https://doi.org/10.31891/csit-2025-3-14
UDC 004.7

Olena SOBKO
Khmelnytskyi National University
Archil CHOCHIA
Tallinn University of Technology

LEGAL AND ETHICAL BASES FOR CREATING REPRESENTATIVE DATASETS TO DETECTING MANIFESTATIONS OF CYBERBULLYING IN TEXT CONTENT

The article is devoted to developing the method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content, considering ethical and legal principles. The primary focus is ensuring fair and equal representation of different demographic groups in text samples, which is critical for creating non-discriminatory and socially responsible artificial intelligence models. Emphasis is placed on compliance with key ethical principles – preventing harm, avoiding bias, and ensuring representativeness – and provisions of international law, particularly the General Data Protection Regulation. Proposed method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content, taking into account ethical principles, which includes the following stages: preliminary processing of text data, analysis of distributions according to ethical aspects (age, gender, religion etc.), and representative adjustment through multi-criteria optimization. Machine learning models are trained on prepared balanced samples using appropriate reference datasets to classify text samples according to ethical criteria. The comparison is based on official demographic data for Ukraine, which ensures the reliability of the assessment of deviations.

As a result of applying the developed method, a representative sample was created with a deviation of the proportions of ethical groups from the target values within 0.00-0.04%. The statistical metrics obtained confirmed the effectiveness of the selected models and demonstrated a high degree of compliance with the ethical responsibility requirements of the results. The analysis showed that the initial datasets contained imbalances, which were successfully eliminated through multi-criteria optimization and data augmentation. The developed approach can be integrated into preparing training samples for ethically oriented artificial intelligence systems that perform automated detection of cyberbullying manifestations in text content, reducing the risks of reproducing social biases and increasing trust in algorithmic decisions.

Keywords: cyberbullying, ethical aspects, legal basis, data representativeness, text content, dataset, discrimination, artificial intelligence, multi-criteria optimization, machine learning.

Олена СОБКО
Хмельницький національний університет
Арчіл ЧОЧІА
Талліннський технічний університет

ПРАВОВІ ТА ЕТИЧНІ ЗАСАДИ ПОБУДОВИ РЕПРЕЗЕНТАТИВНИХ ДАТАСЕТІВ ДЛЯ ВИЯВЛЕННЯ ПРОЯВІВ КІБЕРБУЛІНГУ У ТЕКСТОВОМУ КОНТЕНТІ

Статтю присвячено розробці метод формування репрезентативних датасетів текстових даних для виявлення проявів кібербулінгу у текстовому контенті з урахуванням етичних і правових засад. Основна увага зосереджена на забезпеченні справедливого та рівного представництва різних демографічних груп у текстових вибірках, що є критично важливим для створення недискримінаційних та соціально відповідальних моделей штучного інтелекту. Акцент зроблено на дотриманні ключових принципів етики – недопущенні шкоди, уникненні упередженості та забезпеченні репрезентативності – а також положень міжнародного законодавства, зокрема Загального регламенту про захист даних.

Запропоновано метод формування репрезентативних датасетів текстових даних для виявлення проявів кібербулінгу у текстовому контенті з урахуванням етичних засад, що передбачає такі етапи як попередня обробка текстових даних, аналіз розподілів за етичними аспектами (вік, гендер, релігія) та репрезентативне коригування шляхом багатокритеріальної оптимізації. Для класифікації текстових зразків за етичними ознаками використано навчання моделей машинного навчання на підготовлених збалансованих вибірках із використанням відповідних еталонних датасетів. Порівняння здійснюється на основі офіційних демографічних даних України, що забезпечує достовірність оцінки відхилень.

У результаті застосування розробленого методу сформовано репрезентативну вибірку з відхиленням пропорцій етичних груп від цільових значень у межах 0,00–0,04%. Отримані статистичні метрики підтвердили ефективність обраних моделей і продемонстрували високу відповідність результатів вимогам етичної відповідальності. Аналіз показав, що вихідні датасети містили дисбаланси, які успішно усунуто шляхом застосування багатокритеріальної оптимізації та аугментації даних. Розроблений підхід може бути інтегрований у процеси підготовки навчальних вибірок для етично орієнтованих систем штучного інтелекту, які здійснюють автоматизоване виявлення проявів кібербулінгу у текстовому контенті, знижуючи ризики відтворення соціальних упереджень і підвищуючи довіру до алгоритмічних рішень.

