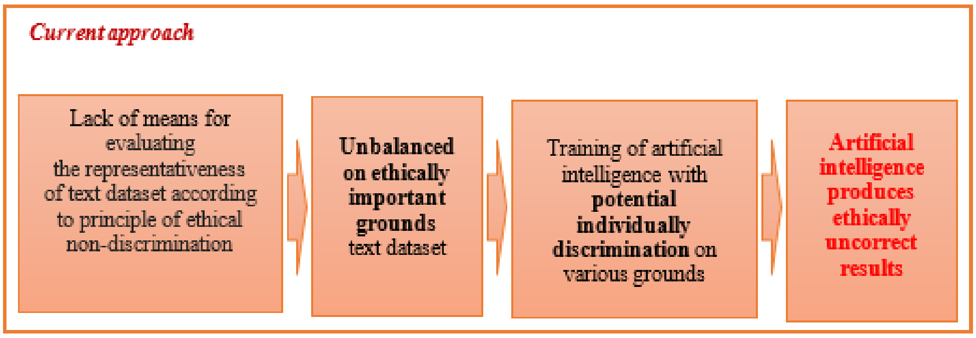

Current approach
Lack of means for evaluating the representativeness of text dataset according to principle of ethical non-discrimination
Unbalanced on ethically important grounds text dataset
Training of artificial intelligence with potential individually discrimination on various grounds
Artificial intelligence produces ethically uncorrect results

Figure 1: An existing approach to training AI models.

The purpose of the work is to ensure compliance with the ethical aspects (gender, religious, age, etc.) of the FATE principle of justice [5] for educational datasets, which consists in creating a method of analysis and formation of representative (according to the specified aspects) text samples of data. To achieve the specified goal, it is necessary to propose a method that will implement the following research tasks:

- to develop an approach to the analysis and formation of relevant representative datasets according to the principle of fairness of FATE for subject areas.
- to investigate the effectiveness of the proposed approach, by using it for the applied analysis of the text dataset and bringing it to a representative view according to the aspects of the FATE principle of justice: gender, age and religion.

3. Method for analysis and formation of representative text datasets

In contrast to the existing approach to training AI models (see Fig. 1), the study proposes a new approach (Fig. 2), which will ensure the representativeness and ethical correctness of the datasets used for training AI models.

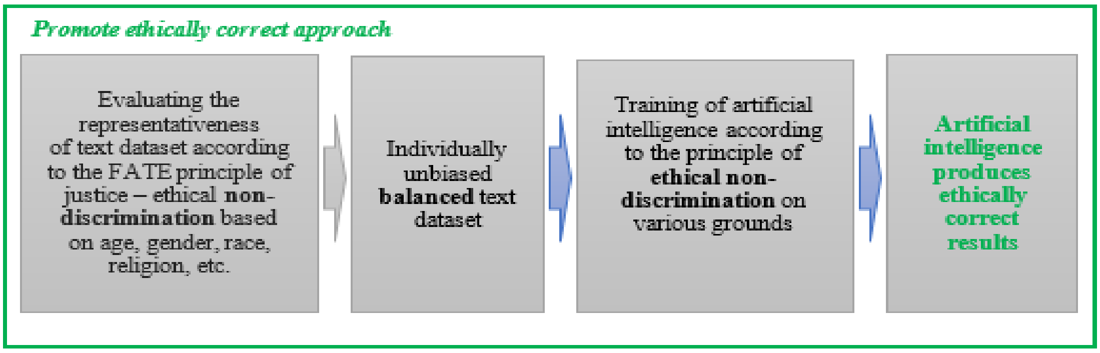

Promote ethically correct approach
Evaluating the representativeness of text dataset according to the FATE principle of justice – ethical non-discrimination based on age, gender, race, religion, etc.
Individually unbiased balanced text dataset
Training of artificial intelligence according to the principle of ethical non-discrimination on various grounds
Artificial intelligence produces ethically correct results

Figure 2: An approach to the formation of ethically representative datasets.

In order to implement the proposed approach, we will present: the information model and presentation of the task of forming representative samples of text data as an optimization task; steps of the method of analysis and formation of representative samples; a way to obtain a typical ML model for the ethical aspect; description of the composition of the datasets for the study.

3.1. Information model

The problem of obtaining a representative, ethically unbiased text dataset can be presented in the framework of an information model of the following form:

$$\{DS, DS', C, A, M, F\}, \tag{1}$$

where *DS* is the text dataset for analysis and correction, *DS′* is the text dataset after correction, *C* is the set of classes of the subject domain of the dataset, *A* is the set of ethical aspects, *M* is the set of trained machine learning models (separate for each ethical aspect), *F* is the objective function minimizing the deviation between current and desired ratios for all ethical aspects.

In (1), the initial dataset *DS* and the corrected dataset *DS′* can be represented as:

$$\{DS\ =\ \{D\ \cup\ Metadata\}, \tag{2}$$

$$DS' = \{D' \cup Metadata\}, \tag{3}$$

where *D* is the set of elements of the *DS* dataset, *Metadata* is the set of metadata of the *DS* dataset, *D′* is the set of elements of the *DS′* dataset, *Metadata′* is the set of metadata of the *DS′* dataset.

Each element of the set of elements of the dataset *D* in (2) and each element of the set of elements of the dataset *D′* in (3) is a tuple of the following form:

$$d\ =\ d'\ =\ (text, c_x, AC_x), \tag{4}$$

where the attribute *text* is the textual content of element d or d'; c_x is the class of the subject area of the dataset to which the element belongs, $c_x \in C$; AC_x is a set of classes of dataset element belonging to ethical aspects.

Thus, in (4) c_x and AC_x are the marking (marking) of the content of the *text* element.

In (4), the set of classes of membership of the dataset element DS or DS' in (1) to the ethical aspects A_x is presented in the form of a tuple:

$$A_x = (a_{1x}, a_{2x}, \dots, a_{kx},), \tag{5}$$

where a_x – classes of element belonging to ethical aspects; k is the number of ethical aspects to be analyzed, $k = |A_x|$.

