Faculty of Science

Department of Information and Computing Science

Business Informatics

Master's Thesis

Bias analysis of NLP models for violence risk assessment

Analysis of gender bias in Dutch NLP models for the violence domain with a real-world ML case study on violence risk assessment

Author:

Joppe Kooistra

Supervisors:

P. Mosteiro Romero

Utrecht University

G. Soganciogl u

Utrecht University

H. Kaya

Utrecht University

December 22, 2023

Abstract

This research delves into gender bias and its mitigation in NLP models for violence risk assessment within the psychiatric care domain. A comparison between transformer-based monolingual and multilingual models is made, as well as a comparison between transformer-based models and a classical machine learning algorithm. First, the dataset is analyzed to gain insights into class balance and gender distribution. Then, the NLP models are trained on the dataset, and their performance and fairness metrics are evaluated. After that, a data augmentation technique is used on the data before running another training round. Finally, the Reject Option Classification method is used in post-processing to optimize performance and fairness. The most important findings are that the monolingual models outperform the multilingual ones, but there is little difference between domain-specific and general models. As with previous work on this topic, the classical machine learning model SVM outperforms the transformer-based models. Furthermore, bias mitigation methods should be carefully chosen, based on the desired metrics to improve, since they often come with trade-offs. Data augmentation led to increased counterfactual fairness for most models, but not for all. However, for some models, this came at the cost of predictive parity fairness, whereas in others this increased. Reject Option Classification showed mixed results as well, improving counterfactual fairness or predictive parity for some models but decreasing it in others. Understanding these trade-offs is the key to successful bias mitigation. This research contributes valuable insights into gender bias mitigation within psychiatric care and offers a thoughtful consideration of trade-offs in adopting bias mitigation strategies. The findings offer a perspective on the dynamics of transformer-based models and classical machine learning algorithms, contributing to the ongoing discourse on responsible and effective AI deployment in mental healthcare contexts.

Keywords: Bias mitigation, NLP, mental healthcare, word embeddings, transformer models, SVM, predictive parity, counterfactual fairness

Contents

1 Introduction **1**
1.1 Research Questions . . . 1
1.2 Relevance . . . 2
1.3 Research outline . . . 3

2 Background **5**
2.1 Word embeddings . . . 5
2.1.1 Pre-trained word embeddings . . . 5
2.1.2 Classical machine learning approaches . . . 8
2.2 Bias in NLP models . . . 8
2.2.1 Pre-processing . . . 9
2.2.2 In-processing . . . 10
2.2.3 Post-processing . . . 10
2.3 Bias measures . . . 10
2.4 Violence risk assessment . . . 11

3 Related work **13**
3.1 Fairness of NLP models in mental healthcare . . . 13
3.2 Measuring bias . . . 14
3.3 Bias mitigation methods . . . 15

4 Methodology **17**
4.1 Data analysis . . . 17
4.2 Violence risk assessment models . . . 17
4.2.1 Pre-processing . . . 17
4.2.2 Training . . . 18
4.2.3 Evaluation . . . 18
4.3 Bias mitigation . . . 19

5 University Medical Center Utrecht Psychiatry department dataset **21**
5.1 Dataset source . . . 21
5.2 Demographics . . . 22
5.3 Training, testing, and validation splits . . . 24

6 Experimental results **25**
6.1 Counterfactual fairness evaluation . . . 25
6.1.1 Original setting . . . 25
6.1.2 Neutralized setting . . . 26
6.1.3 Bias mitigation: data augmentation . . . 26
6.1.4 Bias mitigation: Reject Option Classification . . . 26
6.2 Equalized odds evaluation . . . 27
6.2.1 Original setting . . . 27
6.2.2 Neutralized setting . . . 28

6.2.3 Bias mitigation: data augmentation . . . 28
6.2.4 Bias mitigation: Reject Option Classification . . . 28
6.3 Complete performance overview . . . 29
6.3.1 BERTje . . . 29
6.3.2 MedRoBERTaNL . . . 30
6.3.3 MBERT . . . 31
6.3.4 SVM . . . 32

7 Discussion 34
7.1 Model performance . . . 34
7.2 Fairness . . . 34
7.2.1 Source of bias and bias magnitude . . . 35
7.2.2 Impact on downstream task . . . 35
7.3 Bias mitigation methods . . . 36
7.4 Limitations . . . 36
7.4.1 Generalizability . . . 36
7.4.2 Domain-specific model limitations . . . 37
7.4.3 Manual selection of gendered terms . . . 37
7.5 Ethical considerations . . . 37

8 Conclusion 38
8.1 Future work . . . 38

References V

Appendices VI

A Appendix A: Code VI

1 Introduction

The recent advances in machine learning have paved the way for a plethora of new implementations. These advancements hold immense potential for improving decision-making, automating complex tasks, and enhancing overall efficiency. However, alongside these advancements, concerns regarding fairness and bias have arisen regarding the ethical deployment of machine learning systems. Fairness refers to the concept that decisions made by machine learning models should be fair, impartial, and unbiased towards any specific group or individual based on their inherent or acquired characteristics (Mehrabi et al., 2021). These characteristics may include race, gender, religion, or socioeconomic status, among others. The aim is to prevent any form of discrimination or bias in algorithmic decision-making processes, promoting fairness and equal opportunities for all individuals.

