

National Research University Higher School of Economics

Faculty of Computer Science

Programme: Data Science and Business

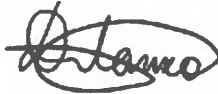
Analytics

BACHELOR'S THESIS

(Research Project)

**Group Fairness for Multiple Group Scenario in
Text Classification**

**Prepared by the student of Group 201, Year 4 (year of study),
Daria Andreevna Lapko**



Thesis Supervisor:

**Candidate of Sciences, First Deputy Dean, Tamara Vasilievna
Voznesenskaya**

Co-supervisor:

**Junior Researcher at ANO "Institute of Artificial Intelligence",
Gleb Yurievich Kuzmin**



Moscow

2024

Abstract

This thesis investigates the problem of intersectional debiasing in text classification models, aiming to mitigate biases that arise from the intersection of multiple demographic attributes like gender, race, age, and country. Existing debiasing techniques often focus on single attributes in isolation, overlooking important sources of unfairness that manifest at the intersections of these attributes. To address this gap, the thesis proposes using a joint attribute that encodes the combinations of protected attributes into a single variable. This allows directly optimizing for fairness across intersectional subgroups rather than just individual attributes. Three debiasing methods are evaluated: Least-squares Concept Erasure (LEACE), Adversarial training (Adv), and Balanced Training with Equal Opportunity (BTEO). These techniques are applied to the Multilingual Twitter Corpus (MTC) dataset, which contains hate speech annotations along with inferred author demographics. The thesis tests three key hypotheses: 1) Debiasing on single attributes is insufficient to substantially improve fairness on the joint attribute, 2) Debiasing on a single attribute can improve fairness on the joint attribute, and 3) There exist correlations between biases in different attributes that can be leveraged for cross-attribute debiasing. Experiments are conducted using the FairLib framework, with accuracy and fairness metrics like TPR-GAP and Distance to Optimum (DTO) evaluated across different debiasing methods and attributes. The results provide insights into the effectiveness of intersectional debiasing and the trade-offs between fairness and performance. The thesis concludes with a discussion of future research directions to further advance the state-of-the-art in fair and inclusive text classification.

Key words: fairness, intersectional debiasing, protected attributes, joint attribute, text classification, natural language processing

Project repository: https://github.com/DLapo4ka/fairness_thesis

Table of Contents

Abstract	2
1.1 Background	4
1.2 Problem Statement	4
1.3 Research Goals and Objectives	5
2. Literature Review	6
2.1 Intersectional Debiasing	6
2.2 Conclusion	7
3. Methods	9
3.1 Fairness Library	9
3.2 Debiasing methods	10
3.2.1 Least-squares Concept Erasure	10
3.2.2 Adversarial Training	11
3.2.3 Balanced Training with Equal Opportunity	11
3.3 Hypotheses	12
3.4 Conclusion	13
4. Experimental Setup	14
4.1 Dataset	14
4.2 Model	15
4.3 Metrics	15
4.3.1 Equal Opportunity	16
4.3.2 Distance to the Optimum	17
4.4 Hyperparameter Optimization	18
4.5 Plan of Experiments	19
4.6 Computational Resources	20
4.7 Conclusion	20
5. Results	22
5.1 Base Model	22
5.2 Debiasing for Joint Attribute	22
5.3 Debiasing for Single Attributes Applied to Joint Attribute	26
5.4 Debiasing for One Single Attribute Applied to Another	28
5.5 Conclusion	31
6. Further Work	32
7. Conclusion	33

1. Introduction

1.1 Background

Text classification is a fundamental task in natural language processing (NLP) that involves categorizing text into predefined classes based on its content. This technique has numerous real-world applications, including spam detection, sentiment analysis, and topic modeling [1]. Text classification models are trained on labeled data, where the input text is assigned a label based on its content. The model then learns to predict the label of unseen text based on the patterns it has learned during training.

However, recent studies have shown that text classification models can exhibit bias, leading to unfair treatment of certain groups. For example, a text classification model trained on a dataset with a disproportionate number of male authors may be biased towards male authors, leading to unfair treatment of female authors. This bias can have serious consequences, particularly in high-stakes applications such as criminal justice, healthcare, and education. Not only is a lack of fairness undesirable from a moral point of view, but it is also prohibited by laws that regulate discrimination (i.e., the US Federal Equal Employment Opportunity collection of laws, the UN General Assembly Convention on the Elimination of All Forms of Racial Discrimination) [2]. In this regard, the fairness of automated decision-making systems is already regulated, and this must be considered when implementing any classification algorithms. This issue is particularly concerning in scenarios involving multiple groups, where the bias is present in several intersecting groups at the same time (for example, if a non-white woman is treated unfairly).

1.2 Problem Statement

Since bias has been found in textual classification models, scientists in the field have been able to develop ways to identify and partially remove it from the model. However, most of the papers focused on fairness with respect to a single group, whether it is race, gender, or else. Cases where several protected groups are subject to bias at the same time have not yet been studied in depth.

This research seeks to address a critical gap in the literature by investigating methods to ensure group fairness in text classification for multiple groups rather than just a single group. By examining the nuances of algorithmic fairness in text classification, this study aims to contribute novel insights and propose methodologies to enhance the equity and transparency of machine learning models. The formal statement of the research question is: What are the effective methods to ensure group fairness in text classification for multiple group scenarios, and how do they impact the bias and accuracy of the text classification models?

1.3 Research Goals and Objectives

The goal of this project is to conduct a comprehensive study on the effectiveness of various debiasing techniques for achieving group fairness in text classification models when dealing with multiple demographic groups. The main objectives of this thesis are to:

1. Investigate the current landscape of fairness in text classification, identifying existing challenges and limitations.
2. Select appropriate methodologies for debiasing text classification models, which can be applied to scenarios involving multiple demographic groups.
3. Experimentally validate chosen techniques using real-world datasets to assess their effectiveness and scalability.
4. Compare the results obtained between the methodologies with relevant benchmarks and identify the strengths and weaknesses of each debiasing technique.
5. Draw conclusions, provide valuable insights based on the conducted analysis and present ideas for further improvements.

