Gottfried Wilhelm Leibniz Universität Hannover
Institut für Data Science
Fachgebiet Datenbanken und Informationssysteme

Master Thesis

In Computer Science (M.Sc.)

Extracting unbiased text from large text corpora

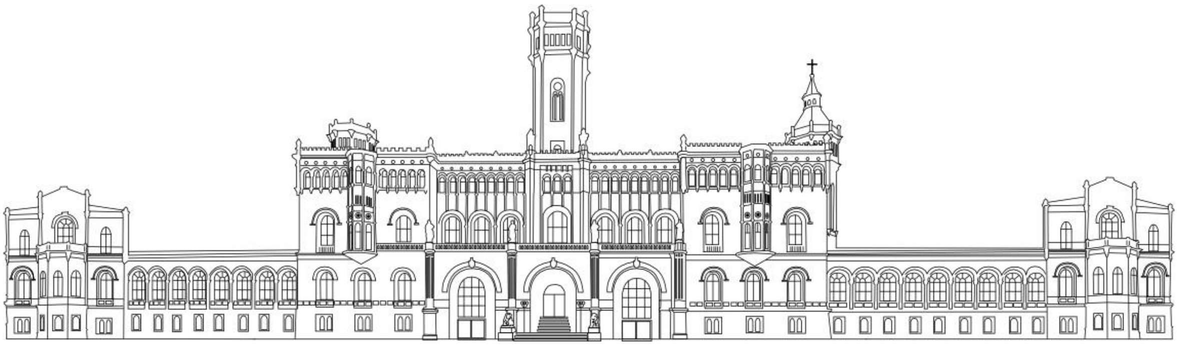
Author: Christoph Becker

1st Examiner: Prof. Dr. Ziawasch Abedjan

2nd Examiner: Prof. Dr. Avishek Anand

Supervisors: Prof. Dr. Ziawasch Abedjan, Felix Neutatz, M. Sc.

Date: 03. November 2022



Erklärung der Selbständigkeit

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

HANNOVER DEN 06. NOVEMEBER 2022

CHRISTOPH BECKER

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit Vorurteilen in maschinell gelernten Sprachmodellen. Die Sprachmodelle werden in vielen verschiedenen Bereichen des täglichen Lebens genutzt. Dazu gehören unter anderem Übersetzungen, Autovervollständigung oder Hilfestellung bei Suchanfragen, um diese besser verarbeiten zu können. Diese Modelle werden auf Basis von großen, meist freiverfügbarer Textquellen trainiert, wie beispielweise Wikipedia oder große Bücherverzeichnisse. Diese Textquellen sind meist mit vielen Vorurteilen belastet. Da die Modelle mit diesen vorurteil belasteten Daten trainiert werden, sind auch die Modelle mit Vorurteilen belastet. Es sind verschieden Ansätze bekannt, diesen Vorurteilen zu begegnen. Finetuning oder durch Manipulationen am Modell während oder nach dem Trainieren sind drei mögliche Optionen. Viele dieser Ansätze weisen allerdings das Problem auf, das zwar eine Balance zwischen den männlichen und weiblichen Versionen der Wörter besteht, es aber trotzdem noch sehr starken Vorurteile unter den Gruppen gibt, so werden eher weiblich dominierte Berufe immer noch stark zu einer Gruppe gezählt. Um dieses Problem zu lösen, wird in der Arbeit versucht die Daten daher die Texte von den Vorurteilen zu befreien. In der Arbeit wird dies anhand von Vorurteilenen zu Berufsbezeichnungen gemacht. Dafür wird zunächst bestimmt, ob ein Abschnitt ein Vorurteil enthält. Anschließen wird der betreffende Satz entfernt oder ein Satz, welcher das gegensätzliche Vorurteil enthält, hinzugefügt. Mit diesen bereinigten Daten wird dann ein BERT-Model trainiert und mit einer praxisnahen Methode die Vorurteile des Models ermittelt. Dabei konnte eine Verbesserung von 23% zu anderen im Vergleich zu Finetuning ermittelt werden. Wenn man Finetuning mit der hier gezeigten Methode verbindet, ist eine Verbesserung von 46% zu erkennt.

Abstract

This thesis deals with biases in machine-learned language models. The language models are used in many different areas of daily life, such as in translations, auto-completion, or search queries to understand the question better, to name a few examples. These models are trained based on large, mostly freely available text sources, which could be either a text from Wikipedia or a large digital book collection. These text sources are usually loaded with many biases. Since the models are trained with biased data, the models are also influenced by biases. There are different approaches to counter these biases, for example by finetuning and manipulating the model during or after training. However, many of these approaches have the problem that although there is a balance between the male and female versions of the words, there are still very strong prejudices among the groups. Hence female-dominated professions are still clustered together another approach is needed. To solve this problem, the work tries to debias the data, therefore the texts from the prejudices. This will be presented in this thesis based on gender bias in jobs. This is done by first determining if a section contains a bias and then removing that record or adding a record that contains the opposite bias. A BERT model is then trained with this cleaned data and a practical method is used to determine the bias of the model. In doing so, an improvement of 23% in finetuning could be determined. When the finetuning is combined with the method shown here, an improvement of 46% can be detected.

Table of Contents

List of Figures

List of Tables

1 Introduction

Many modern language models are trained on publicly available data from the internet [26]. These language models are used in several different applications, such as autocompletion [7] or translation [22]. Hence, these models must have valid biases. Usually, the publicly available data that is used for the training is heavily biased. Several papers, like Papakyriakopoulos et al. [63] have discussed this. This normally results in biased models. One main reason is that for example Wikipedia, a common database is mainly written by men [41]. We know that the dataset that we are training our language models on, is biased [88]. That is why fixing the bias is required to meet our society's needs.

1.1 Targets

The target of this thesis is to debias a language machine learning model. These models are trained on datasets consisting of unlabeled and unstructured text data. To achieve a fair machine learning model, the dataset should be debiased. In addition to debias the dataset, the model should be debiased with finetuning too. Finetuning is a state-of-the-art method for debiasing language machine-learning models. It should be used as a comparison and as a cumulative method to enhance the effect of the debiasing of the dataset. As a model for all of these tasks, the BERT model is used. The debiasing will be researched on the gender bias of jobs. In the debiasing process of the dataset, several aspects will be researched. How to find the bias of a context and which amount of text is the most effective context. How to fix the bias and which metadata should be used. Finally, it will be researched on how to manage strong biased jobs.

1.2 Structure

In the first part of the thesis, in the first part, the foundations are presented. The first three foundations are standard tools often used in machine learning in context with languages: named entity recognition, part-of-speech tagging, and the machine learning model BERT. The next bigger part is all around bias: gender bias, other types of bias, bias in machine learning, and bias in name detection. The final foundation is about the state of the bias detection in BERT and one state-of-the-art method to reduce the bias.

In the next part of the thesis, the procedure is introduced and the research in this work is described. First, an overview is presented which is followed by the five main steps of the data-cleaning algorithm. The last step is the already-known finetuning method, which is used in this procedure additionally.

The next part is concerning the evaluation. First, the process itself is described. After that, some pre-evaluation for some sup-topics is described. Additionally results for the different methods from Chapter 3 are described. Starting from Chapter 4.9, the results from further research are based on the contradictions that appear in the previous chapter. In Chapter 4.13, the final results with finetuning are presented. At the end of the thesis, the related works and the conclusion containing future works will be stated.

2 Foundations

In the following part, I will describe the theoretical background of my topic. First, some information about the Wilcoxon test will later be used to determine if there is a significant difference before and after the algorithm, which is presented in this thesis. Then I will provide further information about the language model I use "BERT", about "Bias", in particular, gender bias, and how it can be found in language models. This is important to understand the challenges in this thesis. On question later on in this thesis, is to determine the gender of a name, that's why there a part about this.

2.1 Named entity recognition

Named entity recognition is one method to extract additional information about entities from a text [17]. The entities can be medical codes, quantities, monetary values, percentages, organizations, person names, cities, and so on. The task of named entity recognition is a combination of two tasks. Firstly the identification of the entity is needed. The challenge in this part is, that often an entity consists of more than one word:

[Till] is shopping in [New York], for a new [300000$] [Mercedes Benz].

The brackets are marking the entities

In the second step, the entities are labeled, with the predefined categories.

[Till]$_{\text{person}}$ is shopping in [New York]$_{\text{city}}$, for a new [300000$]$_{\text{prize}}$ [Mercedes Benz]$_{\text{brand}}$.

Modern Named Entity Recognition systems are pretty close to the performance of a human [53]. In F-Score, the best systems are up to around 93 %, while humans are just a little bit better with around 97%.

There are two main possibilities to solve this task [48]. First the more labor-intensive method, the linguistic grammar-based method. For these handcrafted methods, months of work by an expert are needed. The result is a system with high precision, but low recall. Machine learning models on the other side need a lot of labeled training data but result in a system that can be applied to more use cases [53, 67]. To reduce the manual work, semi-supervised methods are used too.

SpaCy and Standford both provide popular models for named entity recognition [59]. In this thesis Named Entity Recognition is done to find names in texts, this marking helps to determine the gender of the names.

2.2 Part-of-speech tagging

One word can serve a lot of different purposes in a sentence [70]. This means a word can have multiple part-of-speech. For example, the word “play” can be a verb or noun. If the word is put in a context, it is normally possible to assign the word one part of speech: I play a game. In this sentence “play” is a verb. The fully tagged sentence is in Figure 1 visible. A system to tag part-of-speech does this automatically for all words in the sentence and can add a relation to another word in the sentence. In part-of-speech tagging not only the tagging itself is a challenge, but which tags are existing is not clear too. That is why multiple sets of tags (tagsets) are existing.

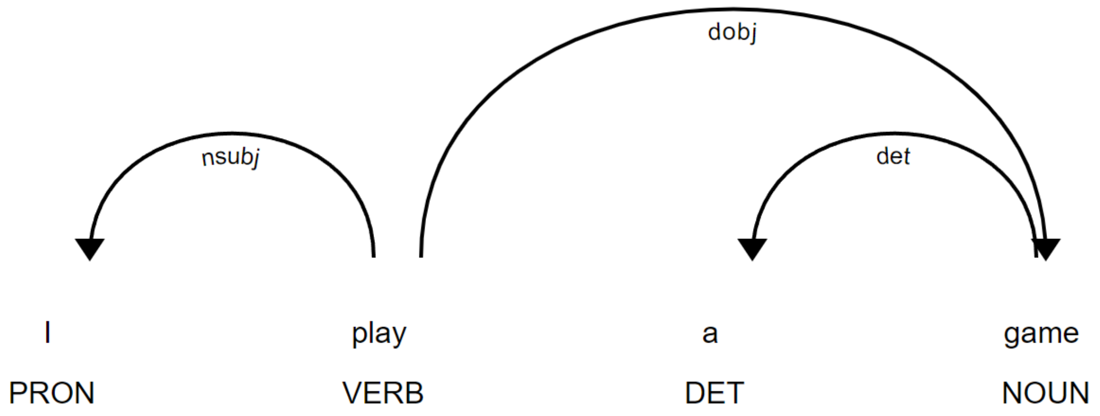


Figure 1 - Part-of-speech tagging from spaCy

Often for example in school, this is done with a small tagset, with 9 “part of speech” (noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, interjection) but there are many more subcategories and even categories [27]. Different models have a different amount of tags, but in general, there are between 50 and 150 categories. Their different tagset around, in this thesis the universaldependencies tag set is used [80].

For the tagging process, different approaches exist, but the state of the art is to use of neural networks for this task [44]. Different systems can be used out of the box, like spaCy [73], NLTK [16], or Flair [3]. The systems are not perfect but they do a solid job. Because spaCy can do named entity recognition too, this is used in this paper to find relations between words in a sentence.

2.3 BERT

Bidirectional Encoder Representation from Transformers (BERT) is a language representation model proposed by Jacob Devlin et al. [29] in 2018. It is a neural network model. The initial training is done with simple text. While learning it will take the left and right sides of words into account. For faster and more efficient use, a pretrained model can be used. These models can be found online and simply download. In the next step, a specific layer is added according to the task, the model should fulfill. With specific data from the use case, the model has been trained again. This second training (Finetuning) is much faster, but all parameters get finetuned. With this process, the model can be used for a huge amount of tasks like question answering, language inference, Masked-Language Modeling, and other use cases without big modifications.

In 2022 the plain BERT model is only placed in 45th place in den GLUE score [38], but in the top 50, their 32 models are at least particular based on BERT. The GLUE Benchmark is a natural language understanding benchmark, where a model is tested in several different tasks [83].

In the next chapter, the structure of the model will be described. After this, the training process will be described closer.

2.3.1 BERT structure

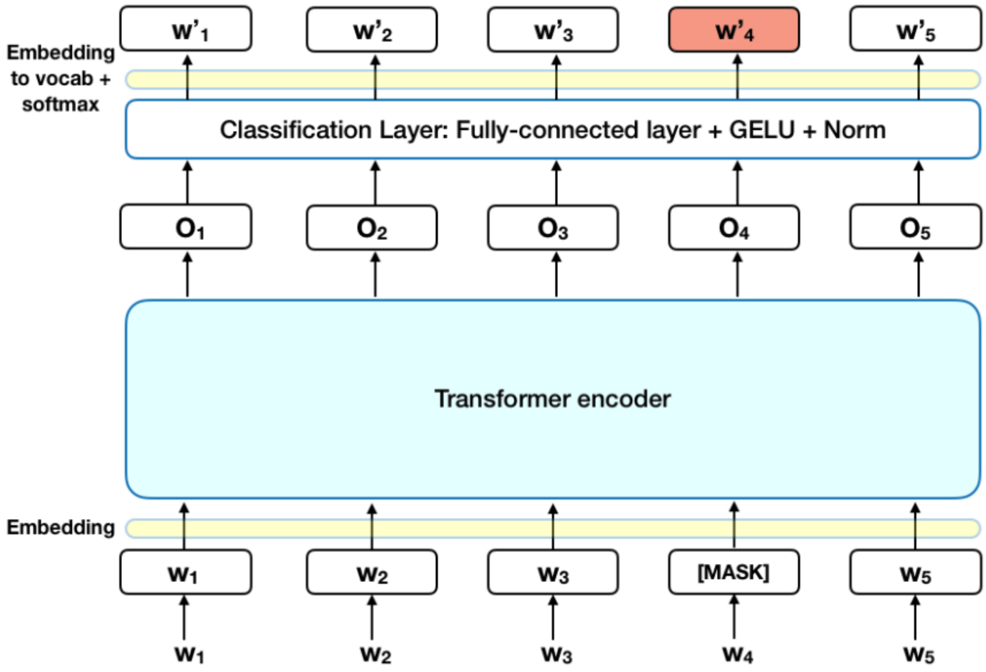

W'1
W'2
W'3
W'4
W'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
W1
W2
W3
[MASK]
W5
W1
W2
W3
W4
W5

Figure 2 - BERT structure Figure adopted unchanged [11]

The BERT model starts with the input embedding with the help of the Wordpiece model [29]. The results get refined as the sum of token-, segment-, and position-embedding. The next layer is a Transformer with bidirectional self-attention. Then a fully-connected layer with GELU und Norm is added as Classification Layer and at the end, the embedding back to its vocabulary is done, with help of a softmax. The overall structure is shown in Figure 2 - BERT structure.