At the same time, in (5) according to (1) $A_x \subset A$, and classes of dataset elements belonging to ethical aspects are elements of the corresponding sets, unique for each of the ethical aspects:

$$a_{\ 1,x} \in A_1, a_{2,x} \in A_2, \dots, a_{k,x} \in A_k, \tag{6}$$

$$A_1 \cup A_2 \cup \dots \cup A_k = A \tag{7}$$

The *Metadata* set of the *DS* dataset in (2) includes:

$$Metadata_{DS} = \{n_{DS}, AN_{DS}, AT_{DS}, n'_{DS}, AN'_{DS}, AT'_{DS}\}, \tag{8}$$

where n_{DS} is the number of elements in D, $n_{DS} = |D|$; AN_{DS} is the set of quantities of dataset elements belonging to each class of each ethical aspect from A_x; AT_{DS} is the set of available proportions of items for each class relative to the total for each ethical aspect from A_x, n'_{DS} is the target number of elements in D'; AN'_{DS} is the set of target quantities of dataset elements belonging to each class of each ethical aspect from A_x; AT'_{DS} is the set of target proportions of elements for each class relative to the total amount for each ethical aspect from A_x.

At the same time, in (8), each element $an_{DS,i}$ of the set AN_{DS} corresponds to a separate i-th ethical aspect and is represented by a tuple of the following form:

$$an_{DS,i} = (n_{DS,i,1}, n_{DS,i,2}, \dots, n_{DS,i,j}, \dots, n_{DS,i,k}), \tag{9}$$

where $n_{DS,i,1}$ is the number of elements in the dataset of the 1st class of the i-th ethical aspect, $n_{DS,i,2}$ is the number of elements in the dataset of the 2nd class of the i-th ethical aspect, $n_{DS,i,j}$ is the number of elements in the dataset of the j-th class of the i-th ethical aspect, k is the number of classes of the i-th ethical aspect.

Similarly to (9), in (8) the proportions of the elements $at_{DS,i}$ of the i-th ethical aspect are represented by a tuple of the following form:

$$at_{DS,i} = (t_{DS,i,1}, t_{DS,i,2}, \dots, t_{DS,i,j}, \dots, t_{DS,i,k}), \tag{10}$$

where $t_{DS,i,1}$ is the ratio of the number of elements in the dataset of the 1st class of the i-th ethical aspect to the total number of elements in the dataset, $t_{DS,i,2}$ is the ratio of the number of elements in the dataset of the 2nd class of the i-th ethical aspect to the total number of elements in the dataset, $t_{DS,i,j}$ is the ratio of the number of elements in the dataset of the i-th class of the i-th ethical aspect to the total number of elements in the dataset.

At the same time, for the values (9) and (10) in accordance with (8) for each i-th ethical aspect, the equality holds:

$$n_{DS,i,1} + n_{DS,i,2} + \dots + n_{DS,i,k} = n_{DS}, \tag{11}$$

$$t_{DS,i,1} + t_{DS,i,2} + \dots + t_{DS,i,k} = 1. \tag{12}$$

In contrast to (8), the set of *Metadata'* of the *DS'* dataset in (3) includes:

$$Metadata'_{DS} = \{n''_{DS}, AN''_{DS}, AT''_{DS}\}, \tag{13}$$

where n''_{DS} is actually the number of elements in D' obtained as a result of adjustment, $n''_{DS} = |D|$; AN''_{DS} is the set actually obtained as a result of adjusting the quantities of dataset elements belonging to each class of each ethical aspect from A_x; AT''_{DS} is the set actually obtained as a result of adjusting the proportions of elements for each class relative to the total number for each ethical aspect from A_x.

Thus, in (8) and (13), (9) and (11) hold for AN'_{DS} and AN''_{DS}, and (10) and (12) hold for AT'_{DS} and AT''_{DS}.

Thus, according to (4), (6) and (7), the text dataset D has the number of elements $n = n_{DS} = |D|$ and can be presented in the form:

$$D = \{d_1, d_2, \dots, d_n,\}, d_i = (text_i, c_i, A_1, A_2, \dots, A_m), i = \overline{1, \dots, n} \tag{14}$$

where $C = \{c_1, c_2, \dots, c_k\}$, where k is the number of classes of dataset D, m is the number of ethical aspects.

According to (6) – (10), the solution of the problem is aimed at obtaining the dataset D', which contains the total number of elements $n' = n'_{DS} = |D'|$, quantitatively balanced according to the ethical aspects Ai from the set of ethical aspects A:

$$A = \{A_1, A_2, \dots, A_m\}, A_i = (C_i, T_{ij}), i = \overline{1, \dots, m}, \tag{15}$$

where each aspect A_i contains classes C_i and target proportions of classes T_{ij} or each element of class C; C is the set of classes of the ethical aspect A_i, $C = \{c_1, c_2, \dots, c_j\}$; j is the number of classes of the ethical aspect of A_i.

To balance the dataset for each ethical aspect, it is necessary to use trained or train an appropriate number of classifier models, which can be as deep learning models, for example, BERT, LSTM, GRU, as well as machine learning models of Logistic Regression, Naive Bayes, Support Vector Machines, k-Nearest Neighbors etc. [17], and according to (1), the set of trained models of classifiers M is presented in the form:

$$M = \{M_1, M_2, \dots, M_m\},\ m = |D|. \tag{16}$$

Thus, within the framework of the proposed information model, it is necessary to perform the transformation $D \Rightarrow D'$ with the condition of maximal correspondence $n''_{DS} \longrightarrow n'_{DS}$, $AN''_{DS} \longrightarrow AN'_{DS}$ та $AT''_{DS} \longrightarrow AT'_{DS}$.

3.2. Idea of the approach

The study proposes to reduce the task of building a representative, ethically unbiased dataset to the task of multi-criteria optimization. The optimization task consists in minimizing the deviation between the current and desired class ratios, taking into account the limitations on the number of samples in the classes and the possibilities of generating synthetic data.

Input data: textual dataset DS, set of ethical aspects A, requirements for representative distribution DS'.

The goal of the problem: to create a representative sample for all ethical aspects that achieves the target class proportions for each ethical aspect $D \Rightarrow D'$.

Variables: x_{ij} – number of samples of class C_j in aspect A_i after sequestration and augmentation.