Keywords: кібербулінг, етичні аспекти, правові засади, репрезентативність даних, текстовий контент, датасет, дискримінація, штучний інтелект, багатокритеріальна оптимізація, машинне навчання.

Received / Стаття надійшла до редакції 02.07.2025
Accepted / Прийнята до друку 22.08.2025

Introduction

Today, artificial intelligence systems used for natural language processing are rapidly developing. Despite the complexity of modern models' architectures, the effectiveness and fairness of their decisions largely depend on

the data they were trained on [1]. This is especially true for tasks involving recognizing socially sensitive patterns, such as hostile language, aggression, or hidden forms of discrimination, as well as sensitive tasks such as detecting cyberbullying in text content [2].

From a legal perspective, building datasets based on text content containing manifestations of cyberbullying requires careful compliance with national and international legislation. First and foremost, this concerns the protection of personal data, as provided for, in particular, by the EU General Data Protection Regulation (GDPR) [3] and relevant Ukrainian legislation, such as the Law of Ukraine "On the Protection of Personal Data" [4]. Processing texts that may contain personalised or sensitive information requires a legal basis and transparency guarantees. In addition, applying machine learning methods to such data raises the issue of copyright, as some of the texts may have been created by third parties and be subject to legal protection that restricts their reuse, even for scientific purposes.

The issue of the legality of processing such data is directly related to the social necessity of developing tools to detect cyberbullying. The relevance of detecting cyberbullying is due to its significant impact on society, in particular on the psychological health of young people [5]. Studies show that a significant proportion of adolescents worldwide encounter insults, threats, or humiliation in the online space. Creating high-quality datasets is an important step in developing tools to help moderate content and protect user rights. Legal frameworks play a key role in ensuring that these tools are based on data collected legally and with respect for human rights. This increases trust in such technologies and promotes their effective implementation.

The ethical foundations of creating datasets, which often remain on the periphery of technical discourse, require systematic consideration. An ethical approach to text data collection should be based on three interrelated principles: preventing harm, avoiding bias, and ensuring representativeness. The principle of preventing harm involves minimizing the risks of re-traumatization or public stigmatization of users whose texts may contain or describe acts of aggression. Avoiding bias is a prerequisite for creating dataset that do not reproduce or reinforce social stereotypes, particularly those based on race, gender, religion, age, etc. [6]. Finally, the principle of representativeness ensures that the collected data adequately reflects the language community's social, demographic, and cultural diversity. Otherwise, even formally correct models become a source of unfair and inaccurate decisions [7].

The main scientific contribution of this study is an approach to creating representative datasets for detecting manifestations of cyberbullying in text content, which is based on compliance with legal norms, in particular regarding the protection of personal data and copyright, and ethical principles, in particular the FATE principle of fairness.

The structure of the article is as follows: the section "Related work" provides an overview of the current state of representativeness of text samples, as well as the fair and non-discriminatory representation of demographic groups in the context of dataset creation; the section "Method" presents the creating of representative text data datasets for detecting manifestations of cyberbullying in text content, taking into account ethical principles; the section "Datasets for research" provides a list of datasets used to study the methodology; the section "Results and discussion" presents the results of an experimental study of the effectiveness of the developed method; the section "Conclusions" systematizes the results obtained and indicates the possibilities for further use of the developed approach in the context of research related to the ethical aspects of artificial intelligence functioning.

Related work

Several recent studies have addressed the issue of representativeness of text samples and fair and non-discriminatory representation of demographic groups in creating datasets for socially sensitive tasks, such as detecting cyberbullying. Representativeness, fairness, and non-discrimination are increasingly considered key prerequisites for creating ethically responsible artificial intelligence models. Existing datasets often have significant gaps, especially in the representation of categories such as gender, age, ethnicity, or religious identity, which are important when analyzing hate speech or discriminatory statements. In addition, the complex, multidimensional nature of demographic variables complicates both their classification and consistent inclusion in datasets, pointing to the need for a systematic approach to creating representative, legally and ethically based datasets.