2. Literature Review

2.1 Intersectional Debiasing

Intersectional debiasing is a critical area of research within the field of natural language processing, focusing on mitigating biases that arise from the complex interaction of multiple demographic variables. It is a widely discussed topic that has gained popularity only recently. One of the main papers in this field, by Subramanian et al., delves into assessing debiasing techniques for intersectional biases, highlighting the pervasive nature of bias in NLP models and the necessity for automatic debiasing methods [3]. The study introduces the concept of ‘gerrymandering’ groups, which include both intersectional and single protected attributes. This phenomenon occurs when performing debiasing towards one attribute worsens the fairness of another, so a model must consider such groups among others too to be truly fair. Moreover, the study introduces a form of bias-constrained model that is novel to NLP, along with an extension of the iterative nullspace projection technique (INLP) capable of handling multiple identities. This study evaluates the effectiveness of various debiasing techniques in mitigating biases arising from the complex interplay of demographic attributes, underscoring the need for sophisticated debiasing strategies to address intersectional biases effectively [3].

There are also some other related studies. Tan & Celis assess social and intersectional biases in contextualized word representations, shedding light on the challenges and implications of biases in machine learning models. Importantly, the authors observe that bias effects for the intersectional groups are aggravated beyond the individual minorities they are composed of [4]. Another paper by Kang et al. introduces InfoFair, an information-theoretic approach for enforcing intersectional fairness in machine learning models. The authors focus on ensuring statistical parity across multiple sensitive attributes simultaneously, and they propose an information-theoretic framework that quantifies intersectional fairness based on mutual information between model predictions and sensitive attributes [5]. Moreover, the paper by Foulds et al. introduces the concept of ‘differential fairness’ that requires the

performance of a machine learning model to be equitable across all intersectional subgroups defined by the protected attributes. They show that their fairness criteria behave sensibly for any subset of sensitive attributes and provide theoretical guarantees around economic, privacy, and generalization properties [6]. Lastly, the paper authored by Gohar and Cheng highlights the growing concerns over the fairness implications in decision-critical domains such as criminal sentencing and bank loans, which have arisen as a result of the extensive utilization of machine learning algorithms. This paper presents a taxonomy for considering fairness in relation to many intersecting attributes and proposes techniques to address biases that affect multiple sensitive attributes simultaneously [7].

Separate attention should be paid to the establishment of a joint attribute – a combination of several protected attributes that should be considered together when evaluating fairness. The paper by Han, Baldwin, et al. suggests setting intersectional classes, which would represent multiple demographic attributes, but leaves this point for further studies [8]. This recent paper by Wang et al. implements this idea, taking the Cartesian product of the values for each protected attribute to generate all possible intersectional groups [9]. For example, taking the gender (male, female) and race (white, non-white) attributes would yield four values in the joint attribute: male & white, male & non-white, female & white, female & non-white. Using the joint attribute is necessary for intersectional debiasing tasks, as it allows to look at disparities between all combinations of protected attributes, not just each attribute independently. Referring to the goal of the current project, the joint attribute with the definition above is used to assess fairness and draw valuable conclusions about the effectiveness of various techniques for removing model bias. This is described in more detail in Section 4.1.

2.2 Conclusion

Based on the literature analysis conducted, approaches to addressing the posed tasks were explored. It has become evident that intersectional debiasing is gaining increasing attention in the machine learning community and certainly requires further

study, which this project can contribute to. Some important conclusions were drawn to be used in the following research. Several studies agreed that extended exploration is needed in the field of intersectional debiasing, while one mentioned that bias in intersectional attributes can essentially be higher than in the single protected attributes it consists of. Some papers outlined that maintaining the limited distances between intersectional groups allows the model to maintain high fairness. Finally, the need for creating a joint attribute that encodes the intersectionality of multiple protected attributes into a single variable was pointed out.

3. Methods

This section explores the use of methods to deal with intersectional debiasing in the current project. The hypotheses that should be checked are also presented in this section.

3.1 Fairness Library

The state-of-the-art, unique Python library FairLib, presented in the paper by Han et al., is a unified framework for assessing and improving fairness in machine learning models. It provides a systematic infrastructure for quickly accessing and evaluating the fairness of models. It includes several fairness metrics, such as demographic parity and equal opportunity, to evaluate the fairness of machine learning models. The framework also presents fairness constraints that limit the amount of bias that can be learned by the model, and fairness regularization techniques are included to encourage the model to learn fair representations of the data. The paper also includes a case study on using FairLib to improve the fairness of a text classification model. The study shows that the framework can effectively improve the fairness of the model while maintaining its performance, which makes it a good starting point for experiments in this project [10].

Even though the FairLib library is quite new and fresh, successful attempts have already been made to refine the library so that it can be used for an even bigger variety of experiments. The work of Kuzmin et al. allows for the use of FairLib with a greater number of popular datasets with demographic data than the initial one, as data loaders are already implemented. The paper itself also explores the impact of fairness considerations on the reliability of debiased machine learning models and assesses uncertainty in debiased models. In collaboration with the authors of this paper, this project has been given the opportunity to receive the newest version of the library that is suitable for the research [11].

3.2 Debiasing methods

As mentioned in the paper by Han et al., various techniques for mitigating bias exist, including debiasing during data preprocessing, model training, and post-processing [10]. To evaluate the dependence of the algorithm type on the efficiency of debiasing and predictive power, each of the three debiasing techniques considered in this thesis belongs to these different types.

3.2.1 Least-squares Concept Erasure

The least-squares concept erasure (LEACE) debiasing technique is a mathematical method that eliminates specific features from a representation. It can be used to prohibit a classifier from considering factors such as gender or race or to remove a concept and analyze how it affects the behavior of a model. The method is introduced in a paper by Belroze et al. and is demonstrated to effectively prevent all linear classifiers from identifying a concept while reducing the amount of change made to the representation. From the inside, LEACE is a unique affine transformation that produces guarded features and minimizes the average squared distance from the original features with regard to all norms that are caused by inner products. The debiasing algorithm is used with the last layer of the model, the classification layer, or sequentially applied to the intermediate representations at each layer of a network by means of ‘concept scrubbing’ [12].

The LEACE method has several strong advantages that were decisive in choosing it over competing techniques. Firstly, it is computationally efficient and fast because it does not require gradient-based optimization and a large number of iterations to take effect. Secondly, the impact of removing specified features on the overall representation is minimized, which means that accuracy is not hurt when one tries to improve fairness. Finally, this method is comfortably scalable, which allows one to try it with non-linear models and assess its effectiveness when used for intersectional debiasing. Since this research is aimed at using the LEACE method at the post-training step, it will only be applied to the last layer of the model.