The objective function F is the minimization of the deviation between the current and desired ratios for all ethical aspects simultaneously, taking into account constraints (18) – (21):

$$F = argmin \sum_{i-1}^{m} \sum_{j-1}^{n_i} \left| \frac{x_{ij}}{n'} - T_{ij} \right|. \tag{17}$$

Limitations of the task:

1) the sum of all class samples within one aspect is equal to the target number of samples for this aspect (4):

$$\sum_{j=1}^{n_i} x_{ij} = n', \forall i \in \{1,2,\dots,m\}, \tag{18}$$

where n_i is the number of classes in the aspectA_i;

2) the number of samples for each class should correspond to the target proportion of classes:

$$\frac{x_{ij}}{n'} \approx T_{ij}, \forall i \in \{1,2,\dots,m\}, \forall j \in \{1,2,\dots,n_i\}; \tag{19}$$

3) the estimated number of samples cannot be negative:

$$x_{ij} \geq 0, \forall i \in \{1,2,\dots,m\}, \forall j \in \{1,2,\dots,n_i\}; \tag{20}$$

4) the ability to add new samples should match the ability to generate new data for each class and aspect:

$$x_{ij} \leq G_{ij}, \forall i \in \{1,2,\dots,m\}, \forall j \in \{1,2,\dots,n_i\}, \tag{21}$$

where G_{ij} is the maximum possible number of samples of class C_j in aspect A_i, that can be added.

Based on the set optimization task of forming a representative dataset (17), we present the steps of the method of analysis and formation of representative samples of text data.

3.3. Main steps of method

Method for analysis and formation of representative text datasets is presented in the form of three consecutive stages: preprocessing, analysis of representativeness according to ethical aspects and representative adjustment of dataset. Each stage consists of its own steps, which are shown in Figure 3.

The input data of the method is the dataset *DS* for analysis, which according to (2) and (8) contains the target number of n'_{DS} elements, the set of ethical aspects *A* with subsets of classes, the target proportions of AT_{DS} classes and the number of AN'_{DS} elements in the classes of ethical aspects, respectively, the trained set of models *M* for each ethical aspect from *A*, which uses balanced samples for each ethical aspect for training.

At *stage 1*, a sample of text data in $D \subset DS$ is pre-processed, namely, the removal of non-informative text fragments such as punctuation marks, numbers and special characters [18]. Removal of emoticons is not performed, as in many cases including emoticons in the analysis improves the accuracy of machine learning models used to classify texts based on emotional or mood content [19]. Incorrect records (empty, uninformative, etc.) are also deleted.

At *stage 2*, an analysis of the representativeness of the sample of textual data is carried out, taking into account ethical aspects. First, it is necessary to vectorize and classify each element $\forall d \in D$ of the data sample using separate machine learning models $m \in M$ for each of the ethical aspects $A_i \in A$. The existing proportions of AN_{DS} and AT_{DS} classes for each of the ethical aspects are determined. The amount of shortage or excess of elements of each class for each of the ethical aspects is calculated. After that, the sufficiency of the data in the sample for augmentation is analyzed (minimum availability of samples of the relevant classes, etc.).

Stage 3 involves a representative adjustment of the data sample to take into account ethical considerations. Adjustments include removing and adding.

The deletion operation is performed to remove redundant elements of each class for each of the ethical aspects with minimal damage to other distributions, for which the optimization problem of selecting redundant elements in the framework of (17), which should be removed to achieve the target proportions of classes, is solved.

The add operation is performed to create new items using one of the known methods, for example, the SMOTE method [20]. Requirements are created in the form of the necessary combination of classes of each of the ethical aspects for each new element, for which the optimization problem of forming requirements for the missing elements is solved within the framework of (17).