In article [8], the authors raise the important issue of sample representativeness in machine learning and artificial intelligence, emphasizing the need for accurate representation of population data. They stress that to ensure high-quality models, it is important to correctly form samples that are as close as possible to the real characteristics of the population. The primary strategy they propose is to use stratified samples, which reduce variability between subgroups and accurately reflect the proportions between different categories in the population.

The study's authors [9] consider biases arising from class imbalance in the data and sensitive (protected) characteristics such as race or gender. The authors propose a new method, Fair Oversampling, which combines the popular SMOTE method for dealing with data imbalance with modifications that help reduce the impact of sensitive characteristics.

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

The article [11] highlights the problem of intersectional bias in natural language processing models, namely the unrepresentative and biased representation of different groups of people in text datasets. The results showed that although existing debiasing methods (e.g., for BERT or RoBERTa) preserve the predictive accuracy of models well, their ability to reduce intersectional bias is limited. In intersectional cases, biases are significantly amplified, and on some tasks, the deviation from fairness can be 20-50% greater compared to individual demographic groups.

In article [12], the authors identify and classify bias in natural language processing. The main model they use for these tasks is based on transformers such as BERT, the standard for working with text data due to its ability to understand context. The authors explore various ways of detecting bias, including the detection of social characteristics such as gender, race, religion, and sexual orientation.

An analysis of current research shows that creating representative and unbiased text samples is one of the key areas in ethically sound artificial intelligence model development. This is especially true for tasks that have social or legal significance, such as detecting cyberbullying in open text content. Without considering ethical norms, artificial intelligence systems remain vulnerable to reproducing social injustice and discrimination.

The main goal of this work is to develop a method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content, which differs from existing ones in its focus on ethical and legal aspects, allowing existing text datasets to be analyzed for the correct representation of different demographic subgroups according to selected aspects of fairness, and to modify text datasets to make them representative in terms of ethical aspects of appearance.

To achieve the stated goal, the following tasks must be solved:

1. Identify issues related to the representativeness of existing open datasets regarding demographic characteristics such as age, gender, and religious identity.
2. Develop a method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content, considering ethical aspects, by solving a multi-criteria optimization problem with the target proportions of demographic subgroups.
3. Create a software implementation of the method based on modern machine learning tools for automated analysis, classification, and augmentation of text samples to achieve a representative sample.
4. Evaluate the proposed method based on public cyberbullying datasets, taking into account the demographic characteristics of the Ukrainian population, and determine the deviation between the existing and target distributions.
5. Evaluate the effectiveness of the representative sample obtained by analyzing the accuracy and balance of distributions in terms of ethical aspects, as well as based on the results of the classifiers.
6. Substantiate the potential of the proposed approach as a tool for increasing the ethical responsibility of artificial intelligence systems used to analyze socially sensitive text content.

Method

The article proposes a method for creating representative text data datasets to detect manifestations of cyberbullying in text content, considering ethical principles. To ensure ethical principles in creating representative text datasets, it is necessary to solve the problem of multi-criteria optimization to maintain a balance between different ethical groups in the data set. This ensures ethical responsibility when constructing representative samples for analyzing manifestations of cyberbullying. The scheme of the method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content is shown in Figure 1.

The input data for the method is a sample of text data for analysis, the target number of elements in the sample, a set of ethical aspects, which also contains classes and target proportions of classes, a set of machine learning models trained for each ethical aspect, which uses balanced samples for each ethical aspect for training.

The first step involves pre-processing a sample of text data, namely removing uninformative text fragments such as punctuation marks, numbers, and special characters [13]. In this case, it is inappropriate to remove emoticons during text pre-processing. Emoticons are important emotional indicators that can significantly change the meaning of a sentence. In many cases, emoticons can serve as a mood or attitude marker [14].

In step 2, we analyse the representativeness of the text sample with regard to ethical aspects. Each sample element is vectorized and then classified according to the relevant ethical criteria using separate machine learning models. Based on the classification results, the actual proportions of classes for each aspect are determined and compared with the target proportions. Deviations are calculated, and the number of missing or redundant elements is determined. This allows us to identify demographic subgroups that need to be balanced. The final step is to assess the sufficiency of the data for augmentation: if specific subgroups are underrepresented, they are targeted for supplementation.