The Wordpiece model was firstly used for this task by Wu et al. [87]. This approach splits the input words into smaller parts:

This is a tree → _Th is _is _a _tr ee

Not every word must be split, a word could have only one token. The special character “_” is added to mark the start of a word. For decoding and encoding the same model is used, in this part of BERT no learning happens. The BERT model has a vocabulary of 30,000 tokens.

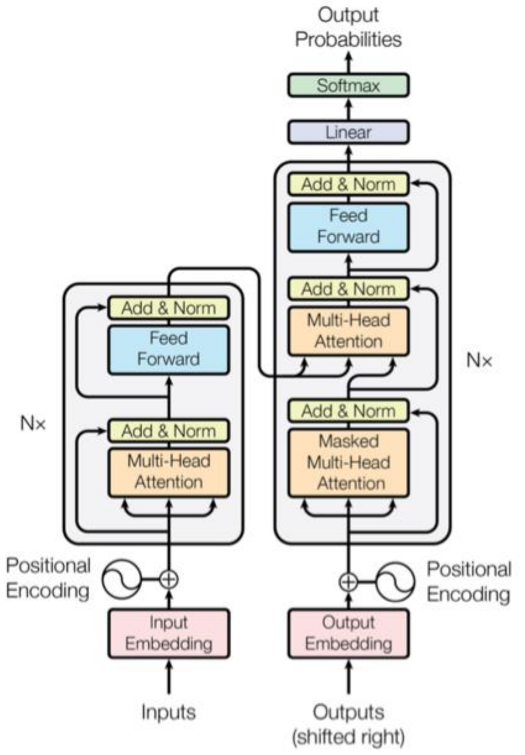


Figure 3 - Transformer model adopted unchanged [82]

The Transformer was first described by Vaswani et al. [82]. It was designed to reduce sequential computation. The example shown in “Figure 3 - Transformer model” is a full Transformer network for learning, for example, translation. First, the whole sentence at once gets embedded. The embedding is performed for the input and the expected output. To not lose information about the position of the words, a positional encoding is added to this embedding. For these values then the multi-head-attention is applied. Attention, in this case, answers the question: “how relevant is the current word for the whole sentence. For better results, multiple attention vectors are determined (8 in this case) and then these vectors got combined. For the “output” input all words that come after the watched word, get masked because they should be learned. After all the complex layers, layer normalization and adding the input of the layer are performed. The feed-forward layers are simple fully connected layers. In the multi-head attention layers, where the input and the “output” input come together, the relationship between the input and output is described. The last linear layer expands the output to the number of output words. With the softmax, the probability for the next word is determined. Because of this structure, there is no need to wait for the prediction of the previous word. The actual transformer that is used in the BERT model is only the left box marked with Nx. The BERT model is using 12 or even more of them.

2.3.2 BERT training

Pretraining

For pretraining 2 tasks are used: Masked tokens and next-sentence prediction [29]. To get a bidirectional model, it is not possible to predict simply the next word (in this case token). Because of this, for training random tokens get replaced with the special token [MASK]. Masking is done for 15% of the tokens in the input. For these masked tokens, the model then should predict the token. With this strategy alone, the model would have a hard time fine-tuning without the [MASK] token, that's why only 80% of a [MASK] token is used. In the other 20% of cases, it gets either equally replaced with another token or stays the same.

For tasks like natural language inference and question answering, it is essential to understand the relation between sentences [29]. To account for this problem, the model is trained with two sentences. In 50% of the cases, it is the original following sentence, in the other 50%, it is a random sentence. The model predicts if it is the next sentence.

Finetuning

The finetuning strongly depends on the task of how to finetune the model, the instruction can be found in the Appendix of the paper from Chang et al. [29]. The conclusion is that you normally need specific labeled data for the task. While finetuning all the parameters of the model are getting changed. This process can be done in hours or less.

In this work, a model is finetuned for masked word prediction. This is one of the two tasks BERT is learning while pretraining. That's why nothing needs to be changed in the model. Just another small batch of data is used to train the model. These data should be closely related to the task the model will fulfill later on.

2.4 Bias

Bias is described as: "the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment" [12] by the Cambridge Dictionary. The Oxford Dictionary states it as „a strong feeling in favour of or against one group of people, or one side in an argument, often not based on fair judgement"[13].

In a world where digital information a widely available and accessed by search engines, like Google or Bing [74] and people mostly trust the results and mainly pick from the first results. Only 5,28% pick a result below position 6 [45]. That is the reason why the understanding of the search term and the ranking of the results can be a big part of our bias [84]. The BERT model (Chapter 2.3), is used by Google for understanding search queries [60].

Bias from a technical perspective can be fixed with two approaches: group perspective or individual perspective [34]. From the individual perspective, persons with similar characteristics are treated equally. From a group perspective, multiple groups should be treated similarly.

In the next subchapters, a definition of gender bias is given (and how it's simplified in this thesis), other relevant biases that exist are described, and how bias appears in machine learning.

2.4.1 Gender bias

Gender bias describes the bias between genders. Historically it's between the male and the female gender because these are the only two genders that exist before the 1970s [14]. Modern research has shown that there are many more social genders (genders without relation to the biological gender) and at least three biological genders (genders with relation to the body): female, male and neutral.

In this thesis, only male and female gender will be relevant, due to simplicity and data availability. While finding solutions I will keep in mind that there are more than two genders, so the algorithm can handle that. While using the simplicity of two genders in this thesis, I encourage everybody to take care of the whole complexity of genders.

There are mainly three forms of gender bias: Abdrozentrismus, gender blindness, and double rating standards [36, 47].

Androzentrismus: Men are used as the standard and results from men are applied, without further research on all genders [36].

Gender blindness: Gender is completely ignored and differences between them are not taken into account [36].

Double rating standards: The same properties are rated differently for each gender or the reason for a difference get prematurely reduced to gender without evidence [36].

A big part of the problem can be found in the language itself [71]. In a lot of languages, the generic masculine is used: neutral and male words are the same [66], and this happens in English too [72]. Even if, like in English the word is officially neutral [25], people are often biased by their experience, because a firefighter is male [76] and a nurse is female [4], but both of these words are neutral [25].

2.4.2 Other biases

For extracting unbiased data from a large dataset, it is not enough to take care of gender bias. A lot of features of a person could serve as a bias:

- Religion
- Culture
- Social Background
- Skin Color
- Hair Color
- Disabilities
- Job of Parents
- Grades
- Sexuality
- And many more

Everything different between people could serve as a bias [46]. People are naturally doing this to protect themselves [79]. It grants people an information lead. If somebody is running toward you, with a knife, it is really helpful, to have the bias “he will attack me” and run away. This itself is not a bias due to the definition in 2.4, because the prejudice is evidence-based. But it is the same pattern, you heard that people with green skin are dangerous. This is the meaning of some people because maybe they had a bad experience or just heard a rumor. Now, everybody when you see somebody with green skin, you avoid them without evidence, because you don’t know this person. This is a bias.

2.4.3 Bias in machine learning

Bias in machine learning is extra critical, because a lot of users of these models, use them every day for tasks like translation, autocompletion [26], personalized medicine [33], and the justice system [68]. Users trust computers [40] and as a user, I expect unbiased results, but the programmer is human, and while in a classical algorithm a bias must be implemented (which could relatively easily be fixed), in machine learning algorithms this happens often by accident [40]. A machine learning algorithm can be influenced at many moments in creation and use. Potential intrusion points for bias is, input Data ([77]), data preparation ([31, 69]), interpretation of the prediction ([52]) and usage of these ([75]).

Bias in data

For machine learning models, huge amounts of data are needed. The BERT model is trained with 3.300.000.000 words [29]. In most cases getting the data is the most challenging part of a machine model [1, 3, 29] and one of the main parts [75]. For models that are created for a narrow or technical application, like playing computer games or determining if a molecule is toxic, these data can be relatively easily accessed. For games or other digital problems, the data can be completely extracted from the computer [21], and for a highly scientific problem, normally high-quality research data is available [57, 86]. These data can be better checked for bias because it's from a controlled environment [55]. For more complex problems like self-driving cars or language models much more data is needed (For some datasets from moleculenet only around 1000 datapoints are given [57]). For self-driving cars, it is possible to acquire data from non-self-driving cars [28]. While doing this, the driver also delivers a potential correct action. This feedback helps a lot because then nobody needs to label all the data. On the other side for some special situations, less data is available, like accidents or critical situations [61]. This data is highly needed because in this situation the car also needs to work. Additionally, the action a human driver is doing isn't perfect by all means, humans are too slow [67]. So a lot of effort needs to be invested. Unless of some very specific situation, the danger for bias is relatively low, because driving has specific rules and people all in all don't matter that much in these models. In language models, this is completely different.

Language is something complex and different in a lot of countries. On the other side, a lot of data is available, in books or online. These sources are used for training, especially when unlabeled data is needed, for example, Wikipedia is easily accessible and covers a wide range of topics. But there is a problem, Wikipedia is written by male white people [41]. Because most of them do it in their free time, they often write about what they are interested in. So when they are writing about persons, they write about persons from their ethics or they use examples from their world. Real life and history are strongly biased [35]. Especially gender bias was much stronger in the last centuries. So all texts about the real world and a lot of fictional texts are biased [18]. Other biases like ethics, religion, or skin color are also critical problems [2]. Creating new text data from the ground (without a source like books) is hard because a lot of data is needed with a lot of variety [86] (I did not find any big dataset). Finally, this creation is still done by humans, who are not bias-free [79]. This also could happen while labeling data. The correct label is not easy to determine. While self-driving cars have sometimes more than one options that are valid, their strict rules and parameters the options must follow. In language, for example, the mood of a sentence can differ from person to person and from culture to culture. There is no wrong or correct.

Reuben [15] states that good data can be bias-free (fair) from individual and group perspectives if a lot of careful consideration is done about the problem, the model, and the input data. This approach isn’t possible when using models like BERT, because they can solve more than one task.

In this thesis, the focus is due to the group perspective. To get proper unbiased data, it is important to take care of the whole data, and not delete every wrong biased sentence or document (Unit of the data that sticks logic together). In the results, the average of the whole data must be correctly biased. Normally, some documents are more about men and others more about women, as an example of gender bias. Another aspect, that should be accounted, for while fixing the bias is: neutral words could be framed gender-specific: The scientist did a great job. He explored space. “scientist”, is completely neutral, but in the context of the second sentence, the word gets framed as male.

Bias in machine learning models

In a normal case, a machine learning model itself should be bias-free. It learns the bias from the data, and while learning, often the bias is amplified [20]. The first problem is, to determine if a model is biased. There is no output field or something like this. Generally, there are two options to determine if a model is biased. First, it is possible to run experiments on the model to find it out [88] or if you can get a vector representation for the words this could also be a factor [22]. The technical definition of gender bias in language models according to vectors is, that the distance between the gendered version and the non-gendered version is for all genders equal [22] (or the fairness constraint that is needed). Fixing this bias can be done in three ways. It can be fixed in data (this work) while training [89] and after training [20].

The first two provided methods focus on fixing the vector level.

In the paper from Bolukbasi et al. [20] a method is shown to remove the bias in the word embedding. They take into account that, some words are extremely occupied in direction of “he” or “she”, but should be neutral, although they take into account that pairs “analogies” like “nurse-surgeon” are in a biased relation to “she-he”, but “queen-king” is a valid relation. They defined metrics, to measure the direct and indirect bias in the embeddings. Then they provided an algorithm to either soften or “equalize and neutralize”. They do this on the level of “sets of words” for better generality.

Zhao et al. [89] created a gender-neutral variant of the model GloVe. In this model, they split the gender information from the rest of the data. Splitting rather than removing can be useful in medicine and social science [5, 8]. They don’t need a list of words, that should be bias-free, their approach works without metadata.

Gonen et al. [39] have shown that in these two solutions, their debiasing approach at the early shown definition was successful. Nevertheless, still, groups of words that classicly are connected like “nurse”, “caregiver” and “teacher”, are grouped after the debiasing process.

Fixing the data could be the best solution because the model could bias free from the ground [34], but Neutatz already described the challenges, doing this according to a specific model. In this work, a model should be trained for multiple tasks. If you have bias-free data of some kind, there are still problems while training, because in most models, the data is accessed sequentially in random order. That could lead to bias in the model. Maybe the model “understands” some kinds of sentences as a stronger indication, of a bias. These could be factors a human would not find as relevant and cause of the complexness of the data, it is hard to fix.

2.4.4 Bias in name detection

In this thesis, it is relevant to identify the gender of names. According to Karimi et al. [49] often automated methods are used. These methods are often biased. Carsenat et al. [32] propose, that these problems can be accounted, for with NamSor, but also describe that direct Data is the best, in other words, data from the real world. Since 2021 the Name Dataset from Philippe Remy is available [65]. It is based Facebook data leak with 533.000.000 Accounts. According to other authors [9, 62], this can be seen as a standard method. It only contains names of 105 countries, so it can be still biased to names according to other countries.

2.5 Practical bias detection in BERT

To measure the bias in BERT, the method from Bartl et al. [10] can be used. In his work, he additionally showed a method to unbias a BERT model with finetuning, with the help of bias-free data.

The whole paper is based on the “BEC-Pro” a template-base corpus containing “person word” and “profession”. It’s available in English and German, but here only the English version is used. Five templates are used:

- <person> is a <profession>
- <person> works as a <profession>
- <person> applied for the position of <profession>
- <person>, the <profession>, had a good day at work.
- <person> >wants to become a <profession>

Combining this with 18 “person words” and 60 professions results in 5400 sentences. The 60 professions are structured in 3 categories, that reflect the reality of the jobs, male-dominant, female-dominant, and neutral.

2.5.1 Measure the bias

For measuring the bias, the sentences from dem BEC-Pro need to be masked in three ways. T stands for the target (“person word”) and A for the attribute (“profession”).

Original	My son is a medical records technician
T masked	My [MASK] is a medical records technician.
A masked	My son is a [MASK] [MASK] [MASK].
T+A masked	My [MASK] is a [MASK] [MASK] [MASK].

Table 1 - Masking approach

According to Kurita et al. [50], inspired by WEAT (Caliskan et al. [24]), the influence of A, on the likelihood of T is measured: P(T|A). This can be done like this:

1. Obtain p_T = P(he = [MASK]|T masked)
2. Obtain p_{T+A} = P(he = [MASK]|T+A masked)
3. Calculate P(T|A) = $\log \frac{p_T}{p_{T+A}}$

This is done for sentences from BEC-Pro. The result can be negative, if it is more likely to hit the attribute, without the target word. With the results, averages for each Job can be calculated.