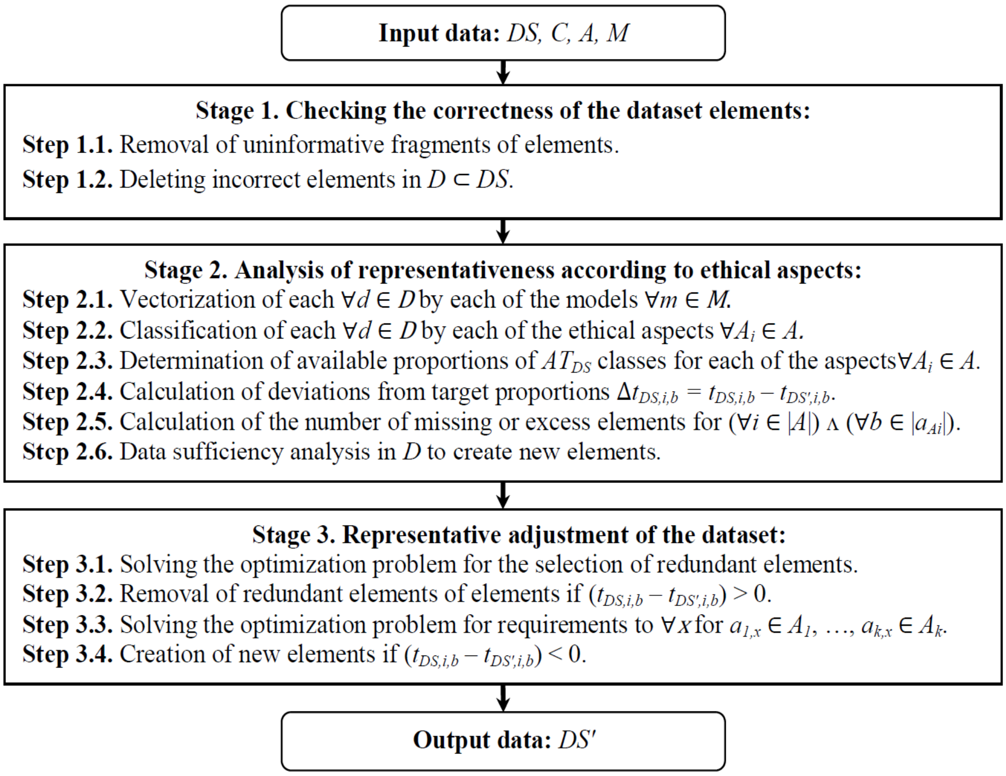

Input data: DS, C, A, M
Stage 1. Checking the correctness of the dataset elements:
Step 1.1. Removal of uninformative fragments of elements.
Step 1.2. Deleting incorrect elements in $D \subset DS$.
Stage 2. Analysis of representativeness according to ethical aspects:
Step 2.1. Vectorization of each $\forall d \in D$ by each of the models $\forall m \in M$.
Step 2.2. Classification of each $\forall d \in D$ by each of the ethical aspects $\forall A_i \in A$.
Step 2.3. Determination of available proportions of AT_{DS} classes for each of the aspects$\forall A_i \in A$.
Step 2.4. Calculation of deviations from target proportions $\Delta t_{DS,i,b} = t_{DS,i,b} - t_{DS',i,b}$.
Step 2.5. Calculation of the number of missing or excess elements for $(\forall i \in |A|) \wedge (\forall b \in |a_{Ai}|)$.
Step 2.6. Data sufficiency analysis in D to create new elements.
Stage 3. Representative adjustment of the dataset:
Step 3.1. Solving the optimization problem for the selection of redundant elements.
Step 3.2. Removal of redundant elements of elements if $(t_{DS,i,b} - t_{DS',i,b}) > 0$.
Step 3.3. Solving the optimization problem for requirements to $\forall x$ for $a_{1,x} \in A_1, \ldots, a_{k,x} \in A_k$.
Step 3.4. Creation of new elements if $(t_{DS,i,b} - t_{DS',i,b}) < 0$.
Output data: DS'

Figure 3: Steps of method for analysis and formation of representative text datasets.

The *output data* of the method is a text dataset $D' \subset DS'$, which has the required volume n'_{DS} and is balanced according to the required proportions AT'_{DS} according to the selected ethical aspects $A_x \subset A$.

The steps of the method of analyzing and generating representative samples of text data will allow you to generate text samples that are non-discriminatory and unbiased and reflect a proportional representation of the sample samples to the actual demographic subgroups of the population, which will affect the accuracy and transparency of training machine learning models for solving various problems.

4. Results and discussion

4.1. Datasets for research

To test the method of analysis and formation of representative samples of textual data, an input dataset was formed based on two datasets "Cyberbullying Classification" [3] and "Cyberbully Detection Dataset" [4]. The "Cyberbullying Classification" dataset contains 46,017 tweets, which are labeled by types of cyberbullying into 6 classes. The "Cyberbully Detection Dataset" contains 99,989 tweets, which is also labeled by type of cyberbullying. Both datasets are unlabeled for gender, age group, religion, and ethnicity of the message author.

To train machine learning models, which will be used to label the input dataset, datasets were used on the example of three ethical aspects of the principle of justice of gender, age and religion.

The English-language dataset "Tweet Files for Gender Guessing" [21], which contains 34,146 unique text entries, which are divided into two classes: female and male, with 17,073 entries in each class, was used to train ML based on the ethical aspect of the gender of the author of the message. On the basis of the English-language dataset "CyberBullying Detection Dataset" [22], which contains 20109 test samples, a sample was created for training the classifier and marking the input dataset according to the religious ethical aspect. The dataset in Italian "TAG-it Dataset Distribution" [23] was translated into English and used to bring to a representative view the working dataset by age and contains 21,948 text messages divided into age classes: 0-19, 20-29, 30-39, 40-49, 50-100 years old.

Since the classes in the given datasets are not balanced and have a different number of samples, which will negatively affect the quality of training of machine learning models, all classes in the datasets were balanced in terms of number. The final number of samples in each class of training samples for ML training according to ethical aspects is shown in Fig. 4.

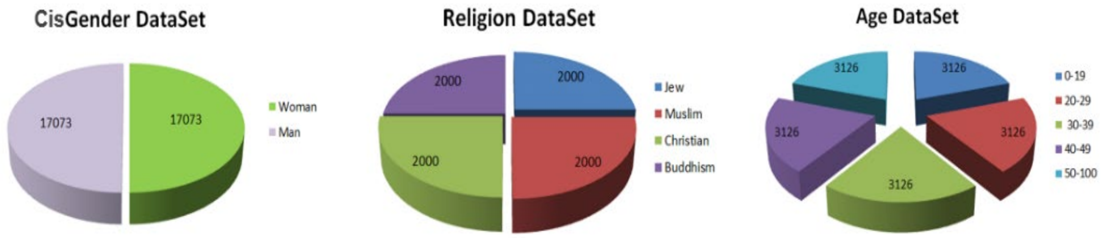
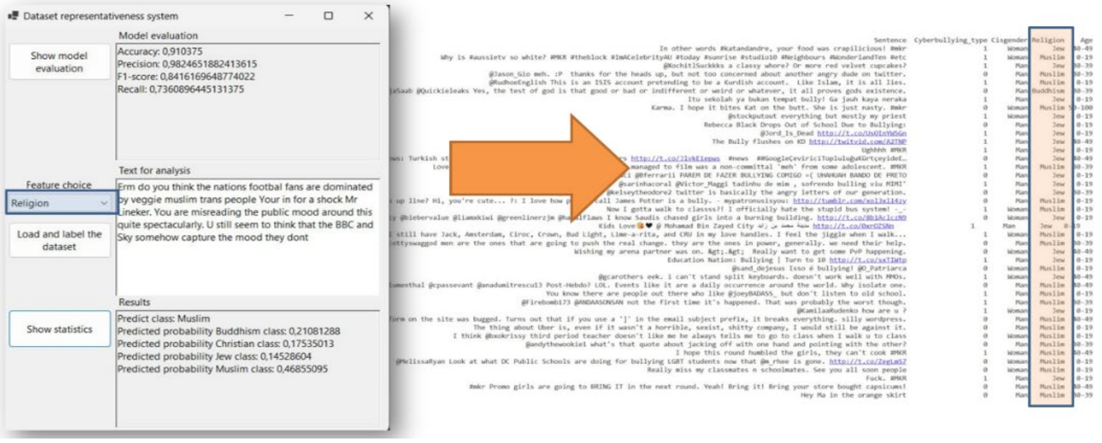


Figure 4: Classes and number of samples in ML training datasets for ethical aspects.

As a result of the work on creating training samples, datasets balanced by the number of text messages in classes were obtained. Such datasets will make it possible to correctly assess the representativeness of working text datasets.

4.2. Software for research

To study the effectiveness of the method of analysis and formation of a representative sample of text data, a software implementation was created using the Python programming language. The tensorflow library (https://www.tensorflow.org/) was used to classify the input dataset on cyberbullying based on gender, age, and religion. In Fig. 5 shows an example of classification based on the religious basis of the FATE-principle of justice.