However, implementing fairness is a challenge since it often requires trade-offs between fairness and other criteria such as efficiency and accuracy, and it is context and use-case specific, meaning that there is no one-size-fits-all solution. As such, it is difficult to achieve fairness without compromising. As such, this has become an area that has gained research popularity.

Despite the growing popularity of machine learning applications in domains such as mental healthcare and risk assessment, research on fairness and bias within these applications remains relatively limited. For instance, predictive models for depression and anxiety have shown promising results in leveraging sensor and usage data from personal digital devices to infer psychological outcomes (Hatton et al., 2019), but the examination of bias and fairness in such models has often been overlooked. Similarly, violence risk assessment models based on natural language processing (NLP) of clinical notes have demonstrated the potential to predict future violent incidents in psychiatric care settings (Mosteiro et al., 2022). Yet, the presence of bias remains unexplored.

This thesis aims to address these gaps in research by focusing on bias and fairness within the Dutch language context. Specifically, it will investigate the bias present in Dutch pre-trained word embeddings for the violence domain, analyzing the potential implications and consequences of such biases in downstream tasks. Furthermore, the thesis will evaluate the fairness of a violence risk assessment model using data from the UMC Utrecht, assessing whether the model exhibits biases against certain groups based on gender, race, or other factors. In doing so, this research seeks to contribute to the understanding of bias in NLP models within the Dutch language and propose suitable mitigation strategies to enhance fairness in algorithmic decision-making processes.

1.1 Research Questions

The main research question is *Do pre-trained Dutch word embeddings carry bias towards protected groups (e.g. gender, race) for the violence risk assessment domain?* To answer this question, the following sub-questions should be answered.

- SQ1: What is the main source of bias in word embeddings?

  SQ1.1 Is bias introduced during pre-training?

  SQ1.2 Is bias introduced during fine-tuning?

- SQ2: How does this bias impact the downstream task of violence risk assessment?
- SQ3: Is there a difference in terms of the magnitude of bias that pre-trained embeddings have?

  SQ3.1 Is there a difference between pre-trained Dutch embeddings compared to multilingual embeddings?

  SQ3.2 Is there a difference between pre-trained embeddings on general data compared to domain-specific data?
- SQ4: Which bias mitigation methods are most efficient to minimize the bias?

Since bias can be caused by multiple different things, and can be introduced at different stages, SQ1 will be answered by analyzing the violence risk assessment ML application and the NLP techniques which are used, its predictions, and the dataset it is trained on. We will first analyze whether the dataset has sampling bias, e.g. the number of male vs female patients in the dataset. Furthermore, we plan to analyze the dataset properties in order to understand the difficulty of the problem, such as the average number of unique words per note, and the readability score of the notes. Then, we will evaluate the bias of contextualized word embeddings, namely BERTje and MBERT, by using available measures in the literature. Related works use measures such as demographic parity, equality of opportunity for the positive class, equality of opportunity for the negative class, WEAT, and log probability score (Zhang et al., 2020; Kurita et al., 2019). To achieve this, we need to craft a dictionary consisting of words/terms in the violence domain. Related literature will be consulted to retrieve suitable bias detection methods. SQ2 will be answered by analyzing the application and its predictions. SQ3 will be answered by analyzing and comparing the NLP techniques, more specifically the multilingual word embeddings from MBERT and the Dutch word embeddings from BERTje. Finally, the results of the previous sub-questions, combined with the results of the literature review on bias mitigation methods, will allow us to answer SQ4 and propose a suitable bias mitigation strategy for the application. Specifically, we plan on using the AIFarness360 library for bias mitigation methods, which includes pre-, in-, and post-processing algorithms.

1.2 Relevance

Currently, most research in this area is focused on English word embeddings Caliskan et al. (2017); May et al. (2019); Bartl et al. (2020). However, very little has actually been researched with regard to lower resource languages, such as Dutch. As such, the main contribution of this work is to provide insights into the bias of Dutch word embeddings, focused on the violence risk assessment domain. The performance of the language-specific Dutch word embeddings from BERTje will be compared to the multilingual version MBERT. Furthermore, many publications regarding NLP use in the mental health care domain either gloss over the topic of bias, or completely omit it, even though it is a relevant topic. Addressing this for an existing application, such as the one used in the case study, is a step in the right direction.

1.3 Research outline

The research outline for this is visualized in Figure 1 and will be further explained here. First, we perform an initial literature study to extract relevant literature topics. These topics are then used to perform a systematic literature study. This will result in a number of bias measures that can be used to quantify bias, and a list of bias mitigation methods that can be used to mitigate bias. Then, using the bias measures, the bias of the UMC dataset, BERT-based models, and the Doc2Vec SVM model (which will all be further explained in the background chapter) is quantified. Based on the bias measurements, a suitable bias mitigation method from the literature study is selected, applied, and tested. Then, the results are reported. This process is repeated for all the bias mitigation methods that were selected from the literature. Depending on the performance of the bias mitigation method, the result can either be positive or negative. The method should reduce bias to an acceptable level, while not having a drastically negative impact on other performance metrics.