3.2.2 Adversarial Training

The adversarial training approach (Adv) is a widely used approach in machine learning and, specifically, natural language processing, for improving model robustness [13]. As a method for debiasing text representations, it is discussed in the studies by Li et al. [14] and Elazar & Goldberg [15]. The technique involves an encoder that learns hidden representations from the input text, a classification head that predicts the target task from the hidden representations, and an adversarial discriminator that tries to predict the sensitive attribute from the hidden representations. The optimization objective is to minimize the task loss while maximizing the adversarial loss, encouraging the encoder to learn representations that are useful for the task but uninformative about the sensitive attribute.

As it is clear from the name, this method is done during model training, which makes it efficient in eliminating bias from virtually every layer of the model. In other words, this method is expected to be the most efficient from the perspective of removing bias. Additionally, this method is expected to act robustly and generalize to unseen data, as often models trained with Adv do so because of learning transferable features [16]. Finally, the regularization present in this technique allows for reduced overfitting and improved generalization.

3.2.3 Balanced Training with Equal Opportunity

The balanced training with equal opportunity method (BTEO) is discussed in the study by Han, Baldwin, et al. The method focuses on equalizing opportunities for positive outcomes between privileged and unprivileged groups by balancing the true positive rate metric. This ensures that the model does not discriminate in its ability to correctly identify positive instances. As mentioned in the paper, BTEO achieves both higher accuracy and reduced bias, resulting in competitive outcomes compared to Adv and competitive trade-offs between performance and fairness compared to other debiasing approaches [8].

As in the paper by Kuzmin et al., BTEO is used as a pre-processing method in the current project. That is, resampling and reweighing are applied to the training set

to ensure it is balanced across protected groups [11]. Compared to other pre-processing techniques, such as dimensionality and noise reduction, BTEO allows for more complex dataset and instance transformations, therefore targeting not only achieving higher model accuracy but also improving fairness.

3.3 Hypotheses

There are a few hypotheses that are expected to be verified to form a complex analysis of applying existing debiasing methods to the intersectional fairness tasks.

To start with, it is required to check if debiasing approaches for single attributes have some positive contribution to solving issue of biasedness in intersecting attributes. Since approaches that only consider one attribute at a time may overlook important sources of unfairness that arise from these intersections, a specific approach to debiasing is likely necessary to ensure equitable fairness across all demographic subgroups. In other words, a model with a debiasing applied to joint attribute should show fairness metric to be better compared to a model without debiasing. This way, we arrive at Hypothesis 1.

Hypothesis 1: Existing debiasing techniques that focus on a single protected attribute will be insufficient to substantially improve fairness in a joint attribute.

The Hypothesis 2 focuses on verifying how well the training model with debiasing method applied on single protected attribute will aid in removing bias in the joint attribute. This statement assumes that the single attribute is included into the joint one. It is expected that doing so may help increase fairness of joint attribute, as the biases between them are expected to be correlated. The hypothesis will be true if joint attribute evaluated on model with debiasing on single attribute brings better fairness metric than the joint attribute of the basic model.

Hypothesis 2: Fairness of joint attribute will increase if it is evaluated using a model with debiasing method trained on constituting single protected attribute.

Lastly, a prediction is made about the influence of bias knowledge acquired by the model with debiasing on one protected attribute on bias reduction of another such

attribute. It is interesting to find out whether there are correlations between biases of different sensitive attributes. This is simply checked by evaluating one attribute using the model trained with debiasing method on another attribute. It is expected that such correlations may theoretically exist.

Hypothesis 3: There exist two single protected attributes that have bias correlation, so training with debiasing on one attribute will show at least some sufficient improvement in fairness if applied to the other single attribute.

3.4 Conclusion

Several methods and tools were observed in this section that may help to prepare and perform experiments smoothly and effectively. It was decided that FairLib will be used as a basis for experiments and to test the effectiveness of debiasing methods. Several debiasing techniques, such as LEACE, Adv and BTEO, were chosen, a comprehensive overview of each was provided, together with the merits and reasons for choice of the specific technique. All these methods will be further used in experiments to evaluate their accuracy and fairness on intersectional debiasing tasks. The section ends with the hypotheses outlined for further study.

4. Experimental Setup

This section explores some of the main details of the experiments, such as the information about the utilized dataset, model and metrics that are required for complete and successful experiments. It also presents a pre-processing step of choosing optimal hyperparameters, a description of the sequence of experiments to be conducted, and a part with a description of technical specifications.

4.1 Dataset

Due to the nature of the task, finding a suitable dataset is challenging. From the outset, the difficulty in evaluating fairness lies in the inability to access personal data about individuals, as this data constitutes protected attributes. In order to preserve privacy, many individuals conceal sensitive data on social media and do not share it publicly. Another issue is the dataset content and fullness, as it should contain intersecting attributes in sufficient quantity so that the size is big enough to train the model and receive meaningful results. Overall, there is a limited number of datasets that meet these requirements. Fortunately, it was possible to obtain one for this project.

The multilingual Twitter corpus (MTC) dataset is a significant resource for text classification, particularly for sentiment analysis tasks in multiple languages. The dataset is presented as a part of a paper by Huang et al. and is said to be the first multilingual hate speech corpus annotated with author attributes aiming for fairness evaluation [17]. It includes tweets in five different languages: English, Italian, Polish, Portuguese and Spanish, with over 1.6 million manually labeled tweets in total. There are 4 demographic attributes present: age, gender, ethnicity (race) and country, with intersections present between most of them. The first three attributes were inferred from images of Twitter users' profiles using a reliable computer vision API, whereas the country attribute was obtained from the coordinates or location indicated in the profiles. The attributes were then binarized, with gender simply splitting into male and female, age being either smaller or larger than the median value, race showing either a white or non-white person, and country being either US or non-US

(accounting for the fact that most of the collected English-based tweets in the dataset are made by US residents). The dataset was retrieved directly from the paper author for educational purposes.

Although the presented dataset has a variety of languages that debiasing can be tested on, a significant number of tweets are still in English, and in combination with the language of this thesis, it was decided to use only the English part of the dataset. A discussion of cross-lingual debiasing and its transferability from one language to another is left for future research. It should also be noted that the joint attribute does not consist of all demographic attributes present in the dataset. Unfortunately, the ethnicity attribute is not included in the joint one, as some intersections are not present in the former that are required for the correct behavior of the latter, which may lead to unexplainable results. Following this, the use of the ethnicity attribute in the experiments is also abandoned, as there is no practical utility in receiving results for another unrelated single attribute, which requires additional experiments and hence time costs.