Model evaluation

Accuracy: 0,910375
Precision: 0,9824651882413615
F1-score: 0,8416169648774022
Recall: 0,7360896445131375

Feature choice: Religion

Text for analysis

Erm do you think the nations footbal fans are dominated by veggie muslim trans people Your in for a shock Mr Lineker. You are misreading the public mood around this quite spectacularly. U still seem to think that the BBC and Sky somehow capture the mood they dont

Results

Predict class: Muslim
Predicted probability Buddhism class: 0,21081288
Predicted probability Christian class: 0,17535013
Predicted probability Jew class: 0,14528604
Predicted probability Muslim class: 0,46855095

Sentence	Cyberbullying_type	Cisgender	Religion	Age
In other words #katandandre, your food was crapilicious! #mkr	1	Woman	Jew	40-49
Why is #aussietv so white? #MKR #theblock #ImACelebrityAU #today #sunrise #studio10 #Neighbours #WonderlandTen #etc	1	Woman	Muslim	0-19
@XochitlSuckkks a classy whore? Or more red velvet cupcakes?	1	Man	Jew	30-39
@Jason_Gio meh. :P thanks for the heads up, but not too concerned about another angry dude on twitter.	0	Man	Muslim	30-39
@RudhoeEnglish This is an ISIS account pretending to be a Kurdish account. Like Islam, it is all lies.	1	Man	Muslim	0-19
jaSaab @Quickieleaks Yes, the test of god is that good or bad or indifferent or weird or whatever, it all proves gods existence.	0	Man	Buddhism	30-39
Itu sekolah ya bukan tempat bully! Ga jauh kaya neraka	1	Man	Jew	0-19
Karma. I hope it bites Kat on the butt. She is just nasty. #mkr	0	Woman	Muslim	50-100
@stockputout everything but mostly my priest	1	Woman	Jew	0-19
Rebecca Black Drops Out of School Due to Bullying:	0	Man	Jew	0-19
@Jord_Is_Dead http://t.co/UsQInYW6Gn	1	Man	Jew	0-19
The Bully flushes on KD http://twitvid.com/A2TNP	1	Man	Jew	40-49
Ughhhh #MKR	1	Man	Jew	0-19
ws: Turkish st s http://t.co/JlvkElepws #news ##GoogleÇeviriciTopluluğuKürtçeyideE…	0	Man	Jew	40-49
Love managed to film was a non-committal 'meh' from some adolescent. #MKR	1	Man	Muslim	30-39
i @Bferrarii PAREM DE FAZER BULLYING COMIGO =(UHAHUAH BANDO DE PRETO	0	Man	Jew	0-19
@sarinhacoral @Victor_Maggi tadinhu de mim , sofrendo bulling viu MIMI'	1	Man	Jew	0-19
@kelseytheodore2 twitter is basically the angry letters of our generation.	0	Man	Jew	30-39
up line? Hi, you're cute... ?: I love how p call James Potter is a bully. - mypatronusisyou: http://tumblr.com/xol3xll4zy	0	Man	Muslim	0-19
Now I gotta walk to classss?! I officially hate the stupid bus system! -_-	1	Woman	Muslim	0-19
y @biebervalue @liamxkiwi @greenlinerzjm @ha lflaws I know Saudis chased girls into a burning building. http://t.co/Bb1AcIczN9	0	Woman	Jew	0-19
Kids Love😘❤ @ Mohamad Bin Zayed City مدينة محمد بن زايد http://t.co/0xrOZSNn	1	Man	Jew	0-19
I still have Jack, Amsterdam, Ciroc, Crown, Bud Light, Lime-a-rita, and CRU in my love handles. I feel the jiggle when I walk...	1	Woman	Muslim	0-19
ottyswaggod men are the ones that are going to push the real change. they are the ones in power, generally. we need their help.	0	Man	Muslim	30-39
Wishing my arena partner was on. &gt;.&gt; Really want to get some PvP happening.	0	Woman	Jew	40-49
Education Nation: Bullying | Turn to 10 http://t.co/sxTIWtp	1	Man	Jew	0-19
@sand_dejesus Isso é bullying! @O_Patriarca	0	Woman	Muslim	0-19
@gcarothers eek. i can't stand split keyboards. doesn't work well with MMOs.	1	Woman	Jew	40-49
lumenthal @cpassevant @anadumitrescu13 Post-Hebdo? LOL. Events like it are a daily occurrence around the world. Why isolate one.	0	Man	Muslim	40-49
You know there are people out there who like @joeyBADASS_ but don't listen to old school.	1	Man	Muslim	0-19
@Firebomb173 @ANDAASONSAN not the first time it's happened. That was probably the worst though.	1	Man	Muslim	30-39
@KamilaaRudenko how are u ?	1	Woman	Jew	0-19
form on the site was bugged. Turns out that if you use a ']' in the email subject prefix, it breaks everything. silly wordpress.	0	Man	Muslim	40-49
The thing about Uber is, even if it wasn't a horrible, sexist, shitty company, I would still be against it.	0	Man	Muslim	0-19
I think @bxokrissy third period teacher doesn't like me he always tells me to go to class when I walk u to class	0	Woman	Muslim	0-19
@andythewookie1 what's that quote about jacking off with one hand and pointing with the other?	0	Man	Muslim	30-39
I hope this round humbled the girls, they can't cook #MKR	1	Woman	Muslim	40-49
@MelissaRyan Look at what DC Public Schools are doing for bullying LGBT students now that @m_rhee is gone. http://t.co/Zegim57	0	Woman	Muslim	0-19
Really miss my classmates n schoolmates. See you all soon people	0	Woman	Muslim	0-19
Fuck. #MKR	1	Man	Jew	0-19
#mkr Promo girls are going to BRING IT in the next round. Yeah! Bring it! Bring your store bought capsicums!	0	Man	Muslim	40-49
Hey Ma in the orange skirt	0	Man	Muslim	30-39

Figure 5: Developed software for classification of dataset by religious ethical aspect.

To form a set of trained ethical machine learning models, which are separate for each ethical aspect, various classifier models were analyzed, and to select the best of them, their quality was evaluated by statistical indicators, such as Accuracy, Precision, Recall, and F1-score [24]. Both deep learning models, such as BERT, GPT, LSTM, GRU, etc., and classifiers such as Logistic Regression, Naive Bayes, Support Vector Machines k-Nearest Neighbors, etc., were studied [25]. After that, the classifier is trained on the selected ML model on the annotated dataset for the ethical aspect.