The third step involves adjusting the data sample concerning ethical considerations to ensure representativeness. First, we solve the optimisation problem of removing redundant items that exceed the target proportions of classes for ethical reasons, while maintaining the sample's internal structure. Next, the requirements for augmentation are formulated to increase the number of elements of underrepresented classes by solving another optimisation problem. Augmentation is carried out without artificially distorting the sample by adding new elements that meet the target proportions and ethical criteria. The result is a balanced sample that meets the target proportions regarding ethical aspects.

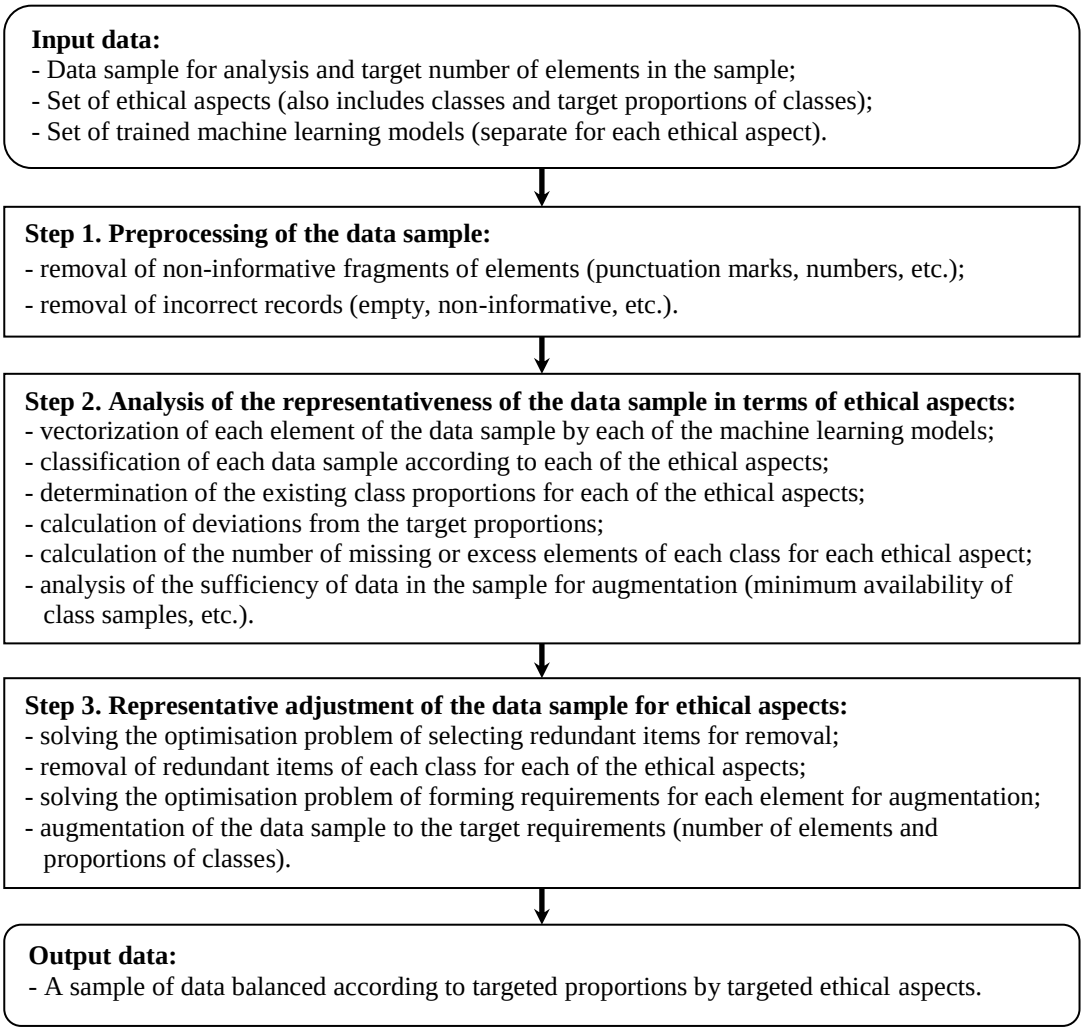


Fig. 1. Steps in the method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content

Therefore, following the steps of the method for creating representative text data datasets to detecting manifestations of cyberbullying in text content will allow us to form text samples that are non-discriminatory and reflect a representation of sample samples that is proportional to real demographic subgroups, which will affect the accuracy and transparency of machine learning models for solving various tasks.