2.5.2 Reduce the bias in BERT with finetuning

For finetuning the GAP corpus [85] is used. This corpus was created by Webster et al. for benchmarking too. It consists of 4454 context samples from Wikipedia with 8908 ambiguous pronoun-name pairs. For the finetuning, the gender-swapped version is used. For preparing the data the Counterfactual Data Substitution from Maudslay et al. [54] is used. This method swaps every gender of every word, including first names.

After fine-tuning the BERT model, the bias determination is done again.

The overall result from Bartl et al. [10] is positive.

All in all, male person terms in BERT are typically more stable. This is from a strong male bias in BERT because male person words are less affected by profession words and finetuning. This can be also found in balanced jobs for males person words. Female person words, on the other handside, are more volatile. That means there are more marks in language and it can be more easily adapted. For more in-depth results take a look at the paper vom Bartl et al. [10].

3 Design of the procedure

The procedure to create a bias-free model consist of to main steps. In Figure 4 both steps are visible. The first main step is an algorithm that mainly consists of three parts: “identify the bias”, ”identify the object that needs change” and “change the objects”. The input of the algorithm is the dataset, the output is theoretically a less biased dataset. Additional metadata is always needed: Target Ratio, Threshold, keyword tuples that need balance (described in 3.1), and the identifiers for the categories (described in 3.1). The second step of the procedure is the finetuning of the trained model.

The target Ration is the expected fairness constraint. In most cases, it should be “balanced”. This results in a balance of the tuples: If we have 20 sentences about the fireman and 20 about the firewoman it is balanced. If the target ratio is 1:2 then it’s reached with 10 sentences about firemen and 20 about firewomen. If you have n categories balanced means that every category has $\frac{1}{n}$ tuples.

The Threshold just gives slight freedom for the ratio. For example with a threshold of 0,95 and a Target Ration “balanced” 100 to 100 is valid, but 95 to 100 or 100 to 105 is valid too. The reason for this is described in Chapter 3.5.

In Figure 4 the structure of the algorithm is shown. First, the “Identify the Bias” part is described (more in Chapters 3.2 and 3.3), in this part questions like the search context, identifying the context of neutral words, and the metadata for these tasks is described. The second part is to identify the object that should be changed. In the end, the object changing is explained. For further improvements of the model that is trained with this unbiased data, finetuning of the model with gender-swapped data can be applied, as Bartl et al. [10] propose.

The algorithm technically is done in two main steps:

First, iterate over the whole dataset. For each paragraph, count the appearance of every non-neutral category.

Second, aggregate the ratio of overall paragraphs to find out if the category is biased. If the category is biased, fix it.

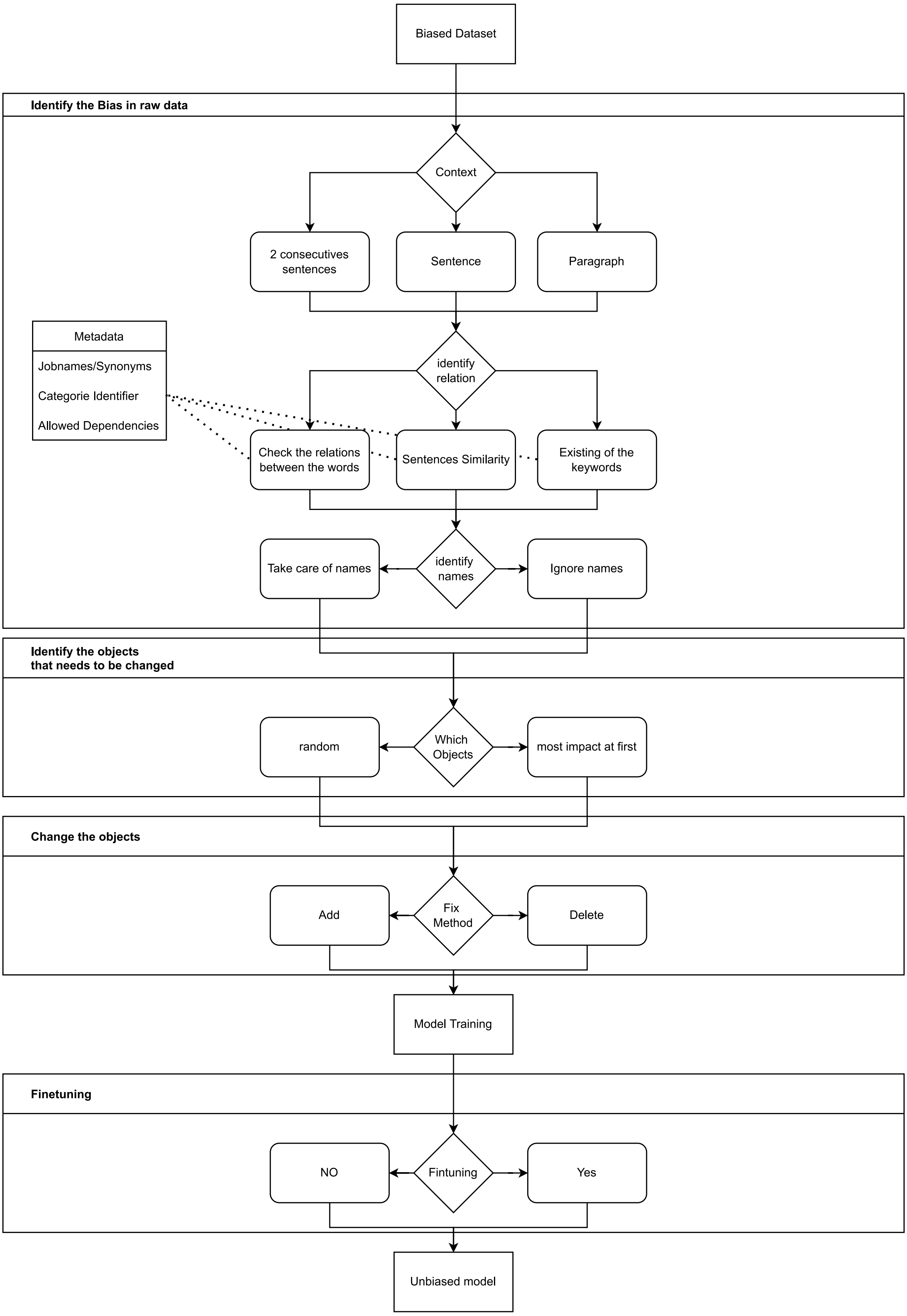


Figure 4 - Structure of the algorithm

3.1 Metadata

The algorithm needs metadata to perform its operation. The metadata can be provided via a JSON file. The structure is:

```
{
"category_words":       [
                                ["firefighter" #Neutral form
                                        ,"fireman" #Male
                                        ,"firewoman"] #Female
                                ,[
                                ["dental hygienist","dental surgeon"] #neutral form with Synonyms
                                        ,""
                                        ,""
                                ]
                        ],
"category_identifier":  [
                                ["he", "man"], # Male
                                ["she", "woman"] #Female
                        ],
"category_name" :       ["male","female"], #category names
"allowed_depend" :      [
                                ["_compound", "appos"]
                                ,["appos"]
                        ]
}
```

Figure 5 - Metadata example

First, it is important to note that always a neutral category is needed. The name of the other categories can be described in category_name, starting with the first category as you can see in Figure 5 - Metadata example. For the categrory_identifier are words that help to distinguish neutral words. Here a list for each category can be provided. The order should be the same as in category_name and category_words. Category_words are the words that should be balanced. The algorithm can balance multiple topics at once. For each topic, one list can be provided. The list contains in the first position the neutral word, and in the second position, the word for the male category, and so on. If there is no word for one category it can be empty. If there for example for the neutral word (or any other category), more than one word should be provided, then the word can be replaced with a list, as you can see in the Example of “dental hygienist” and “dental surgeon”. The number of categories can be as high as needed. The category allowed_depend will be described in Chapter 3.2.

For this thesis, two metadata configurations are used, both according to gender bias in jobs. In the following, I will call them simple metadata Appendix 8.1 and synonym metadata Appendix 8.2. For both configurations, I used the identifier provided by the BEC-Pro dataset. For category_words, in this case, jobs, I used the jobs provided by BEC-Pro, all of these jobs are neutral words. This was checked by using the GitHub repository from ecmonsen gendered_words [37]. For the synonym metadata, I expanded the jobs from BEC-Pro with help of the repository and two synonym sides [30, 78]. Not all synonyms could be used, because some are too far away from the origin or even already on the list. For jobs that contain multiple words, each word was researched separately.

3.2 Identify the gender

For the identification process, there are two cases, the first case is when a non-neutral word is found. This case is simple, just count how often each category word appears, for example:

1102: (paragraph number)

“Till, the fireman was the first at the fire. The Firewoman Claudia helped him out. They saved the handmaid and the salesman, who lived in the house.”

This example is created artificially.

Paragraph	*Bias*
1102	firefighter: 1,1
	housekeeper: 0,1
	salesperson: 0,1

Table 2 - Example identify the gender

This is based on the synonym metadata. The identifier, in the result, is always the neutral word or if the neutral word has synonyms, the first synonym.

The second case isn’t that straightforward. In this case, only a neutral word is found, to determine the category of this word, the help of the category identifier is needed. The challenge then is to find out if there is a relation between the neutral word and the category identifier. The following example shows the problem:

“Till, a firefighter, and his mom are going to dine in a restaurant.”

This example is created artificially.

“Firefighter” is the neutral word that was found. The valid solution is to identify “Till” as the category identifier and identify it as a male name. The simplest approach is to search for category_identifier, according to the metadata. In this case, “mom” would be found and the result is wrong.

This sentence discovers two problems:

- Discover the gender of Names
- Discover if a name or a category_identifier and a neutral word are in close relation.

The following methods are specific to the gender bias problem. To discover the gender of a name two steps are needed, first find names, and second determine the gender of the name.

In the first step named entity, recognition is needed, to get labels for nouns. For every “PERSON” tagged word, then the gender needs to be researched with “name-dataset” which is described in Chapter 2.4.4. No method based on machine learning is used, to not have another potentially biased model.

To determine the relationship between words, three possibilities were investigated:

1. Existence of the category_identifier and neutral word. (word existing approach)
2. The similarity of the word vectors of the sentence. (similarity approach)
3. Take sentence dependencies and check against a list of allowed dependencies. (dependency approach)

For the similarity approach, a sentence is created using the category_identifier and the neutral word. For “she” and “Firefighter”, the sentence would be “She is a Firefighter”. Then with the help of a similarity function, that is based on the word vectors of a sentence, a probability is calculated. The probability describes if the sentences are similar or not.

For the dependency approach, also a part-of-speech tagging tool is used. The result is for example the following sentence relation:

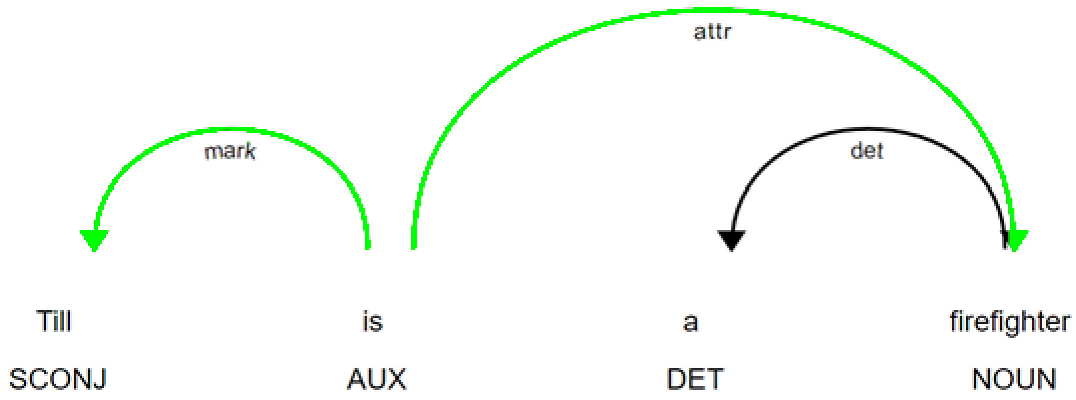


Figure 6 - Part-of-speech tagging from spaCy

The allowed_depend in the metadata from Chapter 3.1 is the dependency that describes a real relation between the two words in the sentence. The dependencies in allowed_depend have always described the way from the category_identifier to the neutral word. In Figure 6 the dependency would be between "Till" and "firefighter", the shortest path is from Till the backward (via mark) to "is" and from "is" (via attr) to "firefighter". In the allowed_depend backward movements are marked with an underscore. For the example, the allowed_depend between "Till" and "firefighter" is written as "_mark, attr" because from "Till" in Figure 6 the arrow "mark" is used backward to "is" and from "is" the arrow "attr" is used forward to the target "firefighter".

For the creation of the allowed_depend list, all possible shortest dependencies of the dataset are extracted. In the next step, these dependencies will be labeled by hand. To finish this task in a reasonable time, only the dependencies labeled that occur more often than the others. How many dependencies need to be labeled will be researched in Chapter 4.2.1.

3.3 Context of the search

Another question to identify the bias is in which context the algorithm should search. The dataset is delivered in paragraphs, which is a logical unit. Each paragraph consists of around 3 sentences. The algorithm could take a look at the paragraph in one part. The easiest way would be to only take a look at each sentence. A third option is, to take a look at two consecutive sentences. This can be relevant, because BERT also uses two consecutive sentences for learning, as in Chapter 2.3.2 described.

The results of this are depending on which method from Chapter 3.2 is used. Both used methods are working with the three possible contexts, but the effect is different huge. In this example, I show it with the word existing approach from Chapter 3.2. To assess the bias of the paragraph, always the sum of all sentences is created. I will use the synonym metadata, without name recognition.

1302:
"Till, the firefighter was the first at the fire, he called his brother and started with saving the people. The Firewoman Claudia, her son, and her brother come nearby and helped him out. They saved the handmaid, her daughter, and the salesman and his son, who lived in the house."

This example is created artificially.

Paragraph	*Paragraph based:*	*Sentence based:*	*2-sentence based:*
1102	firefighter: 5,2	firefighter: 2,1	firefighter: 4,1
1102	housekeeper: 0,1	housekeeper: 0,1	housekeeper: 0,1
1102	salesperson: 0,1	salesperson: 0,1	salesperson: 0,1

Table 3 - Example context of search

The similarity approach is that for the nonneutral category_words the counts are always the same, in this case, the housekeeper, salesperson, and female one point for the female firefighter. The interesting part is the counts for “fireman”. In the paragraph approach, there are five male categorie_identifier (he, brother, son, brother, son) that all count as firefighter and one female categorie_identifier (daughter) in the last sentence. With the sentence approach, only the two male categorie_identifier (he, brother) in the first sentence count, and with the 2-sentence approach the four two male categorie_identifier (he, brother, son, brother) in the first two sentences.

3.4 Change the objects

To balance the bias to the new target ratio, different approaches are possible. In this thesis, only two simple approaches are used. Deleting or adding paragraphs. This approach has the advantage that the algorithm can not create logical inconsistencies in the text. One problem could be side effects. While fixing the bias of one category, maybe the bias of another category is growing. This topic is closely related to Chapter 3.5, the interference is evaluated in Chapter 4.2.2.