As a result, different architectures were chosen as classifiers: FastForest classifiers, SVM and LSTM, BERT deep learning models [26]. Thus, machine learning models such as FastForest, SVM, LSTM, and BERT are effective tools for solving text classification tasks, including determining a person's gender, religion, and age based on user text posts. Classical approaches such as FastForest and SVM have also demonstrated their effectiveness in text classification. FastForest works efficiently with large datasets and prevents overtraining. SVM, in turn, is known for its ability to work with high-dimensional data, which is especially useful for text classification, where each word or phrase can be represented as a separate feature [27]. Deep learning models, such as LSTM and BERT, are able to recognize complex patterns in text sequences, preserving the context at all stages of analysis [28]. A distinctive feature of LSTM is its ability to retain information about previous parts of the text, which makes this model effective for complex classification tasks where the overall context of the message is important. Studies have shown that such a model can achieve an accuracy of up to 92% in text classification tasks [29]. The BERT model, in turn, is characterized by the ability to analyze the text in two directions, that is, to take into account both the previous and subsequent context of words [30].

4.3. Analysis of research results

To analyze and form a representative sample of text data for the target proportions of classes, to form a representative sample of text data by age and gender, the population of Ukraine was taken. According to the M. V. Ptukh Institute of Demography and Social Research of the National Academy of Sciences of Ukraine (https://idss.org.ua/forecasts/nation_pop_proj), as of July 2023, the total population of Ukraine is estimated at 3,559,6216 people. The following number of people is represented in each age subgroup: age group 0-19 years 6,659,068 people, 20-29 years 3,623,143 people, 30-39 years - 6,022,345 people, 40-49 years - 5,431,140 people, 50 -100 years - 13,860,520 people. Regarding the gender structure of the population of Ukraine in 2023: 1,695,1527 are women, and 1,864,689 are men (idss.org.ua). Note that within the scope of this work, the cisgender group is considered in the analysis of the gender ethical aspect.

To study the effectiveness of the method of analysis and formation of a representative selection of text data described in the work, several machine learning models were trained. The results of calculating static metrics such as Accuracy, Precision, Recall and F1-score [24] of machine learning models for the gender, age and religious ethical aspects are shown in Table 1.

Table 1
Statistical metrics Accuracy, Precision, Recall and F1-score of machine learning models by gender, age and religious ethical aspects

ML model	Accuracy	Precision	Recall	F1-score
Gender ethical aspect				
FastForest	0.630	0.640	0.600	0.620
SVM	0.580	0.580	0.580	0.580
LSTM	0.70	0.770	0.670	0.720
BERT	0.690	0.640	0.710	0.670
Age ethical aspect				
FastForest	0.535	0.542	0.504	0.504
SVM	0.815	0.770	0.779	0.770
LSTM	0.590	0.600	0.560	0.580
BERT	0.580	0.430	0.450	0.440
Religious ethical aspect				
FastForest	0.775	0.800	0.762	0.780
SVM	0.825	0.850	0.810	0.829
LSTM	0.850	0.880	0.830	0.854
BERT	0.910	0.980	0.74 0	0.840

For different classes, different levels of linear resolution were obtained: according to religion using the BERT classifier, which showed the best result of the trained machine learning models for the task of classifying text samples according to the religious ethical aspect, the data turned out to be well separated, according to gender using the LSTM classifier, which showed the best performance compared to other models, the data turned out to be moderately separable, and according to age, using the SVM classifier, it was poorly separable.

In addition, it was found that the dataset is not representative, because the classes of various ethical aspects have a number of text samples that do not correspond to the proportions of the demographic subgroups of the population of Ukraine, thus they need balancing to acquire a representative appearance.

Therefore, according to the steps of the method of analysis and formation of a representative sample of text data, a sample of text data needs data augmentation to form a representative sample. For this, it is necessary to solve the optimization problem, for the correct removal of redundant elements of each class according to each of the ethical aspects, with further augmentation of the data sample to the target requirements (number of elements and proportions of classes).

Table 2 presents the percentages of samples by age in the sample of textual data and individuals of the population in age-demographic subgroups, and also calculates the new distribution of the sample classes if only one ethical aspect - age - was taken into account.

Table 2
Percentage ratios of samples by age in the sample of text data and individuals of the population in age demographic subgroups, %

Age demographic subgroups	Percentage of samples by age in text dataset	Percentage of population in age demographic subgroups	Deviation of text samples from subgroups of population	New distribution of sampling classes	Deviation from representative distribution
0-19 years	48.23%	18.71%	29.52%	18.75%	0.04%
20-29 years	2.11%	10.17%	8.06%	10.15%	0.02%
30-39 years	25.28%	16.92%	8.36%	16.87%	0.03%
40-49 years	13.60%	15.26%	1.66%	15.28%	0.02%
50-100 years	10.78%	38.94%	28.16%	38.95%	0.01%

Table 3 presents the percentages of samples by gender in the sample of textual data and individuals of the population in gender demographic subgroups, and also calculates the new distribution of sample classes if only one ethical aspect - gender - was taken into account.

The deviation of the sample distributions by classes of the age-ethical aspect of the dataset, transformed according to the created method, from the ideal representative distribution was obtained: minimum 0.01%, maximum 0.04%, average 0.02%, and for the gender ethical aspect: minimum 0.03%, maximum 0.03%, average 0.03 %.

Table 3
Percentages of samples by gender in the sample of text data and individuals of the population in gender demographic subgroups, %

Gender demographic subgroups	Percentage of samples by gender in text dataset	Percentage of population in gender demographic subgroups	Deviation of text samples from subgroups of population	New distribution of sampling classes	Deviation from representative distribution
Men	58.94%	43.28%	15.67%	43.25%	0.03%
Women	41.06%	56.72%	15.67%	56.75%	0.03%

However, the optimization task of forming a representative sample of textual data is a multi-criteria one, in which the criteria are the formation of a sample based on age and gender ethical aspects, so the goal is to minimize the deviation between the current and desired class ratios, taking into account the limitations on the number of samples and the possibility of generating new data. As a result of solving the optimization problem for the formation of a representative sample by age and gender ethical aspects on the example of demographic subgroups of the population of Ukraine, a representative sample of text data was obtained by augmentation, the balance of classes of which is presented in Table 4, Fig. 6 and Fig. 7.