Datasets for research

To ensure compliance with legal principles, all data sets used were obtained from open sources that provide access to data by the terms of a licence that allows their use in scientific research (in particular, Creative Commons licences or similar). In addition, when working with text data, the requirements of the GDPR were taken into account, which stipulates compliance with the principles of legality, transparency, data minimisation, and ensuring the rights of personal data subjects.

To test the proposed method, an input dataset was created based on “Cyberbullying Classification” [15] and “Cyberbully Detection Dataset” [16]. Neither datasets contain labels for the author’s gender, age, religion, or ethnicity.

The following were used to train machine learning models on ethical aspects (gender, age, religion):

– gender aspect: the “Tweet Files for Gender Guessing” dataset [17]

– religious aspect: the “CyberBullying Detection Dataset” [18];

– for age: the “TAG-it Dataset Distribution” [19].

Since the classes in the datasets are unbalanced, they are equalized by the number of samples. The final number of samples in the classes of training samples for machine learning models by ethical aspects is shown in Figure 2.

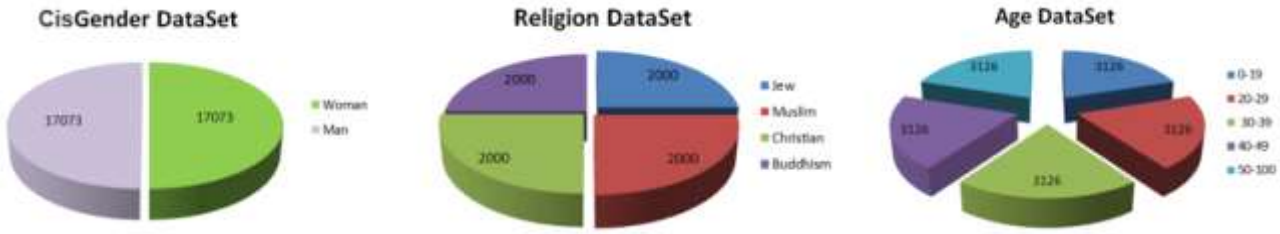


Fig. 2 Classes and number of samples in datasets for machine learning models training by ethical aspects

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

Results and discussion

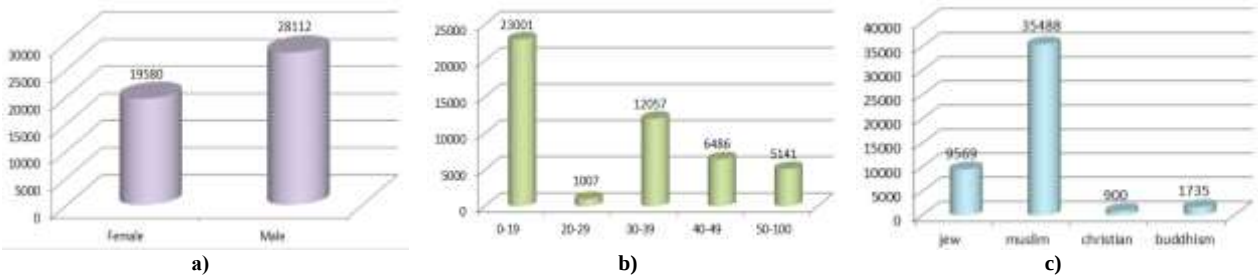
To study the effectiveness method a Python software implementation was created using the TensorFlow library to classify the cyberbullying dataset by gender, age, and religion. Demographic data from Ukraine was used to form a sample according to the target class proportions. According to the M. V. Ptukha Institute of Demography and Social Studies [20], the population of Ukraine as of July 2023 is 35,596,216 people. The population of Ukraine by age subgroups for 2023 is as follows: 0-19 years old – 6,659,068 people, 20-29 years old – 3,623,143 people, 30-39 years old – 6,022,345 people, 40-49 years old – 5,431,140 people, 50-100 years old – 13,860,520 people. Gender structure: women – 16,951,527, men – 18,644,689.