4.2 Model

The current research does not require a complex language model to complete since its goal is to evaluate fairness before and after debiasing, which is done with several methods. Therefore, any accessible model with good overall performance on prediction tasks and suitability for text classification will suffice. The pre-trained BERT model ('bert-base-cased') [18] was chosen as the main model to conduct experiments on. It is simple to start with, though it is still reliable and has excellent potential for high-accuracy predictions. Moreover, it is popular among NLP fairness researchers [11], allowing to compare the results of this project to other papers if needed. The parameters of BERT are fine-tuned on the training set; this process is described in detail in Section 4.4.

4.3 Metrics

The models in this thesis are evaluated using three metrics applicable to classification use cases. Most of these rely on the confusion matrix, which provides a

comprehensive summary of the model's predictions by comparing them to the actual truth values. There are four different combinations of predicted and actual values: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

The first metric is accuracy, a universal performance indicator for machine learning models. *Equation (4.3.1)* shows the formula for its calculation. Note that the value is multiplied by 100 to enhance interpretability.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100 \quad (4.3.1)$$

The second is a gap-based metric for estimating fairness, and the third is a trade-off metric between performance and fairness. These will be explained in detail below.

4.3.1 Equal Opportunity

Fairness metrics are essential for measuring and mitigating bias in AI systems, ensuring that AI models treat all individuals or groups equally, regardless of their underlying characteristics. The equal opportunity fairness metric calculates the true positive rate gap (TPR-GAP), which is used to measure differences in true positive rate (TPR), or recall, between protected attributes [8][12][20]. This metric is simple to obtain yet widely used among NLP fairness researchers, allowing to compare the results that will be received in this project with others reported in scientific literature. The TPR formula can be found in *Equation (4.3.1.1)* [11].

$$TPR = \frac{TP}{TP+FN} \quad (4.3.1.1)$$

Equation (4.3.1.1) is also used as a starting point for the calculation of TPR-GAP. To proceed, one should group-wise aggregate the TPR-GAP according to *Equation (4.3.1.2)*, where C – number of classes, G – number of protected groups, c – class instance, g – group instance, \overline{TPR}_c – averaged $TPR_{c,g}$ across groups.

$$\beta_c = \frac{1}{G} \sum_g |TPR_{c,g} - \overline{TPR}_c| \quad (4.3.1.2)$$

Next, the received β_c is aggregated class-wise, as shown in *Equation (4.3.1.3)*.

$$\delta = \sqrt{\frac{1}{c} \sum_c \beta_c^2} \quad (4.3.1.3)$$

Initially, TPR-GAP values range from 0 to 1. To make its comparison to other metrics easier, the δ is multiplied by 100; see *Equation (4.3.1.4)*.

$$TPR-GAP = 100 \cdot \delta \quad (4.3.1.4)$$

As TPR-GAP approaches 0 (i.e., there is no difference in TPR between the attributes), the model becomes fairer, and vice versa. This way, fairness can be ensured in the model.

4.3.2 Distance to the Optimum

The problem with evaluating debiasing methods is that typically none of them simultaneously achieve the best fairness and performance on a model, which means it is impossible to choose the best method. The last metric chosen allows us to deal with this issue and quantify the trade-off between the two previous metrics. Distance to the Optimum (DTO) is calculated as the normalized Euclidean distance from a model's performance and fairness metrics (accuracy and TPR-GAP, in our case) to an optimal 'utopia' point, where both performance and fairness are maximized [8][20]. An exact formula for DTO used in this project can be found in *Equation (4.3.2.1)* [10]. Note that since accuracy and TPR-GAP given in *Equation (4.3.1)* and *Equation (4.3.1.1)* respectively, were multiplied by 100, they should be returned to their initial ranges for the correct calculation of DTO.

$$DTO = \sqrt{\left(1 - \frac{1}{100} Accuracy\right)^2 + \left(\frac{1}{100} TPR-GAP\right)^2} \quad (4.3.2.1)$$

The metric values range from 0 to 1, and a smaller DTO value indicates that an ordered pair $(Accuracy, TPR-GAP)$ is closer to the maximized pair $(\max(Accuracy), \max(TPR-GAP))$, which means that the trade-off converges to the best.

4.4 Hyperparameter Optimization

Hyperparameter optimization, or tuning, is performed in machine learning to find the optimal combination of hyperparameters for a specific model or method that minimizes the loss function and improves accuracy on unseen data. In this project, hyperparameter tuning is done twice. Firstly, basic BERT is optimized during training with respect to the model accuracy on the validation set. This model has 4 parameters that can be fine-tuned: training batch size, dropout probability, weight decay (L2 penalty) and learning rate. The standard grid search algorithm with a fixed grid was used; see *Table 4.4.1*. The number of epochs is selected by the early stopping parameter of the model.

Table 4.4.1. Grid of parameter values for the base model optimization. The optimal parameters that yielded the highest accuracy are highlighted in bold.

Parameter	Search Range
Batch Size	[16 , 32]
Dropout	[0, 0.1]
Weight Decay	[0, 0.01]
Learning Rate	[0.001, 0.0001, 0.00001, 0.000001]

Secondly, the optimization is done for the hyperparameter of the Adv debiasing method, namely `adv_lambda`. Since the debiasing method works differently in application to different protected attributes, the given parameter should be unique for every respective attribute. This time, the DTO metric is used to assess results and make a choice about the best model. The grid search algorithm with a fixed grid is used again; see *Table 4.4.2*. for results. The number of epochs is selected by the early stopping parameter of the model.

Table 4.4.2. Grid of `adv_lambda` parameter values for different attributes for Adv debiasing method optimization. The optimal parameters that yielded the lowest DTO are highlighted in bold.

Attribute	Search Range
age	[0.0001, 0.001, 0.01, 0.1 , 1, 10, 100]
gender	[0.0001, 0.001, 0.01, 0.1, 1, 10 , 100]
country	[0.0001, 0.001, 0.01, 0.1 , 1, 10, 100]
joint	[0.0001, 0.001, 0.01 , 0.1, 1, 10, 100]

After optimal parameters were found where required, the experiments started to be conducted. Three different seeds are used to prevent outliers and increase the robustness and stability of the model. The results of the experiments on these seeds are then averaged and presented in Section 5.

4.5 Plan of Experiments

The experiment plan is composed in accordance with the hypotheses we want to verify and the methods we would like to use. The starting point is the training of the basic model and the evaluation of its accuracy and fairness in all single protected attributes and joint attribute, then calculation of the DTO out of two metrics. This would allow to see how the model performs prior to debiasing being applied, therefore tagging it as ‘baseline’ and allowing to compare the results with those models where debiasing was performed.