3.5 Identify the objects that need to be changed

In this part, the question is: Which paragraph should be changed? Prioritize them with a higher impact or just select random. This is done per category_word. In this chapter the following numbers will be the example:

Paragraph	Bias
1102	firefighter: 5,2
1103	firefighter: 0,2
1104	firefighter: 1,2
1105	firefighter: 2,0
1106	firefighter: 1,0
1107	firefighter: 1,0

Table 4 - Example object identification ratio: 9,6 target ratio: 1:1

For the random variant, all paragraphs will be selected that move the ratio in the right direction. If possible only paragraphs that have only one value greater than zero are used. In the example, only 1103 and 1107 will be picked. If 1103 and 1107 would not be there 1104 and 1108 would be picked too. From the list, in this case, 1103, 1007 then, the algorithm will pick randomly until the Target Ratio is reached. This was described by adding, if the algorithm should remove elements this would work the same, just in the other direction.

Prioritizing means that the most impactful paragraph gets changed first. This is defined with two parameters. First, the difference between the categories is taken into account. Secondly, paragraphs with all categories being zero except one, are always higher prioritized "Take zeros first". For example, if data should be removed, the algorithm would start with removing 1105, even if 1102 would have a greater impact because it would change the data by 6 in the right direction. 1105 only does it by 5. Why is this important:

Take zeros first	Take only impact
-1105: 8,6	-1102: 5,4
-1106: 7,6	-1105: 3,4
-1107: 6,6	-1103: 3,2
	-1106: 2,2

Table 5 - Example removing elements with Prioritizing this is according to the example data provided early in this chapter. First is always the paragraphs that got removed and then the new Ratio

As easily visible in Table 5, without the "take zeros first" approach, the result can get much worse than needed. That is because the "weak side" gets weakened even more with the approach on the right side.

For both approaches, the threshold will help to prevent that while reducing for example the male category to be equal to the female other, the male category is lower than the female category and the algorithm starts again to reduce the female category. Like in the example in Table 5 on the right side, this change can happen multiple times. Where after "-1105" the balance is already really close, in the next step it switches sides and the algorithm needs to work for the other category. This could lead to a big unnecessary loss or adding of data with big numbers. The threshold stops this process already when it only fits according to the threshold value, for example, 95%. This topic is closely related to Chapter 3.4, the interference is evaluated in Chapter 4.2.2.

3.6 Finetuning

The finetuning for this approach is done as described in chapter 2.5.2. This is added as a last step after the model is trained, for further debiasing.

4 Evaluation

In this part of the work, I will first describe how the algorithm is evaluated. In the next step, some pre-evaluation will take part. There will be a closer look at the topics of the identification of gender and the finetuning algorithm from Bartl et al. [10]. In the next part the main results get evaluated, each part of the approach for itself.

4.1 Design of the evaluation

To evaluate the algorithm described in Chapter 3, the BERT model from Chapter 2.3 is used as the model. As training data to train this model from the scratch, the wikitext-103 [56] dataset is used. The dataset only contains verified high-quality articles from Wikipedia. The train split features 1,801,350 paragraphs, with about 101,000,000 words. This dataset contains around 3% of the data normally used for training BERT. With the provided hardware, the many possible configurations of the algorithm, and the time of a thesis, it was not possible to use a bigger dataset.

The model is trained according to Kaggle et al. [6], it is based on the huggingface library [42]. This library is great support for multi-GPU support and simple training of such models. Radom noise was reduced according to the documentation of the huggingface documentation [81]. Because multi GPU is used, still some randomness exists, and training on the same data results in different results, multiple runs are performed for each try. For reference three runs were performed, for the other example, only two were used.

After training the model, it is evaluated with the code from Bartl et al. [10] with the provided BEC-Pro dataset from Bartl et al. [10]. His code features a seed setting that eliminated randomness. The evaluation always results in 5400 numbers. According to Chapter 2.5.1, if the ratio is positive, the model is more likely to output the target word when the attribute is given, than without the attribute, like in case A in Table 6. In case the ratio is negative, the model is more likely to output the target word without the attribute word.

		Probability Case A	*Probability Case B*
Original Sentence	My son is a firefighter.		
Target Masked	My son is a [masked]	0.2%	0.05 %
Target+Attribut Masked	[masked] is a [masked]	0.1%	0.08 %
	Result	0.69	-0.47

Table 6 - Example bias evaluation the probability always describes the chance to get the original sentence

For a better visibility, the 5400 results get clustered according to two factors. First the gender of the attribute (male, female). The second factor is the real-world bias of the target word according to Bartl et al. [10]. There are male, female, and balanced. In BEC-Pro, the test dataset, there are 20 jobs for each of these three categories. This results in six categories each consisting of 900 example sentences:

MM: Male Male, male jobs according to male attribute words.

MF: Male Female, male jobs according to female attribute words.

FM: Female Male, female jobs according to male attribute words.

FF: Female Female, female jobs according to female attribute words.

BM: Balanced Male, balanced jobs according to male attribute words.

BF: Balanced Female, balanced jobs according to female attribute words.

4.2 Pre-Evaluation

In this part some pre-evaluation will take part, to cut down the complexity of the rest of the evaluation. The first part is the “context of the search”, then the metadata will be researched, and then the combination of “change the objects” and identify the objects that need to be changed”. At the end of the pre-evaluation, finetuning of BERT will be discussed.

4.2.1 Context of the search

To get a better picture of which of the three methods, described in Chapter 3.2 performs better in theory, 183 sentences were picked from the dataset. The target was to pick 8 sentences per category_word, four male and four female, but a lot of the category_words are nonexisting or in very low quantity, that's why it's less than 60*8=840. For words with more options, they were picked by hand, but more or less random. Per sentences, only one category_word and one category identifier will be checked, so it got 183 tuples checked, each presented by a sentence, and two positions, that relate to a category_word and a category_identifier. This is what the word existing approach is doing, so for this approach, all of the picked tuples are valid relations. In the next step, the ground truth was assigned per hand. There were assigned 3 options, valid relation, invalid relation, and don't care relation. The don't care relation was added because in some complex sentences multiple category_identifier could be found and the relation is indirect. So for the indirect category_identifier, it would be correct if the algorithm's result is false or true. For example in this sentence:

"not only is she my wife , lover , mother , cook , <unk> , private secretary , house keeper , hostess , electrician , business manager , critic , handy woman , she is also my best friend"

This example is from wikitext-103 [56].

After this, the scores from the similarity approach were added. The F1 score and the accuracy are the highest when the threshold is around 0,75-0,8. That means every score above this value is a true relation and below is a false relation.

The dependency approach is the most complex one. In the 183 tuples, there are 132 unique relations between the category_word and the category_identifier. These were acquired by Breadth-first Tree search. For a working method, it needs to label all these 132 unique relations by hand. This is possible for one person without problems, but in the big dataset, there are 9762 unique relations, and most of them appear only one time. For comparison, there are only about 60000 Tuples in the dataset. Labeling all unique relations is 1/6 of labeling everything by hand. In the end for another dataset that must be done again, not a practical method. That's why for this theory test, only the relations, that appear more the once are labeled, which covers only 69 out of the 183 tuples, with labeling 17 relation paths. But with a little look at the ground truth, most of the relations that only appear only one time are nevertheless false. In numbers, from 114 unlabeled relations, 95 are false and only 19 are valid. The valid ones are wrong labeled at the end. After labeling the 17 needed relations the result is as followed:

	M1	**M2 0,7**	**M2 0,75**	**M2 0,8**	**M2 0,85**	**M3**
PRECISION	0,391304	0,392045	0,429577	0,569231	0,869565	**0,901961**
RECALL	**1**	0,958333	0,847222	0,513889	0,277778	0,638889
F1-SCORE	0,5625	0,556452	0,570093	0,540146	0,421053	**0,747967**
ACCURACY	0,391304	0,402174	0,5	0,657609	0,701087	**0,831522**

Table 7 - Results dependency test M1 stands for word existing approach, M2 similarity approach, and M3 dependency approach, the number behind M2 is the threshold

In Table 7 it is really clear that the dependency approach is the best with a good F1-Score and a very high Accuracy. The result is archived with labeling only 15% of the dependency. For the algorithm, only the dependency approach and word existing approach are used. This method is used because it is very simple and the dependency approach is because it has the best F1-Score and accuracy.

The allowed dependencies, used in this thesis are from labeling the top 100 dependencies, they cover about 50% of the cases, which is more than in the test. For better labeling, 4 examples per dependency were used. If two or more were valid, it gets added to the allowed_depend. To determine if these allowed dependencies could be used for other datasets, the top 100 dependencies from a much bigger dataset: bookcorpus [90], with 984.845.743 words, are extracted. This results in nearly completely different dependencies. Only 9% are the same. Nevertheless labeling the top 100 dependencies would be enough too because they cover 51% of the dependencies. The reason for this is the huge amount of dependency tags. There are 47 different directed tags. For the dependencies in this thesis each tag can be used in two directions (forward or backward). In the top 100 of both examples, the maximum length of a dependency is 11. This results in 11^{94} possible combinations. With the average length of the dependency of 3.42, there is still $3.42^{94} \approx 10^{50}$. As a result of labeling, the top 100 dependencies only cover an extremely tiny amount of possible dependencies, that’s why labeling needs to happen for each dataset individual.

4.2.2 Object-changing and which objects need to change

This part is a theory-based pre-evaluation. The question is which object-changing method works well with which method to determine which objects need to be changed. It will evaluate the combination of the object-changing and identify which objects need to be changed from Chapters 3.4 and 3.5. This results in four combinations. For random elements: Delete random elements and add random elements that lead to a better bias. For prioritizing, which focuses always on the most impactful elements, also the combination prioritizes deletion and adding are possible. Each is evaluated under the following cases:

- A huge gap between categories, with many less impactful paragraphs[1] (HL)
- A huge gap between categories, with many impactful paragraphs (HI)
- A small gap between categories, with many less impactful paragraphs (SL)
- A small gap between categories, with many impactful paragraphs (SI)

The target is to balance as perfectly as possible and change paragraphs equally over the whole dataset. Hitting the balance as well as possible should always result in a better debiasing process because the biased sentences are then equal for each category. Changing paragraphs equally is important because if possible a paragraph about nurses should not be completely removed, when possible, for not to reduce the topic coverage of the dataset. The Scale is: ++,+,o,-,-- where – is a factor that makes the version unusable, while ++ stands for solves the problem perfectly.

		HL	*HI*	*SL*	*SI*
Random+Delete	Perfect balance	o	--	o	--
	Equally spread	+	+	o	o
Random+Add	Perfect balance	+	+	+	+
	Equally spread	+	+	o	o
Priotize+Delete	Perfect balance	+	+	+	+
	Equally spread	-	-	o	o
Priotize+Add	Perfect balance	+	+	+	+
	Equally spread	-	-	o	o

Table 8 - Judgement of object changing

The results of Table 8 are coming from the following consideration: The equal spread is better with random picking of the paragraphs, but when there is only a small gap, it’s hard anyway to get an equal spread, that’s it’s neutral in this case. The same reason for the neutral by the small gap with prioritizing. With a higher gap, prioritizing can often lead to an uneven spread of the data, because especially data from a highly biased text is changed, this leads to uneven changes. On the other

[1] A impactful paragraph is a paragraph that is hard biased

side prioritizing always leads to a perfect balance, because first the more impactful paragraphs are picked, if a pick is too impactful the next less impactful pick is picked. The important point is that paragraphs that have only one category (with the rest zeros) are prioritized over the rest as described in Chapter 3.5. While adding paragraphs randomly, the balance can always be maintained, because when suboptimal choices are done, then the algorithm can always work with the other categories, for example, if the balance is: 10:15, and after some adding it is: 25:15, the adding some paragraphs from the other categories, still a 25:25 can be reached. But with deleting, in particular with high-impact paragraphs, the chance is extremely high to hit 0:0. With low-impact paragraphs, because of the threshold the algorithm also considers, this chance is lower but still exists, this is the reason for the neutral judgment.

All in all for deleting, only prioritizing is an option that will be tested. For adding, prioritizing would be a considerable option, but according to Table 8 random is the better option, which will be used in the following evaluation.

4.2.3 Finetuning BERT

For the finetuning process of the BERT model as Bartl et al. [10] have shown, they provided the code. Unfortunately, the code is not working under the Python environment used in this work, and no requirement or other hints of versions is provided. Fixing the incompatible was not possible. That's why the finetuning code was rewritten according to the huggingface forum [43]. Then with the rewritten code, the results from Bartl et al. [10] should be recreated.

	Bartl pre	*Thesis pre*	*Bartl post*	*Thesis post*
Male Male	0.16	0.16	0.21	0.07
Male Female	-0.83	-0.83	0.13	-0.08
Female Male	-0.68	-0.68	-0.14	-0.34
Female Female	0.50	0.50	0.36	0.31
Balanced Male	0.05	0.05	0.07	-0.05
Balanced Female	-0.35	-0.35	0.20	0.07

Table 9 - Finetuning comparison to Bartl et al.[10]

As in Table 9 visible, the pre-finetuning results are the same. The results post-finetuning are not identical, notwithstanding the finetuning dataset all parameters a copied exactly. It used the GAP corpus, as described in Chapter 2.5.2. The finetuning was done with the same settings: in 3 epochs, the same seed 42, a learning rate of 0.00002, and so on. Repeating the finetuning leads to the same results, so the seed guarantees repeatability. The results overall are better than those from Bartl et al. [10]. In five out of six cases, the results are closer to zero. Only the Female Male case is worse, but it is still an improvement over the pre-finetuning values.

4.3 Which context is the best

The algorithm can work on different levels of context. There are three option paragraphs, sentence, and 2sentence, which means two consecutive sentences, which were described in Chapter 3.3. The following results were created with the simple Metadata from Appendix 8.1, this is closer described in Chapter 3.1. The bias of a sentence was determined with the word existing approach method, which is described in Chapter 3.2. The balance was created by deleting, which is described in Chapter 3.4. The gender of names was ignored, the problem of names can be found in Chapter 3.2. The reference is from the unchanged dataset, with three runs. The rest of the data in Table 10 is the average of two runs.

	paragraph	*sentence*	*2sentence*	*reference*
male male	0.15	0.16	**0.09**	0.2
male female	**0.05**	0.06	-0.06	0.07
female male	0.54	0.53	**0.5**	0.58
female female	0.76	0.76	**0.7**	0.8
balanced male	0.44	0.45	**0.41**	0.52
balanced female	0.32	0.35	**0.24**	0.38

Table 10 - Context comparison

All in all the results in Table 10 are a little improved compared to the reference. Unexpected the difference between the reference and the improved version, the percentage improvement is much higher by the male real-word biased jobs, despite there have already very small numbers. 2sentence seems to be the best option. In the following chapters often still the sentence version is used, because of the dependency context determination method, which does not work pretty well will paragraph or 2sentence. To find out if 2sentence is significantly better than the other options, further experiments are needed, they will be evaluated in Chapter 0. From a logical perspective, 2sentece should be better, because this is the context the BERT model is using.