Several machine learning models have been trained to investigate the method's effectiveness. The results of static metrics (Accuracy, Precision, Recall, F1-score) for ethical aspects are shown in Table 1, and the actual balance of classes in the cyberbullying sample by gender is shown in Figure 3.

Table 1

Statistical metrics 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



a) **b)** **c)**

Fig 3. Balances of distributions of the input dataset by ethical aspects: a) gender; b) age; c) religion

Different linear resolutions were obtained for different classes: by religion, the BERT classifier showed the best result, with good resolution; by gender, the LSTM classifier was the most effective, with medium resolution; by age, the SVM classifier showed poor resolution. The diagrams in Figure 3 show that the dataset is not representative: the number of text samples on ethical issues does not correspond to the demographic proportions of the Ukrainian population, which requires balancing. In order to form a representative sample, it is necessary to augment the data by solving an optimisation problem: removing redundant elements of classes by ethical aspects and augmenting to the target proportions. Table 2 shows the percentage of samples by age in the textual data sample

and the population in age demographic subgroups, and calculates the new distribution of sample classes if only one ethical aspect – age – were considered.

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 shows the percentages of samples by gender in the textual data sample and of individuals in the population in gender demographic subgroups, and calculates the new distribution of sample classes if only one ethical aspect, gender, were taken into account.

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%

The deviations of the distributions of the transformed dataset samples by age and gender from the ideal representative distribution were as follows: minimum 0.01%, maximum 0.04%, average 0.02%; by gender: minimum 0.03%, maximum 0.03%, average 0.03%.

However, the optimisation task of creating a representative sample of text data is multi-criteria, with criteria for age and gender aspects, and is aimed at minimising deviations between the current and target class proportions, taking into account restrictions on the number of samples and the possibility of generating new data. As a result of solving the problem by augmentation, a representative sample was obtained, the class balance of which is shown in Table 4.

Table 4

Distribution of samples in the creating 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 deviations of the distributions of the transformed dataset samples by age and gender ethical aspects from the ideal representative distribution were: minimum 0.00%, maximum 0.04%, average 0.02%.

As a result of applying the method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content, a non-discriminatory text sample was creating, which proportionally reflects the demographic subgroups of the population of Ukraine.

Conclusions

The article highlights the importance of implementing legal and ethical principles for creating representative datasets to detecting manifestations of cyberbullying in text content, in particular, ensuring fair and equal representation of different groups in text samples. It emphasises that the effectiveness and social legitimacy of artificial intelligence systems depend on compliance with the principles of representativeness, non-discrimination, and respect for human rights.

The current state of legal and ethical principles implementation in creation datasets for detecting cyberbullying is analysed. It has been established that most of the available datasets do not meet the requirements of ethical balance, in particular due to insufficient representation of demographic subgroups and the lack of ethical validation procedures. The study proposes a method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content that considers legal restrictions (in particular, the provisions of the GDPR on the protection of personal data) and ethical aspects, in particular, the avoidance of bias based on age, gender, and religious criteria.

To study the method's effectiveness, software implementations of machine learning models were created, which were trained on ethically labelled data. The appropriate models (SVM, LSTM, BERT) were selected considering the optimal accuracy for each ethical aspect. The current demographic indicators of the Ukrainian population were selected as reference criteria for representativeness, which made it possible to reasonably assess the deviation between the existing and target data distributions.

As a result of multi-criteria optimization, a dataset balanced regarding age and gender aspects was created. The achieved indicators – minimum deviation of 0.00%, maximum deviation of 0.04%, and average deviation of 0.02% – demonstrate the developed approach's effectiveness in ensuring legal and ethical requirements.

The proposed method contributes to the formation of artificial intelligence systems that comply with the principles of ethical responsibility. Further plans to improve the method for creating of representative text data datasets for detecting manifestations of cyberbullying in text content include not only the creating of a non-discriminatory sample in terms of the number of samples, but also the search for and removal of text samples containing biased attitudes towards representatives of various demographic subgroups.