After the baseline model has been evaluated, debiasing is performed using the chosen methods. Everything starts with training, which is performed for every protected attribute separately. At this point, the debiasing technique to be executed is chosen. Importantly, if one uses Adv, the `adv_lambda` parameter must not be forgotten to be indicated before training is executed. As a result, we receive a trained model with weights distributed in a way that reduces the influence of information presented in demographic attributes on the result of classification. Afterwards, the trained models can be used to evaluate the accuracy, TPR-GAP and DTO of the same attribute the training was done in application to, or of other attributes, including joint one. It is particularly important for the research to evaluate a joint attribute on the model with debiasing done with respect to the single protected attributes, a joint

attribute on the model with debiasing on the joint attribute, and a protected attribute on the model with debiasing on the other single attributes, as these are necessary to validate the hypotheses stated in Section 3.3.

4.6 Computational Resources

The experiments were conducted on a high-performance computing cluster at HSE University within the “Debiasing for NLP tasks” project. Three types of computing nodes were utilized; their characteristics are presented in *Table 4.6.1* [21]. 164.36 CPU and GPU hours were spent on conducting experiments in total. The presence of a substantial number of nodes allowed for several experiments to be conducted simultaneously, which reduced overall time spent.

Table 4.6.1. Characteristics of three types of computing nodes in the HSE University computing cluster.

Nodes of type A	
CPU	2 x Intel Xeon Gold 6152 2.1-3.7 GHz (2*22 cores)
GPU	4 x NVIDIA Tesla V100 32 GB NVLink
Nodes of type B	
CPU	2 x Intel Xeon Gold 6152 2.1-3.7 GHz (2*22 cores)
GPU	4 x NVIDIA Tesla V100 32 GB NVLink
Nodes of type C	
CPU	2 x Intel Xeon Gold 6240R 2.4-4 GHz (2*24 cores)
GPU	4 x NVIDIA Tesla V100 32 GB NVLink

4.7 Conclusion

The current section is mostly directed at pre-processing. The dataset was chosen and described, with some attributes of it denied for further use. A base model and metrics were introduced, with detailed calculations of the latter. Next, the hyperparameters of the basic model and one debiasing method were optimized, targeting the highest accuracy and lowest DTO, respectively. A plan of experiments allows to understand in detail how debiasing methods are going to be utilized to

obtain relevant results for the project. Finally, all computational resources used were reviewed.

5. Results

This section describes the experiments conducted and outlines the results. The latter are also discussed, with propositions of the possible reasons for certain behaviors of the model.

5.1 Base Model

The results of the base BERT model trained and evaluated on different attributes are presented in *Table 5.1.1*. Overall, the model shows high accuracy on the given dataset, as well as a good level of fairness for each attribute from the very start. Considering that TPR-GAP is measured between 0 and 100 in this project, the values show that all considered attributes are biased by less than 10% and that the trained BERT is fair enough initially. Still, there is always room for improvement, and it is expected that the chosen debiasing methods will be able to improve fairness even more.

Regarding the fairness of each individual attribute, the joint attribute turns out to be the most discriminatory group, as predicted. Gender, on the other hand, appears to be the least biased. Interestingly, this is not obvious from a real-world perspective, as gender disparities are generally discussed more than age and country. DTO follows the pattern of TPR-GAP, with all attributes put in the same order and joint attribute performing the worst in terms of accuracy-fairness trade-off.

Table 5.1.1. Results for single and joint attributes evaluated on the model without debiasing.

	age	gender	country	joint
Average test accuracy	89.79	89.79	89.79	89.79
Average test TPR-GAP	4.73	2.61	3.51	5.45
Average test DTO	0.1125	0.1054	0.108	0.1158

5.2 Debiasing for Joint Attribute

Now, turn to the results of training the base model with the use of LEACE, Adv and BTEO debiasing techniques applied to the joint attribute and evaluating the model on joint and single attributes afterwards. The results for LEACE are presented

in *Table 5.2.1*. Comparing these to the results of the base model, a good level of accuracy is observed, which increased negligibly. TPR-GAP, unfortunately, has not changed much for the joint attribute as well as DTO, meaning that the LEACE method was not able to grasp the hidden representations and prevent the classifier from using those when making predictions. The values also did not change for the other attributes, evaluated using the debiasing model with LEACE trained on the joint attribute. Overall, LEACE showed itself as a powerless method for intersectional debiasing.

Table 5.2.1. Results for joint and single attributes evaluated on the model with LEACE debiasing applied to the joint attribute.

	age	gender	country	joint
Average test accuracy	89.82	89.82	89.82	89.82
Average test TPR-GAP	4.81	2.58	3.60	5.62
Average test DTO	0.1126	0.1051	0.108	0.1163

The results for another debiasing method, Adv, applied to the joint attribute are shown in *Table 5.2.2*. Compared to the base model, Adv showed disappointing results. The joint attribute fairness decreased, although it is still within the 10% bias range. Hence, the method is not effective for intersectional debiasing, at least in this setting. The accuracy, at the same time, dropped too due to the impact of Adv on the BERT training structure. The adversarial component introduced by Adv constrained the underlying model so as not to learn information about the sensitive attributes, which prevented it from receiving effective training. It may also mean that sometimes sensitive attributes do contain information that is vital to making predictions, but in the Adv setting, these connections are suppressed. A high TPR-GAP in terms of gender and age is a clear indicator of gerrymandering. In other words, trying to remove bias in the joint attribute led to an increase in the bias of several single protected attributes. Besides the aforementioned problems of applied Adv debiasing, a positive outcome is also present in the country attribute. It is better in fairness here than in the base model, which may be a result of the correlation in bias between the joint attribute and country attribute. Due to the fact that the change in accuracy is

larger than the change in TPR-GAP, the DTO is substantially higher than in the initial BERT. From this perspective, the trade-off between performance and fairness does not seem to be the best.