4.4 Best way to determine the bias of the context

The algorithm has two different possibilities to determine the bias of a context, the word existing approach and the dependency approach. By the word existing approach only the existence of the words is checked, by the dependency approach the valid_depend is used, this is described in Chapter 3.3 in more detail. The following results were created with the simple Metadata from Appendix 8.1, described in Chapter 3.1. It used the sentence context because the dependency method supports only this context, more information about contexts is in Chapter 3.3. The balance was created by adding, which is described in Chapter 3.4. The gender of names was ignored, information about names can be found in Chapter 3.2. The reference is from the unchanged dataset, with three runs. The rest of the data in Table 11 is the average of two runs.

	word existing	*dependency*	*reference*
male male	**-0.01**	-0.04	0.2
male female	**-0.01**	-0.05	0.07
female male	**0.33**	0.5	0.58
female female	**0.57**	0.77	0.8
balanced male	**0.19**	0.33	0.52
balanced female	**0.09**	0.28	0.38

Table 11 - Determine the bias of context

In Table 11 a massive improvement with the word existing approach version is visible. The improvement of the dependency version is much lower. The improvement is again relatively unexpected high on already low categories, like “male male” or “balanced female”. In contrast, for “female female” the improvement is only 23%, the extreme counter-example is “male male”, they are the improvement is 95%[2].

Some reasons for the results can be found in appendices 8.4 and 8.3. The word existing approach produced massive more findings. With more findings, even if they are not always correct, it looks like the balancing, in the end, is much more valid. Why the male real-world biased jobs are so much better balanced from a percentage perspective is counterintuitive to the data. Only nine out of the twenty male jobs could be found in the dataset. On the other side, 13 female jobs could be found. The second biggest improvement on the balanced side is unified with the data because the most found jobs are from this category.

4.5 Are names important?

While determining the gender of the context, the algorithm can use the gender of names or ignore it. The following results were created with the simple Metadata from Appendix 8.1, described in Chapter 3.1. It used the sentence context, more information in Chapter 3.3, with the word existing approach method for determining the bias, which is described in Chapter 3.2. To fix the bias, paragraphs that enhance the balance are randomly added, details about this are in Chapter 3.4 and Chapter 3.5. The reference is from the unchanged dataset, with three runs. The rest of the data in Table 13 is the average of two runs.

	without names	*with names*	*reference*
male male	**-0.01**	0.06	0.2
male female	**-0.01**	-0.02	0.07
female male	**0.33**	0.5	0.58
female female	**0.57**	0.66	0.8
balanced male	**0.19**	0.43	0.52
balanced female	**0.09**	0.21	0.38

Table 12 – Impact of leveraging names to identify gender

[2] This is calculated by going from 0.2 to 0 is a difference of 0.2 and the from 0 to -0.01 it is 0.01, which was substracted from the 0.2, because 0 is the perfect sore, so the final improvent is 0.19/0.2 = 0.95

Using names to find the gender of a sentence isn’t an improvement in comparison to not using it. The method with name is still an improvement in comparison to the references. Why this leaves to worse results can have multiple reasons. It could be that the bias detection of the names is not good enough and produces too much noise, but this is unlikely because a valid method is used to determine the gender of names. The reason is probably the combination of the test, which does not include names, and that the BERT model is not combining names and pronouns.

4.6 Adding or removing Paragraphs

The algorithm has two different possibilities to fix the bias in the dataset. By adding paragraphs, picked at random from the paragraphs that are positive for the bias or removing the prioritized paragraphs, more can be found in Chapters 3.3, 3.5, and 4.2.2. The following results were created with the simple Metadata from Appendix 8.1, described in Chapter 3.1. It used the sentence context, more information about this in Chapter 3.3, with the word existing approach method for determining the bias, which is described in 3.2. The gender of names was ignored, more about names is in Chapter 3.2. The reference is from the unchanged dataset, with three runs. The rest of the data in Table 13 is the average of two runs.

	add	*remove*	*reference*
male male	**-0.01**	0.16	0.2
male female	**-0.01**	0.06	0.07
female male	**0.33**	0.53	0.58
female female	**0.57**	0.76	0.8
balanced male	**0.19**	0.45	0.52
balanced female	**0.09**	0.35	0.38

Table 13 - Add or remove paragraphs

The results in Table 13 are showing a strong preference to “add”. Like in the previous results, all values drop relative to the same amount in absolute manners. The reason can be found in Appendix 8.4 and 8.5. In particular in jobs that are heavily biased toward one direction, like secretary with removing over 1200 counts “secretary” are gone, with this a lot of information. While adding no information is lost, only some information is highlighted. This simple reason leads to better results.

4.7 Metadata – Use of Synonyms

In this chapter, the two sets of metadata are compared. These two sets are presented in Chapter 3.1, one is the simple metadata based on the BEC-Pro, and the other is the same with synonym enriched. It uses the sentence context, which is described in Chapter 3.3, with the word existing approach for determining the bias, more about this in Chapter 3.2 and removing to fix the bias, which is described in Chapter 3.4. The gender of names was ignored, information about names is in Chapter 3.2. The reference is from the unchanged dataset, with three runs. The rest of the data in Table 14 is the average of two runs.

	simple	*synonym*	*reference*
male male	0.16	**0.02**	0.2
male female	-0.06	**0.00**	0.07
female male	**0.53**	0.56	0.58
female female	**0.76**	**0.76**	0.8
balanced male	0.45	**0.44**	0.52
balanced female	0.35	**0.28**	0.38

Table 14 - Use of synonyms

These results are ambiguous. The male real-world biased jobs have a good improvement. The female and balanced jobs on the other side have no or a very low improvement. One possible reason for this could be the different amount of synonyms in the metadata for each group. This is not matching, because the balanced group got 15 words added, the male group 19, and the female group 34. So a clear reason for this can be not found. The only observation, that can be done is that the male real-world biased jobs are reacting much better than the rest of the jobs. According to the less computation power and effort needed for simple metadata, this can be preferred.

4.8 Metadata – Which job name should be included?

In this evaluation always the 60 jobs from Bartl et al. [10] are used. Some of these jobs are specific, like “registered nurse”. Because “registered nurse” exists only 20 times, it was replaced in the metadata with “nurse”, which exists 2097 times. For the evaluation of this, the same method as in Chapter 4.1 is used. The context is “sentence” (Chapter 3.3), and the fix method is “adding” (Chapter 3.4). For the metadata, the “simple metadata” from Appendix 8.1 is used, described in Chapter

3.1. Then in the metadata, “nurse” was replaced with “registered nurse”. Then in the evaluation dataset BEC-PRO “nurse” was added. The following results are only the average of sentences with “nurse”/”registered nurse”, according to gender.

	“nurse”-Evaluation		“registered nurse”-Evaluation		average of female real-world biased	
	male	female	male	female	male	female
Uncleaned-reference	1.16	1.76	0.78	1.51	0.57	0.82
Metadata with “registered nurse”	0.90	**1.48**	0.75	**1.34**	0.48	**0.59**
Metadata with “nurse”	**0.63**	1.63	**0.51**	1.47	**0.34**	0.61

Table 15 - Evaluation metadata nurse

The first major observation is, that nurse is an outstanding strongly biased job in comparison to the average of female real-world biased jobs. This does not change after the dataset was cleaned with the approaches of this thesis to fix the bias. The interesting point is that the male value for both “nurse” and “registered nurse“ is better when the data is cleaned with „nurse“ and the female value is better when the metadata is cleaned with „registered nurse“. The difference in the male values is even bigger than the difference for the female. The improvement of the male values is logical because „nurse“ is a word that exists more often in the dataset, than „registered nurse“. Why with less data the female value is lower, the reason can only be that with more data, more error appears. This means when data is cleaned with „nurse“, maybe a lot of paragraphs got identified as male, but they are female. Then these sentences are added and result in this worse result. But overall the value of „nurse“ are better, this using a more generic job title is better.

4.9 Handle jobs with strong bias with the ratio

Some jobs have really bad and extreme numbers in comparison to the average, as visible in Table 15. These values are not improved when the normal ratio is applied. For this reason, in the next example, a 1:10 ratio is applied to maybe fix the bias with this. This ratio was picked, to reach are more extreme ratio than the data has on its own. That means male paragraphs get added until there are ten times more male than female paragraphs. The data is created with simple metadata (described in Chapter 3.1), sentences (described in Chapter 3.3), adding (described in Chapter

3.4), word existing approach (described in Chapter 3.2), and no names (described in Chapter 3.2).

	1:10 cleaned	*balanced*	*reference*
male male	0.08	**-0.01**	0.2
male female	-0.03	**-0.01**	0.07
female male	0.44	**0.33**	0.58
female female	**0.51**	0.57	0.8
balanced male	0.34	**0.19**	0.52
balanced female	0.15	**0.09**	0.38

Table 16 – 1:10 values overview

The overall result is worse than the result of the balanced ratio, as visible in Table 16, but the bias move, as expected toward the male side. For male and balanced real-world biased jobs, the male is more dominant than in the balanced variant. The male bias did not move more toward the male as in the reference. On the other side, in the balanced real-world biased jobs, it percentage difference is higher the in the uncleaned reference. For female real-world jobs, females are still ahead of men, but the distance is smaller.

	"registered nurse"-Evaluation	
	male	female
reference	0.82	1.58
balanced	**0.52**	1.48
1:10 cleaned	0.77	**1.19**

Table 17 – 1:10 values of "nurse"

For "registered nurse" the bias is extremely high in comparison to the average, as discovered in Chapter 4.8. All in all the results in Table 17 are comparable to the results from Table 16, but the bias toward females dropped below the biased from the balanced result. This is a real improvement, but still there strong biased.

Selecting a strong ratio does seem not to work out. In Chapter 4.2.2 the accuracy for a reference is shown, this can be done for males and females separately. The accuracy for males is 44% and for females only 32%. For the nurse, after the balancing algorithm has applied the 1:10 ratio, 3724 male occurrences appear and 376 females. If now the accuracy is applied to these values, from the 3724 male occurrences are 2085 females, and from the 376 females are 255 males. This results in 1894 male and 2206 female occurrences, which is close to a 1:1 ratio and not to a 1:10 ratio. The accuracy from Chapter 4.2.2 does not contain numbers for nurse and are only an average. Nevertheless, a ratio of 1:10 was a step in the right direction, that's why in the next example a 1:100 ratio was used.

	1:100 cleaned	*balanced*	*reference*
male male	0.23	**-0.01**	0.2
male female	-0.13	**-0.01**	0.07
female male	0.83	**0.33**	0.58
female female	0.64	**0.57**	0.8
balanced male	0.58	**0.19**	0.52
balanced female	0.19	**0.09**	0.38

Table 18 – 1:10 values overview

	"registered nurse"-Evaluation	
	male	female
reference	0.82	1.58
balanced	**0.52**	1.48
1:100 cleaned	1.06	**0.87**

Table 19 – 1:100 values of "nurse"

As in Table 18 and Table 19 visible, with a 1:100 ratio the balanced flipped. In all categories the male is more dominant than the female, this comes with the cost of overall higher values.

This approach is fixing the balance between females and males (the values a getting closer to each other) but both values are higher than with a balanced ratio. This approach finally results in stronger bias for males and females. All in all seems an unbalanced dataset leads to bad results, if it's unprocessed or processed toward one gender the result is the same, and the bias grows.

4.10 Why are a lot of the values positive

While evaluating the data, all in all, most of the results of Chapter 4 are positive. In contrast, the results from Bartl et al. [10] have negative values in Chapter 4.2.3. The results from the standard pretrained BERT model are negative and positive too. The reason for this is the extremely high loss of training in this thesis. The evaluation loss in all runs is over 2.5, this means the model is underfitted. The results from this thesis are getting negative if the target and the attribute are masked, and this has a higher chance to appear in comparison to only the target is masked, compare to Chapter 4.1. This is unlikely to happen if the model is underfitted. This is the reason, why the model often predicts the attribute with a higher chance when the attribute is not masked. The meaningfulness of the results in this thesis has only a little impact because the values exist per gender. If the value of a male attribute is high then this only means, the job is more preferred for male attributes, than for other attributes. If the value of the male attribute is negative, then there are other attributes, that are more preferred for example female attributes. If the female attribute is preferred this is then visible in the female result.

4.11 Why are male jobs small and debias so well

In nearly all results the real-world male-biased jobs have a very small bias, additional this bias often improves relative extreme in contrast to the other jobs. When analyzing the data from Appendix 8.4, the male jobs have the least appearances in the dataset, with only 1301 appearances. The female jobs have 3072 and the balanced jobs 1985 appearances. This low number of s could lead to less effective training for these jobs and could result in less extreme numbers. Together with the balanced jobs, the male jobs have the same ratio, $1/_4$ female appearances and $3/_4$ male appearances on the other side female jobs are less extreme ratio with $1/_3$ female and $2/_3$ male appearances. As a result of this more extreme split in the male and balanced jobs, the algorithm adds more jobs, if the target ratio is 50:50. The results overall represent this, the bias is going down much more by male and balanced jobs. The combination of these two facts could lead to the small and very well-improved values of real-world male-biased jobs.

4.12 Interference between different categories

In Chapter 3.4 the problem of interference between different categories (jobs) was found. This problem appears in a theoretical context. A paragraph can include multiple jobs and when data should be added or removed because of one job, this could influence the other job. The problem is even worse when the implantation is added. The algorithm is working in two steps. First, it scans the dataset for the statics, for better performance, in this single run all jobs are done. In the second step with this produced statistic, for each job, the elements get added or removed. This is done successively for each job. The only interaction between these jobs is, if the paragraph is already removed or added, it will not be changed again.

	Paragraph	*Bias firefighter*	*Bias nurse*		*Job[3]*	*Action*	*Sum firefighter*	*Sum nurse*
Data	1102	5,2	4,1	Actions	firefighter	-1105	8,6	7,8
	1103	0,2	0,2			-1106	7,6	7,8
	1104	1,2	3,1			-1107	6,6	7,8
	1105	2,0	2,1		nurse	-1103	6,6	7,6
	1106	1,0	0,2			-1105	6,6	5,5
	1107	1,0	0,0					

Table 20 - Example interference between categories

In Table 20 an example dataset statistic is presented on the data side. On the action side potential removal process, with the results, the algorithm would calculate. The real results are different, for “firefighters” it would be: 6,4 for “nurse” it is: 7,2. While the algorithm fixes firefighters, the “nurse” is not taken into account. After this, “nurse” is fixed, based on the evaluation from the beginning. Removing -1103 works normally, while removing 1105, the algorithm registers that it’s already removed, and only changes the score.

It can be that a job that is processed at the end changes paragraphs from another earlier processed job and the algorithm will not recognize this. To take a look if this happens while using the algorithm, the data from Appendix 8.4 and 8.5 are evaluated. The after values are collected with a second run of the statistic generation part of the algorithm.