Table 4
Distribution of samples in the formed representative sample after data augmentation as a result of solving a multi-criteria optimization problem

Age demographic subgroups	0-19 years	20-29 years	30-39 years	40-49 years	50-100 years
Percentage ratio of demographic groups by gender and age in the population of Ukraine					
Men	9.67%	5.64%	8.96%	7.79%	15.56%
Women	9.04%	4.53%	7.96%	7.47%	23.38%
Percentage ratio of demographic groups by gender and age in the text sample					
Men	9.65%	5.62%	8.94%	7.80%	15.57%
Women	9.05%	4.57%	7.97%	7.45%	23.38%
The resulting deviation from a representative distribution					
Men	0.02%	0.02%	0.02%	0.01%	0.02%
Women	0.01%	0.04%	0.01%	0.02%	0.00%

The deviation of the sample distributions by classes of age and gender ethical aspects of the dataset simultaneously, transformed according to the created method, from the ideal representative distribution was obtained: minimum 0.00%, maximum 0.04%, average 0.02%.

So, as a result of performing the steps of the analysis method and forming representative samples of text data, a text sample was formed, which is non-discriminatory and unbiased and reflects the representation of sample samples proportional to the real demographic subgroups of the population of Ukraine.

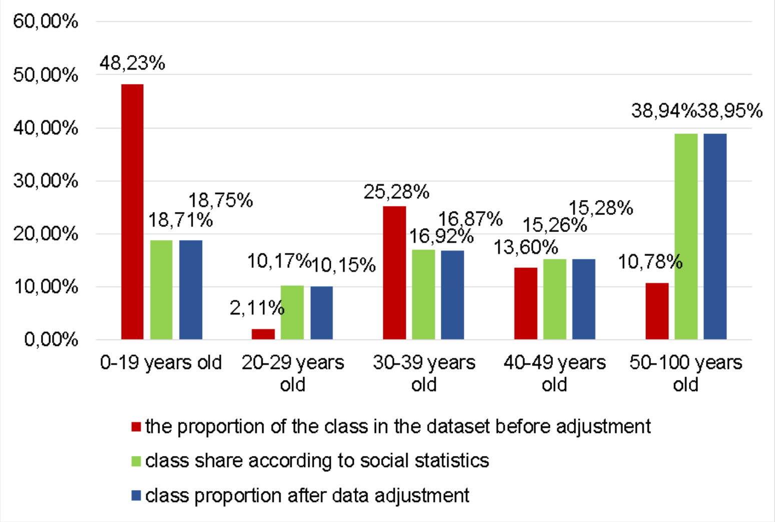


Figure 6: The balance of the distribution of the input dataset according to the age-ethical aspect of the FATE-principle of justice.

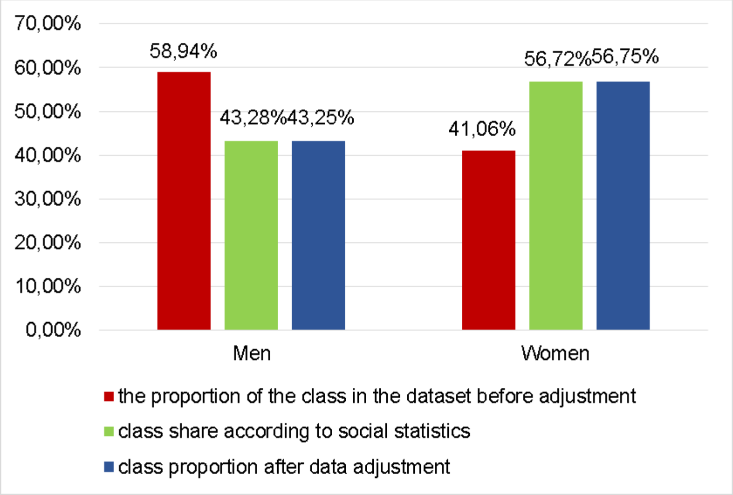


Figure 7: The balance of the distribution of the input dataset according to the gender ethical aspect of the FATE-principle of justice.

5. Conclusion

Thus, the goal of the study was achieved through the development of the method for analysis and formation of representative text datasets, designed for the analysis and formation of representative text samples of data according to the principle of fairness of FATE for subject areas.

To investigate the effectiveness of the analysis method and the formation of a representative presentation of the text dataset, software was created that uses machine learning models to classify texts according to various ethical aspects - age, gender, religion, ethnicity, etc. Thus, to classify the text samples in the sample according to the age-ethical aspect, SVM was used, LSTM was used for gender, and BERT was used for religious ones, which are the best indicators of statistical metrics.

As a result of the practical application of the developed method, it was established that the available dataset is not representative compared to the objective data of demographic statistics, so a multi-criteria optimization problem was solved and the dataset was transformed into a representative one in terms of age and gender ethical aspects. The obtained deviations of the sample distributions by classes of ethical aspects of the dataset transformed according to the created method from the ideal representative distribution were: minimum 0.00%, maximum 0.04%, average 0.02%, under the conditions of the initial volume of the dataset 47,692 elements, the minimum initial number of samples in the class 1007 elements, the maximum initial number of samples in the class is 28,112 elements. The studied efficiency proves that the developed method allows performing the analysis of the representativeness of text datasets and bringing them to a representative form according to various aspects of the FATE fairness principle.

The obtained results contribute to improvement of representativeness of text datasets and fair and unbiased representation of demographic groups in them, which increases trust in decisions made by artificial intelligence, and complies with goals SDG3 (good health and well-being), SDG4 (quality education) and SDG16 (peace, justice, and strong institutions).

Further plans for improving the method of analysis and formation of representative samples of text data are the formation of not only a non-discriminatory sample by the number of samples, but also the search and removal of samples of text samples that contain a biased attitude towards representatives of various demographic subgroups, according to the ethical aspects of the FATE-principle of justice.

Also, the prospects for further research are the use of the developed method for adjusting textual datasets of subject areas and their use for solving applied problems, such as detection and classification of cyberbullying, analysis of the emotional tonality of messages, detection of the physical and mental state of users based on their posts, etc. Detecting performance gains from using ethically balanced text datasets will provide feedback for improving the developed method.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check; DeepL Translate in order to: some phrases translation into English. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Shah, N. Sureja, A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions, Arch Computat Methods Eng (2024). doi:10.1007/s11831-024-10134-2

[2] Y. Yusyn, N. Rybachok, dictionary-based deterministic method of generation of text CORPORA, Computer systems and information technologies 3 (2024) 67–73.. doi:10.31891/csit-2024-3-9.