Table 5.2.2. Results for joint and single attributes evaluated on the model with Adv debiasing applied to the joint attribute.

	age	gender	country	joint
Average test accuracy	67.99	67.99	67.99	67.99
Average test TPR-GAP	5.77	4.55	2.58	6.18
Average test DTO	0.3259	0.3234	0.3213	0.3267

The third debiasing method, BTEO, is also tested when used to reduce bias in the joint attribute. The results are presented in *Table 5.2.3*. Compared to the base model, the joint attribute fairness is better, which is a good point. However, there are some concerns that this result is reliable. At the experiment time, the model yielded extremely varying results on different seeds, with TPR-GAP either zeroed and accuracy lowered to unpractical values or TPR-GAP increased towards the Adv value and accuracy preserved at a base model level. At this point, it is hard to say whether it is a failure in experiment or a strong method of debiasing in relation to the chosen dataset. It is possible that the intragroup recall converged to 0 and groups became almost identical in bias, but the likelihood is very small. This may also, possibly, be explained by the difference in data distribution with respect to attribute-class combinations in the validation and test sets. Notably, the strange results stated above are only present for the test set, while at the validation set, the results for accuracy and fairness are much better. As with all post-processing methods, BTEO is vulnerable to data discrepancies, which imitate the distribution of real data without strict stratification. The accuracy of the model is even lower than that of Adv, but it frequently happens if fairness soars. This could be explained by the accuracy-fairness trade-off; however, DTO shows high values again because the change in accuracy outweighs the change in fairness. That is, for such a small change in fairness, a subsequent change in accuracy is not worth it. Paying attention to the change in the other protected attributes' values, their fairness metric is also substantially lower than

in the base model. The issue with these is the same as with the joint attribute, and the same reason why it possibly happens. DTO for single attributes also suggests that the accuracy-fairness tradeoff is the worst compared to the base model and other methods used. Therefore, we can't conclude that the BTEO method is robust for dealing with intersectional debiasing, though there is some evidence of successful bias mitigation.

Table 5.2.3. Results for joint and single attributes evaluated on the model with BTEO debiasing applied to the joint attribute.

	age	gender	country	joint
Average test accuracy	61.99	61.99	61.99	61.99
Average test TPR-GAP	3.43	1.58	2.3	4.13
Average test DTO	0.3847	0.381	0.3822	0.3869

The final distribution of models according to their debiasing results on joint attribute is presented in *Figure 1*. As observed, LEACE and the base model have minimal differences in TPR-GAP, while accuracy is approximately the same. Adv owns the highest TPR-GAP, which indicates the lowest fairness and semi-good prediction performance. Adv is certainly not a choice for intersectional debiasing. BTEO presented arguable results, showing high fairness on the one hand and having a data distribution difference problem distorting the results on the other. It is not recommended to rely on this method for joint attribute debiasing.

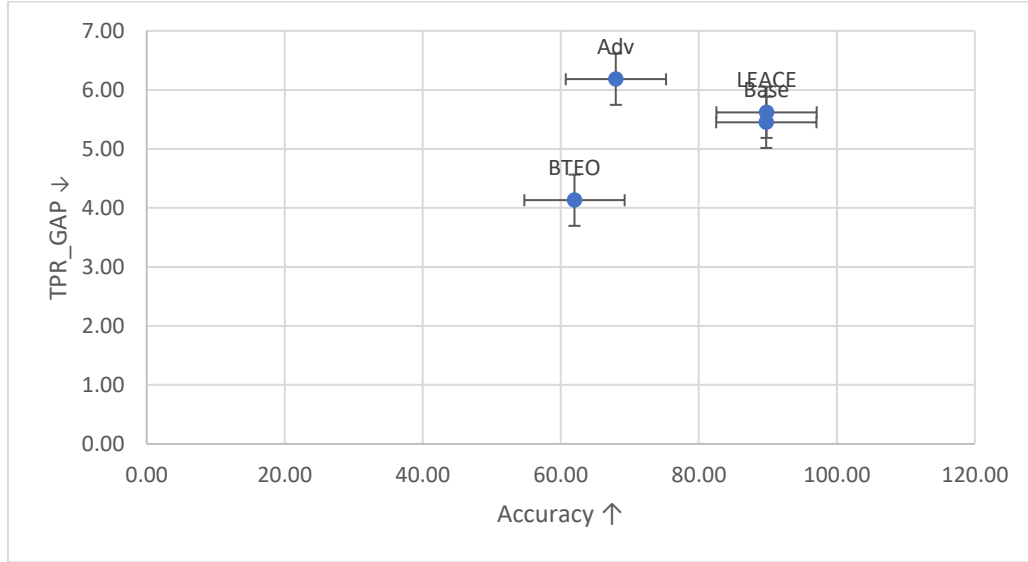


Figure 1. Joint attribute debiasing model performance.

5.3 Debiasing for Single Attributes Applied to Joint Attribute

As a next step in the analysis, it is desired to reveal whether using trained models with debiasing on single attributes can improve the fairness of joint attribute. Starting with the age variable, results are shown in *Table 5.3.1*. For the first two methods, the result is negative: fairness decreased to varying degrees in LEACE and Adv compared to the base model. This means that the knowledge about bias in age did not help predict and remove bias in the joint attribute, even though this single attribute is contained in it, and the issue of gerrymandering appears. The BTEO method, as indicated in the previous section, has some unexplained variability in results between validation and test sets, so conclusions about it should be made with care. Nevertheless, BTEO seems to catch the bias knowledge in age such that it helps to decrease joint attribute TPR_GAP sufficiently compared to the base model. This, however, happens not without losses: accuracy is low, so predictions made are often erroneous. Following this, the DTO is very high compared to the values of other methods, indicating that the usefulness of BTEO for the current task is questionable.

Table 5.3.1. Results for the joint attribute evaluated on the models with different debiasing methods applied to the age attribute.

	LEACE	Adv	BTEO
Average test accuracy	89.96	73.6	44.99
Average test TPR-GAP	6.28	5.97	3.87
Average test DTO	0.1185	0.2726	0.5517

The details about the model with gender debiasing are presented in *Table 5.3.2*. The results for LEACE and Adv are negligibly better than those for age attribute debiasing. The gerrymandering issue is also present, as fairness values have worsened compared to the base model. BTEO was able to perform at its best again, surpassing the TPR-GAP value in the base model. This result, however, is lower than when joint attribute is evaluated on an age-debiased model, meaning that there are more correlations between the biases of joint and gender attributes. The BTEO accuracy is not that bad this time, compared to the one in *Table 5.3.1*, but it is still not high enough for this method to be practically useful. DTO also demonstrates the inefficiency of the accuracy-fairness trade-off of BTEO compared to other methods.

Table 5.3.2. Results for the joint attribute evaluated on the models with different debiasing methods applied to the gender attribute.