[3] The job, the algorithm tries to fix in this phase.

job	*male adding*	*female adding*	*% difference adding*	*male removing*	*female removing*	*% difference removing*
kindergarten teacher	3	3	1	6	6	1
dental hygienist	1	1	1	3	2	**1.5**
nurse	241	231	1.04	382	375	1.02
dental assistant	1	1	1	2	1	2
secretary	416	435	0.96	1722	1708	1.01
medical assistant	1	1	1	3	3	1
hairdresser	23	20	**1.15**	26	26	1
dietitian	2	2	1	3	3	1
paralegal	2	2	1	5	5	1
receptionist	28	27	1.04	49	49	1
housekeeper	72	69	1.04	78	78	1
registered nurse	3	3	1	5	4	**1.25**
bookkeeper	16	15	**1.07**	26	27	0.96
health aide	0	0	0	0	0	0
taper	0	0	0	0	2	0
steel worker	1	1	1	1	2	0.5
roofer	0	0	0	0	7	0
electrician	9	8	**1.13**	40	40	1
conductor	80	83	0.96	281	281	1
plumber	12	12	1	39	39	1
carpenter	99	104	0.95	255	248	1.03
mason	97	102	0.95	365	363	1.01
firefighter	6	5	**1.2**	19	19	1

salesperson	0	2	0	5	5	1
crossing guard	0	0	0	0	3	0
photographer	115	121	0.95	272	272	1
lifeguard	8	8	1	20	20	1
sales agent	0	0	0	1	0	0
insurance underwriter	0	0	0	0	0	0
medical scientist	0	0	0	0	1	0
statistician	1	1	1	8	8	1
judge	310	328	0.95	1151	1151	1
bartender	15	15	1	61	59	1.03
dispatcher	1	1	1	1	10	**0.1**
order clerk	0	0	0	0	0	0
mail sorter	0	0	0	0	0	0

Table 21 - Interference between different categories after the algorithm balanced them

As in Table 21 visible, nearly all jobs are in balance if the threshold from 0.95 is taken into account. Only four jobs by adding and three jobs by removing are not hitting the ratio. For five of these seven problematic jobs, the reason is clear, the numbers are simply so small that just a difference of one is resulting in not hitting the threshold. The other two jobs have very small numbers too. These are the visible and expected interference between the jobs. There can be more, but the other interferences are so small that they are not relevant. All in all, it is the right decision to process the as performant as possible and ignore the interference. With the threshold and the huge data in comparison to the founded jobs, this is not problematic.

4.13 Influence of Finetuning

In this chapter, the best result of the previous is compared to the finetuning method from Bartl. et al. [10]. Most of the possible strategies are tested in the previous chapters. The best method was simple metadata, the sentence context with the word existing approach for determining the bias and adding to fix the bias. The gender of names was ignored. Direct comparisons, this is the best method in chapters, besides the context comparison. In Chapter 4.3 the two-consecutive sentences approach is better than the sentence approach. This is the reason why in Table 22 these two contexts are again compared against each other, with the rest of the parameters set two the best values (simple metadata, adding, word existing approach, no names)

	add s	*add 2s*	*reference*
male male	**-0.01**	0.21	0.2
male female	**-0.01**	0.15	0.07
female male	**0.33**	0.53	0.58
female female	**0.57**	0.78	0.8
balanced male	**0.19**	0.44	0.52
balanced female	**0.09**	0.36	0.38

Table 22 – Comparison of best methods, add s stands for adding with sentence context, add 2s is adding with 2 consecutive sentences

Regarding Table 22, the best method for the context is the sentence approach. In Table 23 finetuning is applied to the reference and the best approach from the data-cleaning algorithm (simple metadata (described in Chapter 3.1), sentence (described in Chapter 3.3), adding (described in Chapter 3.4), word existing approach (described in Chapter 3.2), no names (described in Chapter 3.2)).

	reference	*reference + finetuning*	*add*	*add + finetuning*
male male	0.20	0.19	**-0.01**	-0.05
male female	0.07	0.17	**-0.01**	-0.07
female male	0.58	0.30	0.33	**0.12**
female female	0.80	0.47	0.57	**0.21**
balanced male	0.52	0.36	0.19	**0.11**
balanced female	0.38	0.33	0.09	**0.02**
absolute average	0.43	0.30	0.20	**0.10**

Table 23 - Influence of finetuning

Table 23 shows that the combination of finetuning and cleaning the dataset is the best method. The male real-world biased jobs are cleaned to a level, where it is just noise. The jobs are reaching very good levels too. Another clear point is that cleaning the dataset is more effective than just finetuning. The reason is the cleaned amount of data. For finetuning only a small dataset with 4453 sentences is used, while the algorithm from this thesis cleans the whole dataset with around 5,000,000 sentences. The combination is working because these two methods are working on different points of the model training process. As result, they can be stacked and enhanced each other. Additionally, they work with different data, so they do not overlap each other with the cleaning input.

5 Related Works

In this part, I will discuss and position my work in contrast to other papers.

Large language models, with a lot of tokens and trained on a very big dataset can be effective, as Brants et al. [22] propose. In this thesis, smaller models, with still big datasets are used. These models should use real-world training data, as Aye et al. [7] have shown with the example of autocompletion, which is much better when trained on real-world data, this is used in this thesis too. But when real-world data is used on large scale, a big bias comes with this, as Papakyriakopoulos et al. [63] have shown for the example of word embedding. Garg et al. [35] have shown that gender and ethnic stereotypes are a big problem in word embeddings. He could compare the bias in the models with bias in the real world of the last 100 years. Caliskan et al. [24] have shown these problems not exits only in the word embedding, but also in the whole model. They could accurately find all the historic biases in the models. Mandis [52] has shown that the bias is even stronger in the model than in the data, he trained a network with 1.000 random Wikipedia articles and found that the gender and race bias more than double up in the model in comparison to the data. Mandelbaum et al. [51] discussed in 1949 that it is important to understand that bias can affect the psychological status of different groups. That is why bias is a problem in machine learning and in this thesis, an approach is shown to fix this problem.

Carlini et al. [26] have taken care of another risk. With these large models, it is possible to extract the data, where the model was trained, even names and related information are possible. This problem is treated in this thesis. Argrawal et al. [1] work on a future database system to provide more easy access to data for training machine learning models. This should improve the quality of these and reduce bias. This is not used here, because the target is to work with real-world data.

For the proper ratio of bias, or fairness, according to Binns et al. [15], there is group and individual fairness. These are different perspectives that can be used on the ratio of bias. Goodman et al. [39] agree with this. Cowgill et al. [19] proposed another fairness definition: the results should be compared to a counterfactual ideal case, and then it should be decided if the algorithm is biased. Donini et al. [31] researched the algorithmic fairness of machine learning algorithmic. Neutatz et al. [34], have shown some solutions to these problems, but only for models with specific tasks. In this work, other solutions are needed cause, it is about models that can be used for multiple problems.

Pedreshi et al. [64] discussed the discrimination in datasets about different topics, for example, credit scores. They postulated that simply removing the discrimination attributes isn't a solution, that's why manipulating the model for less bias isn't appropriate and in this thesis, the data is modified. Brunet et al. [23] backtraced the bias from the word embedding to the dataset and removed the bias parts. He was very successful on Wikipedia and the New York times copras. In this paper, this should be done only based on the dataset, without including the model in the debias process. For fixing the bias problem Bolukbasi et al. [20] have shown how to reduce bias in word embedding after training. He used metrics to balance the word embedding. Zhao, Wang et al. [88] extracted the bias to extra dimensions while training so can be used whenever needed or ignored if an unbiased model is needed. But Gonen et al. [39] discussed that these approaches are not working properly, because there are still groups of words that are close together, and so are still biased, this is the reason why another debias approach is needed.

In this thesis, the BERT model from Devlin et al. [29] is used. Bartl et al. [10] provide the test dataset BEC-Pro, based on Maudslay et al. [54]. The finetuning dataset is taken from Webster et al. [85]. It is a former test dataset based on real-life data from Wikipedia. The genders were swapped, so training with the dataset develop a more bias-free model. Bartl et al. used the dataset for the first time for finetuning. The bias measuring method from Bartl et al. [10], which is also used in this thesis, is a very practical approach, based on the work of Kurita et al. [50]. Zhaom, Zhou et al. [89]. Nangia et al. [58] provide other bias benchmarks. This is based on the challenge dataset CrowS-Pairs, which they introduce. It can evaluate nine types of bias. For keeping the test in this thesis towards one bias, this dataset was not used.

For identifying the gender of a sentence named entity recognition is used. The base for this was created by Black et al. [17]. He used a rule-based system without any learning. The next steps with an enhanced technique (machine learning) were done by Kapetatnios et al. [48]. Nasiboglu et al. [59] did a comparison of spaCy and the method from Standford. This was the reason for selecting spaCy which is used in this thesis. Another needed technique is the determination of the gender of names. This was discussed by Karimi et al. [49] with some initial methods. Carsenat et al. [32] proposed methods that could work better, but also explained that methods based on real data are better. A method based on real data then is used in this thesis. To understand relations between words in a sentence, part of speech tagging is used. Schmid et al. [70] presented the first neural network-based approach and compared it to two non-neural network approaches. Jin et al. [44]. Compared different modern tools for part of speech tagging. He criticized their unreflected use because these systems are by far not perfect. Based on his paper, spaCy was picked.

6 Conclusion

In this final part of the thesis, I will summarize the results and then present the next logical steps and further ideas.

6.1 Summary

The target was to reduce the bias in a machine learning model, by fixing the bias in the dataset. All in all the results are promising, with the provided example the bias could be improved by 77%. This result could be reached with one of the simplest configurations, described in this thesis. By just checking the existence of the provided keywords in a sentence and adding the paragraphs, in a way the bias is reduced. After training finetuning was added. In absolute numbers, the bias was 0.43 without any cleaning. This means that simplified[4] it is about 50% more likely to get a specific job when you provide the pronoun than to get the same job without providing a pronoun. With finetuning, the reference method, the bias goes down to 0.3, with only data-cleaning, it went down to 0.2, and with the combination of data-cleaning and finetuning it's only 0.1, which is an improvement of 77%. With the best-cleaned method, it's about a 10% higher chance to get a specific job with the given pronoun. The more complex solutions, like taking the names into account or using the relation of the words to each other, did not provide any improvement, they made the results worse and extend the computation time.

The result is good but created with a small dataset. This results in unstable training with the BERT model because the model can not be trained to the point where it reliably works. This could lead to an under or overestimation of the effect of the dataset balancing because the impact on the result by the size of the dataset is unknown.

6.2 Future work

This work has a lot of room for future work. First, the algorithm should be tested on a big enough dataset, to train a valid model. Then it should be tested with different models. These tests could provide information about the universal usage of the algorithm and if could perform in the real world as well as in theory. The next task could be further research according to extreme jobs. Maybe it is possible to predict the ratio that hits the balance perfectly and reduces the balance. Then use a per-

[4] This average is a absolut average. This means for some genders with some real-world biased job the change can be negativ. This would mean it's 50% less likely to get a specific job when the pronoun is provided in comparison to a sentence without the pronoun.

job ratio to balance each better job better. In the next steps, it could be useful to check the model with more tests if the bias is reduced. It could be also useful to check if adding only the sentence instead of the whole paragraph is an option too. While these tests the general performance of the model should be monitored. The next step could be the extension of the metadata, first to cover more jobs, and later to cover more topics.

I have shown in this thesis, that balanced dataset results in a balanced model. Instead of fixing the bias in the dataset with selective changes, maybe it is easier to just multiply each sentence by all genders (ethics, religions). This could result in different improvements:

- More data, for a better training performance of the models.
- Possibility to introduce genders into jobs with really low data or the possibility to cover all genders/ethics groups even if there is no data available.
- Theoretical absolute balanced models.
- This could lead to incorrect information or “fake news”, the sentence “Women give birth” would result in “Men give birth”, which is clearly wrong, but the model would learn it.

7 References

[1] Agrawal, A., Chatterjee, R., Curino, C., Floratou, A., Godwal, N., Interlandi, M., Jindal, A., Karanasos, K., Krishnan, S., Kroth, B., Leeka, J., Park, K., Patel, H., Poppe, O., Psallidas, F., Ramakrishnan, R., Roy, A., Saur, K., Sen, R., Weimer, M., Wright, T., and Zhu, Y. 2020. Cloudy with high chance of DBMS: a 10-year prediction for Enterprise-Grade ML. In *10th Conference on Innovative Data Systems Research, CIDR 2020, Amsterdam, The Netherlands, January 12-15, 2020, Online Proceedings*. www.cidrdb.org.

[2] Ahn, J. and Oh, A. 2021. Mitigating Language-Dependent Ethnic Bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 533–549. DOI=10.18653/v1/2021.emnlp-main.42.

[3] Akbik, A., Blythe, D., and Vollgraf, R. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics, 1638–1649.

[4] Andrea.zawacki. 2012. *Male Nurses vs Female Nurses: What's the Difference [Infographic]*. https://carrington.edu/blog/male-vs-female-nurse/. Accessed 13 August 2022.

[5] Anna C. McFadden, George E. Marsh, Barrie Jo Price, and Yunhan Hwang. 1992. A STUDY OF RACE AND GENDER BIAS IN THE PUNISHMENT OF SCHOOL CHILDREN. *Education and Treatment of Children* 15, 2, 140–146.

[6] arnabs007. 2020. Pretrain a BERT language model from scratch. *Kaggle* (Jul. 2020).

[7] Aye, G. A., Kim, S., and Li, H. 2021. Learning Autocompletion from Real-World Datasets. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice*. *ICSE-SEIP 2021 : proceedings : virtual (originally Madrid, Spain), 25-28 May 2021*. IEEE Computer Society, Conference Publishing Services, Los Alamitos, California, 131–139. DOI=10.1109/ICSE-SEIP52600.2021.00022.

[8] Back, S. E., Payne, R. L., Simpson, A. N., and Brady, K. T. 2010. Gender and prescription opioids: findings from the National Survey on Drug Use and Health. *Addictive behaviors* 35, 11, 1001–1007.

[9] Bagdasaryan, E. and Shmatikov, V. 2022. Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures. In *43rd IEEE*

Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022. IEEE, 769–786. DOI=10.1109/SP46214.2022.9833572.

[10] Bartl, M., Nissim, M., and Gatt, A. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT’s Gender Bias. *CoRR* abs/2010.14534.

[11] 2022. *BERT Explained: State of the art language model for NLP*. Accessed 9 May 2022.

[12] 2022. *bias*. https://dictionary.cambridge.org/de/worterbuch/englisch/bias. Accessed 4 August 2022.

[13] 2022. *bias_1 noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com*. https://www.oxfordlearnersdictionaries.com/definition/english/bias_1?q=bias. Accessed 4 August 2022.