References

1. Memarian B., Doleck T. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. Computers and Education: Artificial Intelligence. 2023. P. 100152. URL: https://doi.org/10.1016/j.caeai.2023.100152 (date access: 16.06.2025).
2. Cyberbullying: Research Challenges and Opportunities / S. Nizam et al. 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 9–10 February 2024. 2024. URL: https://doi.org/10.1109/ic2pct60090.2024.10486805 (date access: 16.06.2025).
3. General Data Protection Regulation. URL: https://gdpr-info.eu/ (date access: 16.06.2025).
4. Zakon Ukrainy "Pro zakhyst personalnykh danykh". URL: https://zakon.rada.gov.ua/laws/show/2297-17 (date access: 16.06.2025).
5. Yengejeh A. A., Combating Cyberbullying on Social Media: A Machine Learning Approach with Text Analysis on Twitter. Data Science and Data Mining, 15. 2024. URL: https://core.ac.uk/download/pdf/599808315.pdf (date access: 16.06.2025).
6. Chen H., Ji Y., Evans D. Addressing Both Statistical and Causal Gender Fairness in NLP Models. Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico. Stroudsburg, PA, USA, 2024. URL: https://doi.org/10.18653/v1/2024.findings-naacl.38 (date access: 16.06.2025).
7. Elazar Y., Goldberg Y. Adversarial Removal of Demographic Attributes from Text Data. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/d18-1002 (date access: 16.06.2025).
8. Clemmensen L. H., Kjærsgaard R. D., Data Representativity for Machine Learning and AI Systems, arXiv preprint arXiv:2203.04706 (2022). URL: https://doi.org/10.48550/arXiv.2203.04706 (date access: 16.06.2025).
9. Dablain D., Krawczyk B., Chawla N. Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. Discover Data. 2024. Vol. 2, no. 1. URL: https://doi.org/10.1007/s44248-024-00007-1 (date access: 16.06.2025).
10. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias / R. K. E. Bellamy et al. IBM Journal of Research and Development. 2019. Vol. 63, no. 4/5. P. 4:1–4:15. URL: https://doi.org/10.1147/jrd.2019.2942287 (date access: 16.06.2025).
11. Benchmarking Intersectional Biases in NLP / J. Lalor et al. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States. Stroudsburg, PA, USA, 2022. URL: https://doi.org/10.18653/v1/2022.naacl-main.263 (date access: 16.06.2025).
12. Evans A. S., Moniz H., Coheur L., A Study on Bias Detection and Classification in Natural Language Processing, arXiv preprint arXiv:2407.01595 (2024). URL: https://doi.org/10.48550/arXiv.2407.01595 (date access: 16.06.2025).
13. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network / O. Zalutska, M. Molchanova, O. Sobko, O. Mazurets, O. Pasichnyk, O. Barmak, I. Krak. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 344–356. URL: https://ceur-ws.org/Vol-3387/paper26.pdf (date access: 16.06.2025).
14. EMFSA: Emoji-based multifeature fusion sentiment analysis / H. Tang et al. PLOS ONE. 2024. Vol. 19, no. 9. P. e0310715. URL: https://doi.org/10.1371/journal.pone.0310715 (date access: 16.06.2025).
15. Cyberbullying Classification. Kaggle.com. 2021. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download (date access: 16.06.2025).
16. CyberBullying Detection Dataset. Kaggle.com. 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification (date access: 16.06.2025).
17. Tweet Files for Gender Guessing. Kaggle. URL: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing (date access: 16.06.2025).
18. CyberBullying Detection Dataset. Kaggle.com. 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification (date access: 16.06.2025).
19. TAG-it Dataset Distribution. Live European Language Grid. URL: https://live.european-language-grid.eu/catalogue/corpus/8112/download/ (date access: 16.06.2025).
20. Natsionalni demohrafichni prohnozy. Idss.org.ua. URL: https://idss.org.ua/forecasts/nation_pop_proj (date access: 16.06.2025).

Olena Sobko **Олена Собко**	Postgraduate student of the Department of Computer Science, Khmelnytskyi National University https://orcid.org/0000-0001-5371-5788 e-mail: olenasobko.ua@gmail.com	Аспірантка кафедри комп'ютерних наук, Хмельницький національний університет
Archil Chochia **Арчіл Чочіа**	Ph.D., Senior Researcher, TalTech Law School, Tallinn University of Technology, Estonia. e-mail: archil.chochia@taltech.ee https://orcid.org/0000-0003-4821-297X	Доктор філософії, старший науковий співробітник, Юридична школа TalTech, Талліннський технологічний університет, Естонія.