	LEACE	Adv	BTEO
Average test accuracy	89.63	89.87	65.76
Average test TPR-GAP	5.93	5.88	4.15
Average test DTO	0.1195	0.1171	0.3503

The third protected attribute to perform debiasing on is country; the results are available in *Table 5.3.3*. The situation here is much better than for the previous attributes. The LEACE method was able to outperform the base model in TPR-GAP and DTO, whereas accuracy stayed the same. Even though the increase in the fairness value is not very large, this shows that LEACE has great potential and is capable of finding at least some correlations between attributes' biases. Turning to other methods applied to the country, Adv received worse fairness than the base model while also losing some accuracy, and BTEO had an insane increase in the fairness of

the joint attribute compared to the initial model but also lost a large amount of accuracy.

Table 5.3.3. Results for the joint attribute evaluated on the models with different debiasing methods applied to the country attribute.

	LEACE	Adv	BTEO
Average test accuracy	89.79	73.02	30.09
Average test TPR-GAP	5.08	5.83	1.1
Average test DTO	0.114	0.2776	0.6914

The analysis of the use of a model with a debiasing algorithm applied to some protected attribute in order to promote fairness in the joint one yielded a promising result. To be precise, when trained on the country attribute, LEACE was able to deal with a significant amount of bias in the joint attribute. The method's ability to find hidden correlations between biases could possibly be enhanced and scaled. These results also mean that the country attribute in the MTC dataset has a significant correlation with the joint attribute in the sense of bias, and this information can be used to design more effective debiasing methods specifically for this dataset. Talking about BTEO, even though the fairness level it shows is decent in most cases, it is not likely that the model with such a frequent appearance of low accuracy would be used for real tasks.

5.4 Debiasing for One Single Attribute Applied to Another

The last analytical part of the results section is dedicated to checking whether a model with debiasing done on some protected attribute can transfer knowledge and improve the fairness of another attribute that is evaluated on it, and if these single attributes have bias correlations. *Table 5.4.1* shows the results of evaluating several protected attributes on the model using different debiasing methods on the age attribute. To start with, the LEACE did not show improvements in attributes' fairness values compared to the base model. The accuracy is higher, but insignificantly, so DTO has also increased a little for gender and country. The Adv method has lost some accuracy and yielded bad results for gender. Clearly, a gerrymandering issue

occurs here, forcing gender attribute fairness to fall when age attribute bias is decreased. For the country, however, Adv performed quite well, somehow overtaking the base value of fairness. It can be stated that there is a correlation between age and country that allows for the removal of bias in the latter when debiasing the former. In BTEO, correlations with age bias are also found in the country, while gender is experiencing gerrymandering again. However, the accuracy level is disappointing, so even with good fairness values, DTO yields the worst accuracy-fairness tradeoff for BTEO among all methods.

Table 5.4.1. Results for the single attributes evaluated on the models with different debiasing methods applied to the age attribute.

	LEACE			Adv			BTEO		
	age	gender	country	age	gender	country	age	gender	country
Average test accuracy	89.96	89.96	89.96	73.6	73.6	73.6	44.99	44.99	44.99
Average test TPR-GAP	5.41	2.82	3.82	5.27	4.23	3.16	3.55	2.87	1.78
Average test DTO	0.1141	0.1043	0.1075	0.2709	0.2676	0.2672	0.5516	0.5511	0.5504

The results for the debiasing model trained on gender are shown in *Table 5.4.2*. The LEACE method is a complete outsider here, as it did not show any positive results. Adv reached high accuracy this time, which even surpassed both the accuracy of the base model and the LEACE model. Unfortunately, there was no correlation between gender bias and other attributes' biases either. BTEO achieved good fairness for every attribute by sacrificing accuracy level again, and DTO also showed a bad trade-off between accuracy and fairness once more.

Table 5.4.2. Results for the single attributes evaluated on the models with different debiasing methods applied to the gender attribute.

	LEACE			Adv			BTEO		
	age	gender	country	age	gender	country	age	gender	country
Average test accuracy	89.63	89.63	89.63	89.87	89.87	89.87	65.76	65.76	65.76
Average test TPR-GAP	4.95	2.72	3.78	5.17	2.76	3.79	3.15	1.45	2.40
Average test DTO	0.1149	0.1072	0.1104	0.1138	0.1083	0.2672	0.3471	0.3435	0.3453

The final data to observe is placed in *Table 5.4.3*. The model with the LEACE method now showed a little lower TPR-GAP for the age attribute than the base model, which indicated that country and age are indeed bias-correlated. The accuracy is similar to the base model, so the DTO value is lower. The Adv technique did not perform well this time, experiencing both growth in TPR-GAP and a drop in accuracy, and gerrymandering occurred. BTEO produced extremely low TPR-GAP values, which appear unrealistic. In fact, during experiments, 2 seeds out of three had zeroed TPR-GAP, and the train accuracy was a few times smaller than the validation one. This is another instance of a problem, possibly caused by a discrepancy between the data distributions of the test and validation sets outlined in Section 5.2.

Table 5.4.3. Results for the single attributes evaluated on the models with different debiasing methods applied to the country attribute.

	LEACE			Adv			BTEO		
	age	gender	country	age	gender	country	age	gender	country
Average test accuracy	89.79	89.79	89.79	73.02	73.02	73.02	30.89	30.89	30.89
Average test TPR-GAP	4.60	2.65	3.47	5.03	4.26	3.2	0.96	0.73	0.51
Average test DTO	0.112	0.1054	0.1078	0.2755	0.2734	0.273	0.6913	0.6912	0.6911

In the end, we have a clear correlation between age and country attributes' bias, while gender bias is uncorrelated with any of the other attributes. Not every debiasing method, however, can find the hidden correlations, so it is important to differentiate the bad performance of bias knowledge transfer because of an unsuitable debiasing technique and a lack of this knowledge.

5.5 Conclusion

To conclude this section, a presentation and deep analysis of all the obtained results were conducted. It started with a description of the base model results that are used as a baseline. Then, three approaches towards intersectional debiasing and bias knowledge transfer were considered to answer the hypotheses set in Section 3.3. The relevant conclusions were drawn about the debiasing methods and protected attributes used in experiments for this project.

6. Further Work

After the results were studied, some omissions and limitations of this project became clear. Some bright ideas were also postponed due to a lack of time for their execution. The points on how to solve the emerging problems and make the research more comprehensive are presented in this section.

One way to improve the analysis of debiasing and make the results clearer in terms of algorithmic effectiveness is to choose a dataset that does not present high fairness values at the start. The problem with this dataset, which hampered the assessment of the performance-fairness trade-off, was that the change in accuracy was generally greater than the change in fairness, and the fairness metric had limited room for improvement, possibly devaluing the work of debiasing methods. Choosing datasets with more severe bias in the groups and their intersections may help to understand better how well the debiasing techniques deal with the bias in different attributes. It was also noted in Section 4.1 that discussion of multilingual debiasing and its transferability within several languages is left for future research.

