Olena Sobko
Teacher of Computer Science Department
Khmelnytskyi National University, Ukraine

PRACTICE IMPLEMENTATION OF THE METHOD FOR ANALYSIS AND FORMATION OF REPRESENTATIVE TEXT DATASETS

Proposes the practice implementation 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 FATE principle of fairness for subject areas. The studied efficiency proves that developed method allows performing the analysis of the representativeness of text datasets and bringing them to representative form according to various aspects of the FATE fairness principle.

If text 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 [1]. The representativeness of text datasets according to ethical principle of FATE is achieved by correct balancing according to various ethical aspects: age, gender, religious, etc. [2].

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 1.

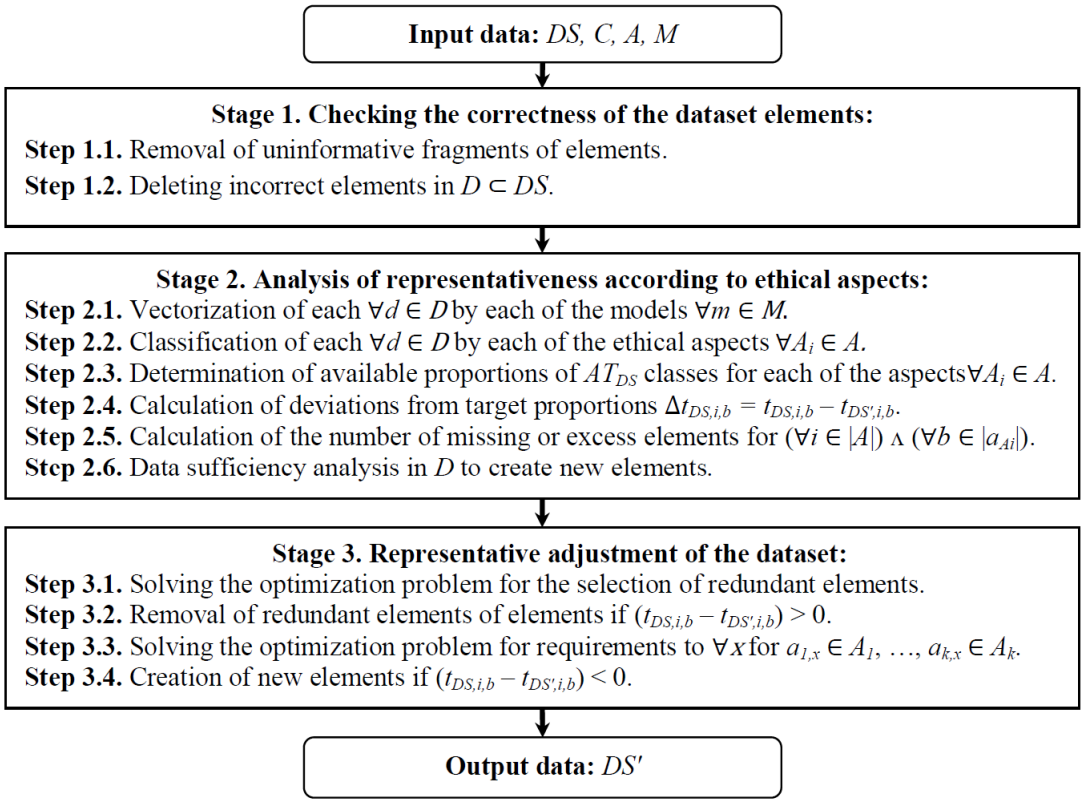

Figure 1 – Scheme of method for analysis and formation of representative datasets

To study of the method, software implementation was created using the Python programming language. The tensorflow library was used to classify the input dataset on cyberbullying based on gender, age, and religion. 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.

However, the optimization task of forming a representative sample of text 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 Figure 2.

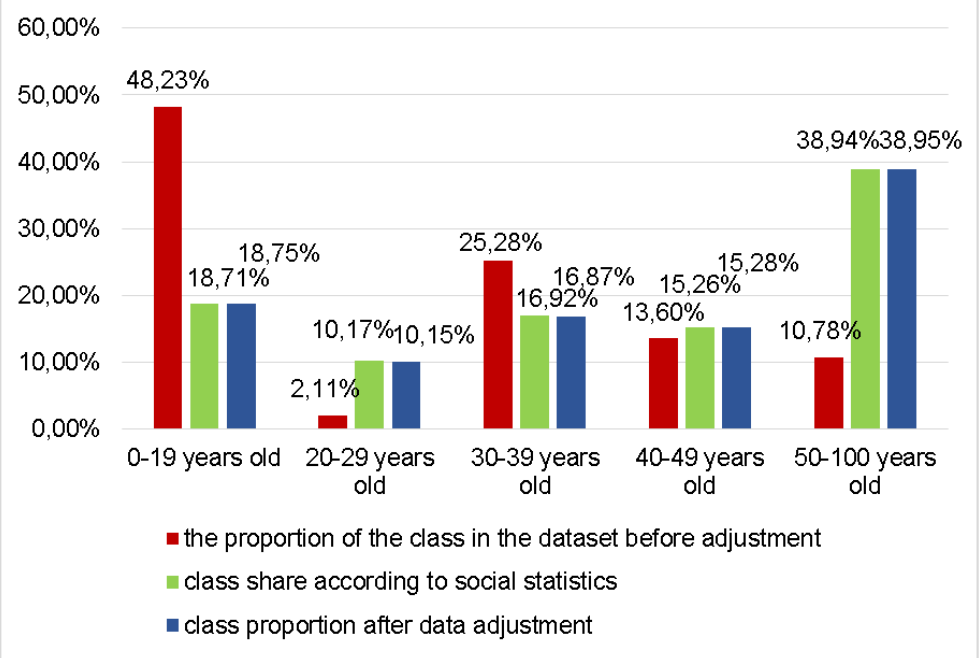

Figure 2 – The balance of the distribution of the input dataset

The obtained deviations of sample distributions by classes of ethical aspects of dataset transformed according to 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.

References:

1. V. Slobodzian, O. Kovalchuk, M. Molchanova, O. Sobko, O.Mazurets, O.Barmak, I. Krak, Text Data Vectorization Model of Ukrainian-Language Internet Communication Content, in: CEUR Workshop Proceedings, 2022, vol. 3171, pp. 561–571.

2. I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, R. Bahrii, O. Sobko, O. Barmak, Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network, in: CEUR Workshop Proceedings, vol. 3688, 2024, pp. 16–28.