[14] Binder, B. 2017. (Europäische) Ethnologie: reflexive Ethnografien zu Geschlecht und Geschlechterverhältnissen. In *Handbuch Interdisziplinäre Geschlechterforschung*, B. Kortendiek, B. Riegraf and K. Sabisch, Eds. Geschlecht und Gesellschaft 65. Springer Fachmedien Wiesbaden, Wiesbaden, 1–9. DOI=10.1007/978-3-658-12500-4_120-1.

[15] Binns, R. 2020. On the Apparent Conflict Between Individual and Group Fairness. In *FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. ACM, 514–524. DOI=10.1145/3351095.3372864.

[16] Bird, S. 2006. NLTK: The Natural Language Toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics. DOI=10.3115/1225403.1225421.

[17] Black, W. J., Rinaldi, F., and Mowatt, D. 1998. FACILE: Description of the NE System Used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

[18] Blaicher, G. 1987. *Erstarrtes Denken*. *Studien zu Klischee, Stereotyp und Vorurteil in englisch-sprachiger Literatur*. Narr, Tübingen.

[19] Bo Cowgill, C. T. 2017. *Algorithmic Bias: A Counterfactual Perspective.*

[20] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 4349–4357.

[21] Bowling, M., Fürnkranz, J., Graepel, T., and Musick, R. 2006. Machine learning and games. *Mach Learn* 63, 3, 211–215.

[22] Brants, Thorsten, Popat, Ashok C., Xu, Peng, Och, Franz J., Dean, and Jeffrey. 2007. Large Language Models in Machine Translation. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*. ACL, 858–867.

[23] Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. 2019. Understanding the Origins of Bias in Word Embeddings. In *Proceedings of the 36th International Conference on Machine Learning*. Proceedings of Machine Learning Research. PMLR, 803–811.

[24] Caliskan, A., Bryson, J. J., and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.)* 356, 6334, 183–186.

[25] Cambridge Dictionary. 2022. *Nouns and gender*. https://dictionary.cambridge.org/de/grammatik/britisch-grammatik/nouns-and-gender. Accessed 13 August 2022.

[26] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*. USENIX Association, 2633–2650.

[27] Cukr, M. 2018. POS tags. *Lexical Computing CZ s.r.o.* (Mar. 2018).

[28] 2022. *Data Monetization – Die Grundsätze von Tesla für datengestützten Erfolg*. https://blog.usu.com/de-de/die-grundsaetze-von-tesla-fuer-datengestuetzten-erfolg. Accessed 14 August 2022.

[29] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 4171–4186. DOI=10.18653/v1/n19-1423.

[30] 2022. *Dictionary by Merriam-Webster: America's most-trusted online dictionary*. https://www.merriam-webster.com/. Accessed 22 September 2022.

[31] Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. 2018. Empirical Risk Minimization under Fairness Constraints. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2796–2806.

[32] Elian Carsenat. 2019. *Inferring gender from names in any region, language, or alphabet. DOI*=10.13140/RG.2.2.11516.90247.

[33] Emmert-Streib, F. and Dehmer, M. 2019. A Machine Learning Perspective on Personalized Medicine: An Automized, Comprehensive Knowledge Base with Ontology for Pattern Recognition. *Mach. Learn. Knowl. Extr.* 1, 1, 149–156.

[34] Felix Neutatz, Z. A. 2022. What is ”Good” Training Data? - Data Quality Dimensions that Matter for Machine Learning.

[35] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA* 115, 16, E3635-E3644.

[36] Gesellschaft für Sozialwissenschaftliche Frauenforschung e.V. Wissensnetz Gender Mainstreaming für die Bundesverwaltung.

[37] GitHub. 2022. *GitHub - ecmonsen/gendered_words: Dictionary of English words tagged with their natural gender*. https://github.com/ecmonsen/gendered_words. Accessed 27 August 2022.

[38] 2022. *GLUE Benchmark*. https://gluebenchmark.com/leaderboard. Accessed 6 May 2022.

[39] Gonen, H. and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Workshop on Widening NLP@ACL 2019, Florence, Italy, July 28, 2019*. Association for Computational Linguistics, 60–63.

[40] Gulati, S., Sousa, S., and Lamas, D. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology* 38, 10, 1004–1015.

[41] Harvard Business Review. 2016. *Why Do So Few Women Edit Wikipedia?* https://hbr.org/2016/06/why-do-so-few-women-edit-wikipedia. Accessed 10 April 2022.

[42] 2022. *Hugging Face – The AI community building the future*. https://huggingface.co/. Accessed 25 August 2022.

[43] Hugging Face Forums. 2021. *Fine-tuning BERT Model on domain specific language and for classification - 🤗Transformers - Hugging Face Forums*. https://discuss.huggingface.co/t/fine-tuning-bert-model-on-domain-specific-language-and-for-classification/3106. Accessed 16 September 2022.

[44] Jin, S., Chen, S., and Xie, X. 2021. Property-based Test for Part-of-Speech Tagging Tool. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 1306–1311. DOI=10.1109/ASE51524.2021.9678807.

[45] John E Lincoln. 2020. *Google Click-Through Rates (CTR) By Ranking Position [2020]*. https://ignitevisibility.com/google-ctr-by-ranking-position/. Accessed 4 August 2022.

[46] Jonas, K., Stroebe, W., and Hewstone, M. 2014. *Sozialpsychologie*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[47] Jutta Kühl. Geschlechtsbezogener Verzerrungseffekt (Gender Bias).

[48] Kapetanios, E., Tatar, D., and Sacarea, C. 2014. *Natural language processing*. *Semantic aspects*. CRC Press, Boca Raton, Florida.

[49] Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., and Strohmaier, M. 2016. Inferring Gender from Names on the Web. In *WWW'16 companion*. *Proceedings of the 25th International Conference on World Wide Web : May 11-15, 2016, Montreal, Canada*. International World Wide Web Conferences Steering Committee; ACM, Republic and Canton of Geneva, New York, NY, 53–54. DOI=10.1145/2872518.2889385.

[50] Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. *CoRR* abs/1906.07337.

[51] Mandelbaum, D. G. and Sapir, E., Eds. 1949. *Selected writings of Edward Sapir in language, culture and personality*. Univ. of Calif. Pr, Berkeley [u.a.].

[52] Mandis, I. S. 2021. Reducing Racial and Gender Bias in Machine Learning and Natural Language Processing Tasks Using a GAN Approach. *IJHSR* 3, 6, 17–24.

[53] Marsh, E. and Perzanowski, D. 1998. MUC-7 Evaluation of IE Technology: Overview of Results. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

[54] Maudslay, R. H., Gonen, H., Cotterell, R., and Teufel, S. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 5266–5274. DOI=10.18653/v1/D19-1530.

[55] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, 1–35.

[56] Merity, S., Xiong, C., Bradbury, J., and Socher, R. 2017. Pointer Sentinel Mixture Models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

[57] moleculenet. 2021. *Datasets*. https://moleculenet.org/datasets-1. Accessed 13 August 2022.

[58] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. 2020. *CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models*.

[59] Nasiboglu, R. and Gencer, M. 2021. COMPARISON OF SPACY AND STANFORD LIBRARIES’PRE-TRAINED DEEP LEARNING MODELS FOR NAMED ENTITY RECOGNITION. *Journal of Modern Technology and Engineering* 6, 2, 104–111.

[60] Nayak, P. 2019. Understanding searches better than ever before. *Google* (Oct. 2019).

[61] 2022. Nearly 400 car crashes in 11 months involved automated tech, companies tell regulators. *NPR* (Jun. 2022).

[62] Nicoll, S., Douglas, K., and Brinton, C. 2022. Giving Feedback on Feedback: An Assessment of Grader Feedback Construction on Student Performance. In *LAK22 conference proceedings*. *The Twelfth International Conference on Learning Analytics & Knowledge : learning analytics for transition, disruption and social change : March 21-25, 2022, Online, Everywhere*. ICPS. The Association for Computing Machinery, New York, New York, 239–249. DOI=10.1145/3506860.3506897.

[63] Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., and Marco, F. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 446–457. DOI=10.1145/3351095.3372843.

[64] Pedreshi, D., Ruggieri, S., and Turini, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACMKDD International Conference on Knowledge Discovery & Data Mining*. *Las Vegas, NV, USA, August 24-27, 2008*. ACM Press, New York, NY, 560. DOI=10.1145/1401890.1401959.

[65] Philippe Remy. 2021. *Name Dataset*. GitHub. *GitHub repository*.

[66] Posch, C. 2011. Mitgefangen – Mitgehangen? Generisches Maskulinum und Normen geschlechtergerechten Sprachgebrauchs.

[67] RoadSafetyFacts.eu. 2019. *How can automated and connected cars improve road safety? | RoadSafetyFacts.eu*. https://roadsafetyfacts.eu/how-can-automated-and-connected-vehicles-improve-road-safety/. Accessed 14 August 2022.

[68] SANDRA G . MAYSON. Bias In, Bias Out. *THE YALE LAW JOURNAL*.

[69] Schelter, S., He, Y., Khilnani, J., and Stoyanovich, J. 2020. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. In *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT 2020, Copenhagen, Denmark, March 30 - April 02, 2020*. OpenProceedings.org, 395–398. DOI=10.5441/002/edbt.2020.41.

[70] Schmid, H. 1994. Part-of-Speech Tagging with Neural Networks. In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, 172–176.

[71] Selzer, A. 2009. A beginner's guide to language and gender, by Allyson Jule. *Gender and Education* 21, 3, 343–344.

[72] Silveira, J. 1980. Generic masculine words and thinking. *Women’s Studies International Quarterly* 3, 2-3, 165–178.

[73] 2022. *spaCy · Industrial-strength Natural Language Processing in Python*. https://spacy.io/. Accessed 22 September 2022.

[74] Statista. 2022. *Marktanteile der Suchmaschinen - Mobil und stationär 2022 | Statista*. https://de.statista.com/statistik/daten/studie/222849/umfrage/marktanteile-der-suchmaschinen-weltweit/. Accessed 4 August 2022.

[75] Stoyanovich, J., Abiteboul, S., Howe, B., Jagadish, H. V., and Schelter, S. 2022. Responsible data management. *Commun. ACM* 65, 6, 64–74.

[76] Tesser, M. 2018. Males VS females - Is there gender equality in the Fire Service? *Emergency Live* (Jan. 2018).

[77] The Official Microsoft Blog. 2016. *Learning from Tay’s introduction - The Official Microsoft Blog*. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/. Accessed 18 August 2022.

[78] 2022. *Thesaurus | Check Thesaurus Online for Free*. https://www.thesaurus.net/. Accessed 22 September 2022.

[79] Uhrig, S. 2021. Darum haben wir alle Vorurteile. *Quarks* (Apr. 2021).

[80] 2022. *Universal Dependency Relations*. https://universaldependencies.org/u/dep/. Accessed 9 May 2022.

[81] 2022. *Utilities for Trainer*. https://huggingface.co/docs/transformers/internal/trainer_utils. Accessed 25 August 2022.

[82] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

[83] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for

Computational Linguistics, Stroudsburg, PA, USA, 353–355. DOI=10.18653/v1/W18-5446.

[84] 2008. *Web Search*. Springer, Berlin, Heidelberg.

[85] Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. 2018. *Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns*.

[86] Wikipedia. 2022. *List of datasets for machine-learning research*. https://en.wikipedia.org/w/index.php?title=List_of_datasets_for_machine-learning_research&oldid=1100503668. Accessed 13 August 2022.

[87] Wu, Y., Schuster, M., Chen, Z., Le V, Q., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144.

[88] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of*. Association for Computational Linguistics, Stroudsburg, PA, USA, 15–20. DOI=10.18653/v1/N18-2003.

[89] Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Association for Computational Linguistics, 4847–4853.

[90] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *The IEEE International Conference on Computer Vision (ICCV)*.

8 Appendix

8.1 Simple metadata

```
{
"categorywords":[ ["kindergarten teacher","",""]
,["dental hygienist","",""]
,["nurse","",""]
,["speech-language pathologist","",""]
,["dental assistant","",""]
,["childcare worker","",""]
,["medical records technician","",""]
,["secretary","",""]
,["medical assistant","",""]
,["hairdresser","",""]
,["dietitian","",""]
,["vocational nurse","",""]
,["teacher assistant","",""]
,["paralegal","",""]
,["billing clerk","",""]
,["phlebotomist","",""]
,["receptionist","",""]
,["housekeeper","",""]
,["registered nurse","",""]
,["bookkeeper","",""]
,["health aide","",""]
,["taper","",""]
,["steel worker","",""]
,["mobile equipment mechanic","",""]
,["bus mechanic","",""]
,["service technician","",""]
,["heating mechanic","",""]
```

```
,["electrical installer","",""]
,["operating engineer","",""]
,["logging worker","",""]
,["floor installer","",""]
,["roofer","",""]
,["mining machine operator","",""]
,["electrician","",""]
,["repairer","",""]
,["conductor","",""]
,["plumber","",""]
,["carpenter","",""]
,["security system installer","",""]
,["mason","",""]
,["firefighter","",""]
,["salesperson","",""]
,["director of religious activities","",""]
,["crossing guard","",""]
,["photographer","",""]
,["lifeguard","",""]
,["lodging manager","",""]
,["healthcare practitioner","",""]
,["sales agent","",""]
,["mail clerk","",""]
,["electrical assembler","",""]
,["insurance sales agent","",""]
,["insurance underwriter","",""]
,["medical scientist","",""]
,["statistician","",""]
,["training specialist","",""]
,["judge","",""]
,["bartender","",""]
,["dispatcher","",""]
```

```
,["order clerk","",""]
,["mail sorter","",""]],
"categoryidentifier" :  [
                        ["he", "man", "brother", "son", "husband", "boyfriend", "father",
"uncle", "dad"],
                        ["she", "woman", "sister", "daughter", "wife", "girlfriend",
"mother","aunt", "mom"]
                      ],
"categoryname" : ["male","female"],
"allowed_depend" : [
["_compound", "appos"]
,["appos"]
,["_compound", "compound"]
,["_compound", "_pobj", "_prep"]
,["_pobj", "_prep"]
,["compound"]
,["_nsubj", "attr"]
,["_nsubj", "prep", "pobj"]
,["_compound", "_nsubj", "nsubj"]
,["_nsubj", "nsubj"]
,["_compound", "relcl", "attr"]
,["_compound", "conj"]
,["_nsubj", "prep", "pobj", "conj"]
,["_compound", "_appos"]
,["_appos"]
,["_compound", "appos", "compound"]
,["_compound", "nmod"]
,["nmod"]
,["_nsubjpass", "prep", "pobj"]
,["_poss", "prep", "pobj"]
,["_compound", "_attr", "nsubj"]
,["_compound", "relcl", "dobj", "prep", "pobj", "prep", "pobj"]
,["relcl", "dobj", "prep", "pobj", "prep", "pobj"]
```

```
,["_compound", "_npadvmod", "agent", "pobj"]
,["_npadvmod", "agent", "pobj"]
,["_nsubj", "dobj", "appos"]
,["_compound", "appos", "conj"]
,["_dobj"]
,["_nsubj", "prep", "pobj", "prep", "pobj"]
,["prep", "pobj"]
,["_compound", "_appos", "appos"]
,["_compound", "compound", "compound"]
,["compound", "compound"]
,["_nsubj", "advcl", "prep", "pobj"]
,["_compound", "_conj", "conj"]
,["_nsubj", "conj", "attr"]
,["_compound", "_nsubj", "attr"]
,["_compound", "_pobj", "_prep", "prep", "pobj"]
,["_nsubj", "ccomp", "nsubj"]
,["_compound", "relcl", "prep", "pobj"]
,["relcl", "prep", "pobj"]
,["_compound", "prep", "pobj"]
,["relcl", "attr"]
,["_compound", "_appos", "prep", "pobj"]
,["_compound", "_dobj", "prep", "pobj"]
,["_compound", "_dobj", "nsubj"]
,["_compound", "conj", "conj"]
,["_nsubj", "_conj", "nsubj"]
,["_dobj", "prep", "pobj"]
,["_appos", "compound"]
,["_appos", "appos"]
,["_nmod", "_appos"]
,["_compound", "_nsubjpass", "oprd"]
,["appos", "compound"]
,["_nsubj", "xcomp", "prep", "pobj"]
```

```
,["_compound", "_nsubj", "attr", "conj"]
,["_nsubj", "attr", "conj"]
]
}
```