Another improvement that can possibly be made is the enhancement of the LEACE method for a more effective search of hidden representations. As mentioned in Section 5.3, LEACE was able to deal with finding underlying correlations that existed in the biases of two attributes. The ability to identify such connections proves that LEACE can be improved, helping it to become a more robust debiasing technique.

Lastly, slight changes can be made to the Adv method used. This specific version of adversarial training was not robust enough to mitigate bias in the presented attributes while still maintaining the accuracy level. It can be replaced by a more effective and improved analog, such as the diverse adversaries approach (DAdv) [22].

Overall, the study of debiasing methods that can be applied for intersectional debiasing purposes should be continued, with experiments nourished by new models, metrics and methods.

7. Conclusion

This comprehensive investigation into intersectional fairness and debiasing techniques has shed critical light on the limitations of existing approaches that consider protected attributes in isolation. The findings clearly demonstrate that true fairness cannot be achieved by optimizing for single-attribute metrics alone, as this leaves the model vulnerable to gerrymandering. As there are currently no methods designated specifically to deal with bias in intersecting attributes, this project contributes to the field by examining the fairness and performance of several basic debiasing methods if applied to the intersection of groups. A plethora of such techniques was studied to choose the best candidates based on different approaches. The focus was put on three types of debiasing methods that differ depending on the time of execution in relation to training: pre-training, at-training and post-training. As a result, the three techniques were chosen: BTEO, Adv and LEACE, each of different respective type.

The experiments were conducted based on the chosen dataset, model and metrics. Some pre-processing steps were also applied to the model to ensure the highest initial performance. The results were received and discussed, with several important conclusions to be drawn. Firstly, the chosen methods are suitable for improving fairness in some single attributes with correlated bias, but they are not performing robustly and consistently. These methods still can't beat the base model in the accuracy-fairness trade-off. Hence, Hypothesis 1 is confirmed. Secondly, there are attributes, including the joint one, that exhibit correlated biases, so debiasing one of them may have a positive effect on the fairness of the other. Examples where learning single attributes positively influences the evaluation of joint attributes can be observed in Section 5.3. At this point, Hypothesis 2 is verified too. Also, there are examples showing single attributes positively influencing the evaluation of other single attributes, which are not interconnected intuitively, in Section 5.4. Precisely, we were able to deduce that age and country attributes in the MTC dataset have correlated biases. This allows us to confirm Hypothesis 3. Finally, the use of pre-training methods is not favored in intersectional debiasing tasks with reality-based

datasets. This type of method, which includes BTEO used in this project, is very vulnerable to model results if the data distribution is different in validation and test sets. They are also balancing samples in such a way that the training set becomes smaller, leading to a limited number of samples of joint attribute available.

As machine learning systems become increasingly ubiquitous, especially in high-stakes domains, ensuring intersectional fairness is not just a moral imperative but a crucial step towards building trustworthy and equitable AI that serves all members of society fairly. This work lays a strong foundation for future research to further advance the state-of-the-art in fairness-aware machine learning, paving the way for a more inclusive and just future powered by responsible AI technologies.

8. Bibliography

1. Text classification: What it is & how to get started // Levity URL: <https://levity.ai/blog/text-classification> (дата обращения: 25.04.2024).
2. Protected Attributes and “Fairness through Unawareness” // MIT OpenCourseWare URL: <https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-three-framework/protected-attributes/> (дата обращения: 25.04.2024).
3. Subramanian S. et al. Evaluating debiasing techniques for intersectional biases //arXiv preprint arXiv:2109.10441. – 2021.
4. Tan Y. C., Celis L. E. Assessing social and intersectional biases in contextualized word representations //Advances in neural information processing systems. – 2019. – Т. 32.
5. Kang J. et al. Infofair: Information-theoretic intersectional fairness //2022 IEEE International Conference on Big Data (Big Data). – IEEE, 2022. – С. 1455-1464.
6. Foulds J. R. et al. An intersectional definition of fairness //2020 IEEE 36th International Conference on Data Engineering (ICDE). – IEEE, 2020. – С. 1918-1921.
7. Gohar U., Cheng L. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges //arXiv preprint arXiv:2305.06969. – 2023.
8. Han X., Baldwin T., Cohn T. Balancing out bias: Achieving fairness through balanced training //arXiv preprint arXiv:2109.08253. – 2021.
9. Wang Y. et al. Intersectional Two-sided Fairness in Recommendation //Proceedings of the ACM on Web Conference 2024. – 2024. – С. 3609-3620.
10. Han X. et al. Fairlib: A unified framework for assessing and improving fairness //Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. – 2022. – С. 60-71.
11. Kuzmin G. et al. Uncertainty Estimation for Debaised Models: Does Fairness Hurt Reliability? //Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). – 2023. – С. 744-770.
12. Belrose N. et al. Leace: Perfect linear concept erasure in closed form //Advances in Neural Information Processing Systems. – 2024. – Т. 36.
13. Yoo J. Y., Qi Y. Towards improving adversarial training of NLP models //arXiv preprint arXiv:2109.00544. – 2021.
14. Li Y., Baldwin T., Cohn T. Towards robust and privacy-preserving text representations //arXiv preprint arXiv:1805.06093. – 2018.
15. Elazar Y., Goldberg Y. Adversarial removal of demographic attributes from text data //arXiv preprint arXiv:1808.06640. – 2018.

16. Terzi M. et al. Adversarial training reduces information and improves transferability // Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – Т. 35. – №. 3. – С. 2674-2682.
17. Huang X. et al. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition // arXiv preprint arXiv:2002.10361. – 2020.
18. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. – 2018.
19. A gentle introduction to ML fairness metrics // The observability blog URL: <https://superwise.ai/blog/gentle-introduction-ml-fairness-metrics/> (дата обращения: 25.04.2024).
20. Han X., Baldwin T., Cohn T. Fair enough: Standardizing evaluation and model selection for fairness research in NLP // arXiv preprint arXiv:2302.05711. – 2023.
21. HSE University HPC Cluster "cHARISMa" // Supercomputer Modeling Unit URL: <https://hpc.hse.ru/en/hardware/hpc-cluster/> (дата обращения: 25.05.2024).
22. Han X., Baldwin T., Cohn T. Diverse adversaries for mitigating bias in training // arXiv preprint arXiv:2101.10001. – 2021.