[3] Kaggle.com, Cyberbullying Classification, 2021. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download

[4] Kaggle.com, CyberBullying Detection Dataset, 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification

[5] B. C. Stahl, D. Eke, The ethics of ChatGPT – Exploring the ethical issues of an emerging technology, International Journal of Information Management 74 (2024), p. 102700. doi:10.1016/j.ijinfomgt.2023.102700

[6] E. Manziuk, I. Krak, O. Barmak, O. Mazurets, V. Kuznetsov, O. Pylypiak, Structural alignment method of conceptual categories of ontology and formalized domain, CEUR Workshop Proceedings 3003 (2021) pp. 11–22

[7] O. Barmak, O. Mazurets, I. Krak, A. Kulias, A. Smolarz, L. Azarova, K. Gromaszek, S. Smailova, Information technology for creation of semantic structure of educational materials, Proceedings of SPIE – The International Society for Optical Engineering 11176 (2019), pp. 147–156. doi:10.1117/12.2537064

[8] L. K. H. Clemmensen, D. K. Rune, Data Representativity for Machine Learning and AI Systems, 2022. URL: https://ar5iv.labs.arxiv.org/html/2203.04706

[9] D. Dablain, B. Krawczyk, N. Chawla, Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. Discov Data 2, 4 (2024). doi:10.1007/s44248-024-00007-1

[10] R. K. E. Bellamy, AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development, volume 63, 4/5, (2019) pp. 1-15. doi:10.1147/JRD.2019.2942287

[11] J. Lalor, Y. Yang, K. Smith, N. Forsgren, A. Abbasi, Benchmarking Intersectional Biases in NLP, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computa-tional Linguistics: Human Language Technologies, Association for Computational Linguistics (2022), pp. 3598–3609. doi:10.18653/v1/2022.naacl-main.263

[12] S. Raza, D.J. Reji, C. Ding, Dbias: detecting biases and ensuring fairness in news articles. International Journal of Data Science and Analytics, volume 17 (2024), pp. 39–59. doi:10.1007/s41060-022-00359-4

[13] H. Chen, Y. Ji, D. Evans. Addressing Both Statistical and Causal Gender Fairness in NLP Models. In Findings of the Association for Computational Linguistics: NAACL 2024,

Association for Computational Linguistics (2024), pp. 561–582. doi: 10.48550/arXiv.2404.00463

[14] K. Zhou, J. Wen, N. Yang, D. Yuan, Q. Lu, H. Chen, Fairpriori: Improving Biased Subgroup Discovery for Deep Neural Network Fairness (2024). doi:10.48550/arXiv.2407.01595

[15] A. S. Evans, H. Moniz, L. Coheur, A Study on Bias De-tection and Classification in Natural Language Processing (2024). doi:10.48550/arXiv.2408.07479

[16] A. Orelaja, C. Ejiofor, S. Sarpong, S. Imakuh, C. Bassey, I. Opara, J. N. A. Tettey, O. Akinola, Attribute-Specific Cyberbullying Detection Using Artificial Intelligence, Journal of Electronic & Information Systems, volume 6(1) (2024), pp. 10–21. doi:10.30564/jeis.v6i1.6206

[17] M. Zulqarnain, R. Sheikh, S. Hussain, M. Sajid, S. N. Abbas, M. Majid, U. Ullah, Text Classification Using Deep Learning Models: A Comparative Review, Cloud Computing and Data Science (2024), pp. 80-96. doi: 10.1007/s10115-023-01856-z

[18] O. Zalutska, M. Molchanova, O. Sobko, O. Mazurets, O. Pasichnyk, O. Barmak, I. Krak, Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network, CEUR Workshop Proceedings 3387 (2023), pp. 344–356. doi:10.15407/jai2024.02.085

[19] H. Tang, W. Tang, D. Zhu, S. Wang, Y. Wang, L. Wang, EMFSA: Emoji-based multifeature fusion sentiment analysis. PLoS One. 19(9) (2024). doi: 10.1371/journal.pone.0310715

[20] T. He. S. Wongvorachan, O. Bulut, A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining, Information, 14(1), 54 (2023). doi:10.3390/info14010054

[21] Kaggle.com, Tweet Files for Gender Guessing, 2019. URL: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing

[22] Kaggle.com, CyberBullying Detection Dataset, 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification

[23] Live.european-language-grid.eu, TAG-it Dataset Distribution, 2024. URL: https://live.european-language-grid.eu/catalogue/corpus/8112/download/

[24] K. Nazeri, H. Thasarathan, M. Ebrahimi, Edge-informed single image super-resolution, Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops , 2019. URL: https://openaccess.thecvf.com/content_ICCVW_2019/html/AIM/Nazeri_Edge-Informed_Single_Image_Super-Resolution_ICCVW_2019_paper.html

[25] O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina, Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets, Lecture Notes on Data Engineering and Communications Technologies 149 (2023) pp. 591–607. doi:10.1007/978-3-031-16203-9_33

[26] R. Sadeghi, A. Akbari, M.M. Jaziriyan, ExaAUAC: Arabic Twitter user age prediction corpus based on language and metadata features. Discover Artifcial Intelligence 4, 48 (2024). doi: 10.1007/s44163-024-00145-0

[27] V. Sheth, U. Tripathi, A. Sharma. A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. Procedia Computer Science (2022), pp. 422-431. doi: 10.1016/j.procs.2022.12.044

[28] Sh. Minaee, E. Cambria, J. Gao. Deep Learning-based Text Classification. ACM Computing Surveys (CSUR), 54 (2020), pp. 1–40. doi:10.1145/3439726

[29] H. Sa'diah, M. R. Faisal, A. Farmadi, F. Abadi, F. Indriani, M. Alkaff, V. Abdullayev, Gender Classification on Social Media Messages Using fastText-base Feature Extraction and Long Short-Term Memory, Journal ofElectronics, Electromedical Engineering, and Medical Informatics, volume 6(3) (2024), pp. 243–252. doi: 10.35882/jeeemi.v6i3.407

[30] A. Alqahtani, K. H. Ullah, Sh. Alsubai, M. Sha, A. Almadhor, T. Iqbal, S. Abbas. An efficient approach for textual data classification using deep learning. Frontiers in Computational Neuroscience, 16 (2022). doi:10.3389/fncom.2022.992296