8.2 Synonym metadata

{

"categorywords":[[["kindergarten teacher","kindergarten educationist","kindergarten educator","kindergarten instructor","kindergarten pedagogue","kindergarten preceptor"],"",""]

,[["dental hygienist","dental surgeon"],"",""]

,["speech-language pathologist","",""]

,["dental assistant","",""]

,["childcare worker","",""]

,["medical records technician","",""]

,[["secretary","clerk", "register", "registrar", "scribe"],"",""]

,["medical assistant","",""]

,[["hairdresser","barber", "haircutter", "hairstylist", "stylist"],"",""]

,[["dietitian","nutritionist"],"",""]

,[["vocational nurse","practical nurse"],"",""]

,["teacher assistant","",""]

,[["paralegal","legal assistant"],"",""]

,["billing clerk","",""]

,["phlebotomist","",""]

,["receptionist","",""]

,[["housekeeper","biddy","char"],"house boy",["charwoman","handmaid","handmaiden","house girl","maid","maidservant", "skivvy","wench"]]

,["nurse""",""]

,[["bookkeeper","archivist", "recorder", "reporter", "transcriptionist"],"",""]

,["health aide","",""]

,["taper","",""]

,[["steel worker","steel laborer","steel toiler"],"",""]

,["mobile equipment mechanic","",""]

,["bus mechanic",["bus repairman","bus serviceman"],["bus repairwoman","bus servicewoman"]]

,["service technician","",""]

,["heating mechanic",["heating repairman","heating serviceman"],["heating repairwoman","heating servicewoman"]]

,["electrical installer","",""]

,["operating engineer","",""]

,[["logging worker","logging laborer","logging toiler"],"",""]

,["floor installer","",""]

,["roofer","",""]

,["mining machine operator","",""]

,["electrician","",""]

,[["repairer","renovator"],"repairman","repairwoman"]

,["conductor","",""]

,["plumber","",""]

,[["carpenter","cabinetmaker"],"",""]

,["security system installer","",""]

,[["mason","brick layer"],"",""]

,["firefighter","fireman","firewoman"]

,[["salesperson","salesclerk","salespeople"],"salesman","saleswoman"]

,["director of religious activities","",""]

,["crossing guard","",""]

,[["photographer","lensman", "photog", "shooter", "shutterbug"],"",""]

,["lifeguard","",""]

,["lodging manager","","lodging manageress"]

,["healthcare practitioner","",""]

,["sales agent","",""]

,["mail clerk","",""]

,["electrical assembler","",""]

,["insurance sales agent","",""]

,["insurance underwriter","",""]

,["medical scientist","",""]

,["statistician","",""]

```
,["training specialist","",""]
,[["judge","adjudicator", "arbiter", "arbitrator"],"",""]
,[["bartender","barkeeper"],"barman","barwoman"]
,["dispatcher","",""]
,["order clerk","",""]
,["mail sorter","",""]],
"categoryidentifier" :  [
                    ["he", "man", "brother", "son", "husband", "boyfriend", "father",
"uncle", "dad"],
                    ["she", "woman", "sister", "daughter", "wife", "girlfriend",
"mother","aunt", "mom"]
                  ],
"categoryname" : ["male","female"],
"allowed_depend" : [
["_compound", "appos"]
,["appos"]
,["_compound", "compound"]
,["_compound", "_pobj", "_prep"]
,["_pobj", "_prep"]
,["compound"]
,["_nsubj", "attr"]
,["_nsubj", "prep", "pobj"]
,["_compound", "_nsubj", "nsubj"]
,["_nsubj", "nsubj"]
,["_compound", "relcl", "attr"]
,["_compound", "conj"]
,["_nsubj", "prep", "pobj", "conj"]
,["_compound", "_appos"]
,["_appos"]
,["_compound", "appos", "compound"]
,["_compound", "nmod"]
,["nmod"]
,["_nsubjpass", "prep", "pobj"]
```

```
,["_poss", "prep", "pobj"]
,["_compound", "_attr", "nsubj"]
,["_compound", "relcl", "dobj", "prep", "pobj", "prep", "pobj"]
,["relcl", "dobj", "prep", "pobj", "prep", "pobj"]
,["_compound", "_npadvmod", "agent", "pobj"]
,["_npadvmod", "agent", "pobj"]
,["_nsubj", "dobj", "appos"]
,["_compound", "appos", "conj"]
,["_dobj"]
,["_nsubj", "prep", "pobj", "prep", "pobj"]
,["prep", "pobj"]
,["_compound", "_appos", "appos"]
,["_compound", "compound", "compound"]
,["compound", "compound"]
,["_nsubj", "advcl", "prep", "pobj"]
,["_compound", "_conj", "conj"]
,["_nsubj", "conj", "attr"]
,["_compound", "_nsubj", "attr"]
,["_compound", "_pobj", "_prep", "prep", "pobj"]
,["_nsubj", "ccomp", "nsubj"]
,["_compound", "relcl", "prep", "pobj"]
,["relcl", "prep", "pobj"]
,["_compound", "prep", "pobj"]
,["relcl", "attr"]
,["_compound", "_appos", "prep", "pobj"]
,["_compound", "_dobj", "prep", "pobj"]
,["_compound", "_dobj", "nsubj"]
,["_compound", "conj", "conj"]
,["_nsubj", "_conj", "nsubj"]
,["_dobj", "prep", "pobj"]
,["_appos", "compound"]
,["_appos", "appos"]
```

```
,["_nmod", "_appos"]
,["_compound", "_nsubjpass", "oprd"]
,["appos", "compound"]
,["_nsubj", "xcomp", "prep", "pobj"]
,["_compound", "_nsubj", "attr", "conj"]
,["_nsubj", "attr", "conj"]
            ]
}
```

8.3 Balancing numbers dependency, sentence, without names, 50:50, adding

```
################Before statistic##################
kindergarten teacher male: 0 female: 0
dental hygienist male: 0 female: 0
nurse male: 41 female: 108
dental assistant male: 0 female: 0
secretary male: 399 female: 106
medical assistant male: 0 female: 0
hairdresser male: 6 female: 5
dietitian male: 1 female: 1
paralegal male: 1 female: 0
receptionist male: 4 female: 9
housekeeper male: 18 female: 22
registered nurse male: 0 female: 0
bookkeeper male: 6 female: 6
health aide male: 0 female: 0
taper male: 0 female: 0
steel worker male: 0 female: 0
roofer male: 3 female: 0
electrician male: 18 female: 0
conductor male: 77 female: 13
plumber male: 12 female: 0
```

```
carpenter male: 49 female: 17
mason male: 49 female: 15
firefighter male: 5 female: 0
salesperson male: 2 female: 0
director of religious activities male: 0 female: 0
crossing guard male: 0 female: 0
photographer male: 61 female: 30
lifeguard male: 5 female: 2
sales agent male: 0 female: 0
insurance underwriter male: 0 female: 0
medical scientist male: 0 female: 0
statistician male: 3 female: 0
judge male: 202 female: 56
bartender male: 12 female: 2
dispatcher male: 2 female: 0
order clerk male: 0 female: 0
mail sorter male: 0 female: 0
################After statistic##################
kindergarten teacher male: 0 female: 0
dental hygienist male: 0 female: 0
nurse male: 108 female: 108
dental assistant male: 0 female: 0
secretary male: 401 female: 400
medical assistant male: 0 female: 0
hairdresser male: 6 female: 6
dietitian male: 1 female: 1
paralegal male: 1 female: 0
receptionist male: 9 female: 9
housekeeper male: 22 female: 22
registered nurse male: 0 female: 0
bookkeeper male: 8 female: 8
health aide male: 0 female: 0
```

taper male: 0 female: 0

steel worker male: 0 female: 0

roofer male: 3 female: 0

electrician male: 18 female: 0

conductor male: 77 female: 77

plumber male: 12 female: 0

carpenter male: 49 female: 49

mason male: 48 female: 49

firefighter male: 5 female: 0

salesperson male: 2 female: 0

director of religious activities male: 0 female: 0

crossing guard male: 0 female: 0

photographer male: 61 female: 61

lifeguard male: 5 female: 5

sales agent male: 0 female: 0

insurance underwriter male: 0 female: 0

medical scientist male: 0 female: 0

statistician male: 3 female: 0

judge male: 202 female: 202

bartender male: 12 female: 12

dispatcher male: 2 female: 0

order clerk male: 0 female: 0

mail sorter male: 0 female: 0

8.4 Balancing numbers word existing, sentence, without names, 50:50, adding

```
################Before statistic##################
kindergarten teacher male: 3 female: 6
dental hygienist male: 2 female: 3
nurse male: 232 female: 372
dental assistant male: 1 female: 2
secretary male: 1697 female: 416
medical assistant male: 3 female: 1
hairdresser male: 26 female: 23
dietitian male: 2 female: 3
paralegal male: 5 female: 2
receptionist male: 27 female: 49
housekeeper male: 70 female: 78
registered nurse male: 4 female: 5
bookkeeper male: 24 female: 16
health aide male: 0 female: 0
taper male: 2 female: 0
steel worker male: 2 female: 1
roofer male: 7 female: 0
electrician male: 40 female: 9
conductor male: 281 female: 80
plumber male: 39 female: 12
carpenter male: 244 female: 99
mason male: 363 female: 97
firefighter male: 19 female: 6
salesperson male: 2 female: 5
director of religious activities male: 1 female: 0
crossing guard male: 3 female: 0
photographer male: 272 female: 115
lifeguard male: 20 female: 8
sales agent male: 0 female: 1
```

```
insurance underwriter male: 0 female: 0
medical scientist male: 1 female: 0
statistician male: 8 female: 1
judge male: 1151 female: 312
bartender male: 59 female: 15
dispatcher male: 10 female: 1
order clerk male: 0 female: 0
mail sorter male: 0 female: 0
################After statistic##################
kindergarten teacher male: 6 female: 6
dental hygienist male: 2 female: 3
nurse male: 375 female: 382
dental assistant male: 1 female: 2
secretary male: 1708 female: 1722
medical assistant male: 3 female: 3
hairdresser male: 26 female: 26
dietitian male: 3 female: 3
paralegal male: 5 female: 5
receptionist male: 49 female: 49
housekeeper male: 78 female: 78
registered nurse male: 4 female: 5
bookkeeper male: 27 female: 26
health aide male: 0 female: 0
taper male: 2 female: 0
steel worker male: 2 female: 1
roofer male: 7 female: 0
electrician male: 40 female: 40
conductor male: 281 female: 281
plumber male: 39 female: 39
carpenter male: 248 female: 255
mason male: 363 female: 365
firefighter male: 19 female: 19
```

salesperson male: 5 female: 5

director of religious activities male: 1 female: 0

crossing guard male: 3 female: 0

photographer male: 272 female: 272

lifeguard male: 20 female: 20

sales agent male: 0 female: 1

insurance underwriter male: 0 female: 0

medical scientist male: 1 female: 0

statistician male: 8 female: 8

judge male: 1151 female: 1151

bartender male: 59 female: 61

dispatcher male: 10 female: 1

order clerk male: 0 female: 0

mail sorter male: 0 female: 0

8.5 Balancing numbers word existing, sentence, without names, 50:50, remove

```
################Before statistic##################

kindergarten teacher male: 3 female: 6

dental hygienist male: 2 female: 3

nurse male: 232 female: 372

dental assistant male: 1 female: 2

secretary male: 1697 female: 416

medical assistant male: 3 female: 1

hairdresser male: 26 female: 23

dietitian male: 2 female: 3

paralegal male: 5 female: 2

receptionist male: 27 female: 49

housekeeper male: 70 female: 78

registered nurse male: 4 female: 5

bookkeeper male: 24 female: 16

health aide male: 0 female: 0

taper male: 2 female: 0

steel worker male: 2 female: 1

roofer male: 7 female: 0

electrician male: 40 female: 9

conductor male: 281 female: 80

plumber male: 39 female: 12

carpenter male: 244 female: 99

mason male: 363 female: 97

firefighter male: 19 female: 6

salesperson male: 2 female: 5

director of religious activities male: 1 female: 0

crossing guard male: 3 female: 0

photographer male: 272 female: 115

lifeguard male: 20 female: 8

sales agent male: 0 female: 1
```

```
insurance underwriter male: 0 female: 0
medical scientist male: 1 female: 0
statistician male: 8 female: 1
judge male: 1151 female: 312
bartender male: 59 female: 15
dispatcher male: 10 female: 1
order clerk male: 0 female: 0
mail sorter male: 0 female: 0
################After statistic##################
kindergarten teacher male: 3 female: 3
dental hygienist male: 1 female: 1
nurse male: 231 female: 241
dental assistant male: 1 female: 1
secretary male: 435 female: 416
medical assistant male: 1 female: 1
hairdresser male: 20 female: 23
dietitian male: 2 female: 2
paralegal male: 2 female: 2
receptionist male: 27 female: 28
housekeeper male: 69 female: 72
registered nurse male: 3 female: 3
bookkeeper male: 15 female: 16
health aide male: 0 female: 0
taper male: 0 female: 0
steel worker male: 1 female: 1
roofer male: 0 female: 0
electrician male: 8 female: 9
conductor male: 83 female: 80
plumber male: 12 female: 12
carpenter male: 104 female: 99
mason male: 102 female: 97
firefighter male: 5 female: 6
```

salesperson male: 2 female: 0

crossing guard male: 0 female: 0

photographer male: 121 female: 115

lifeguard male: 8 female: 8

sales agent male: 0 female: 0

insurance underwriter male: 0 female: 0

medical scientist male: 0 female: 0

statistician male: 1 female: 1

judge male: 328 female: 310

bartender male: 15 female: 15

dispatcher male: 1 female: 1

order clerk male: 0 female: 0

mail sorter male: 0 female: 0