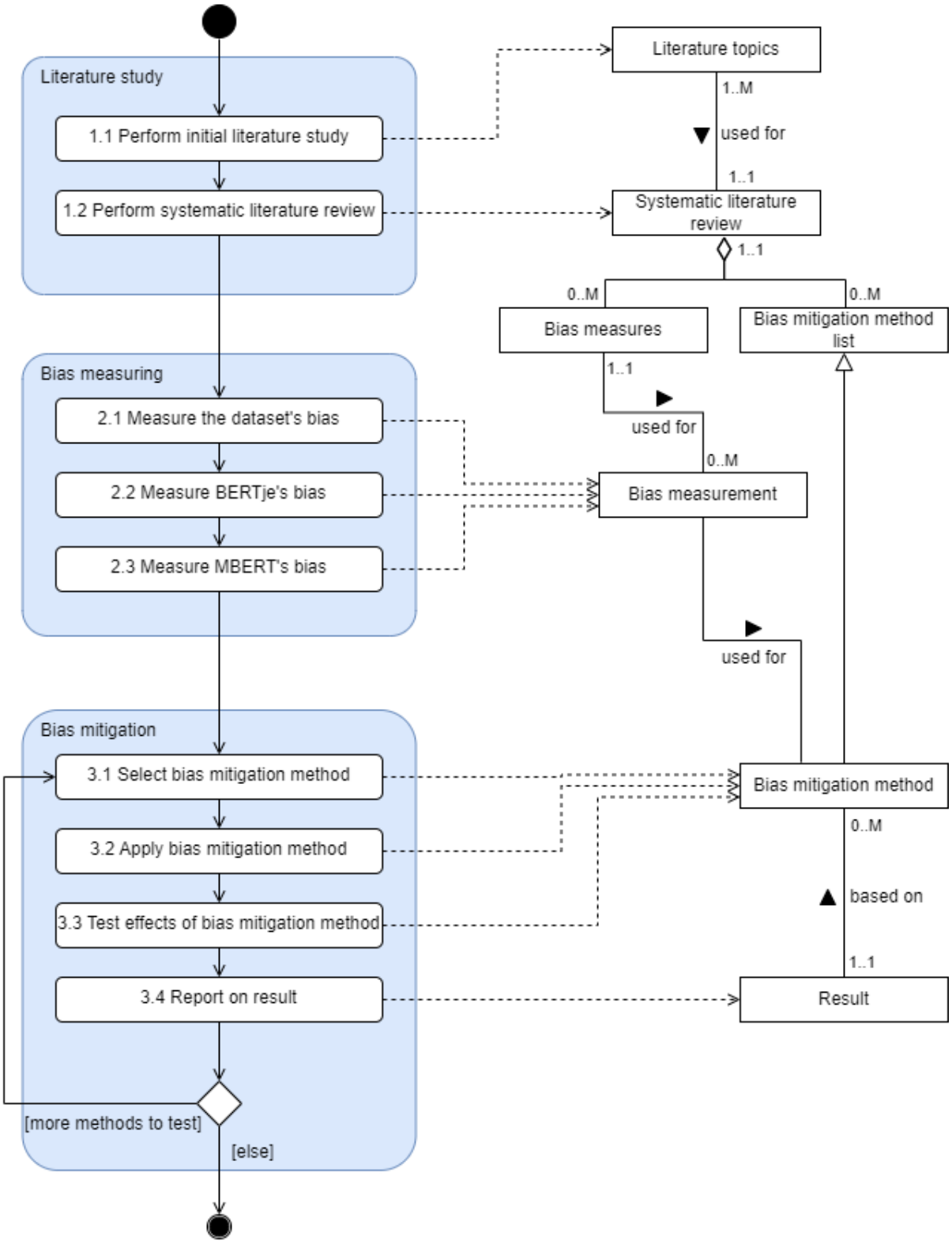

Figure 1: PDD of research outline

2 Background

This chapter delves into the background topics that build a framework for the rest of the research. First, we will discuss word embeddings and their variants, followed by bias in NLP models, bias measures to minimize said bias, and finally violence risk assessment.

2.1 Word embeddings

A word embedding is a vector that represents information about a word, such as its semantic and syntactic properties as found in a text or corpus. Word embeddings are successful at the semantic representation of words, but can potentially exhibit bias due to the dataset on which they are trained. This, in turn, can cause harm in downstream applications that use these embeddings (Sogancioglu et al., 2022). Word embeddings can be grouped into two categories: contextual word embeddings and context-independent word embeddings. Contextual word embeddings look at the context of a word, e.g. the surrounding words or the entire sentence. As such, the embedding of a single word can differ depending on the context in which it is used. On the contrary, contextual-independent word embeddings only look at the words on a global level, and as such the embedding of a single word is always the same, regardless of the context (Miaschi & Dell'Orletta, 2020). One of the most well-known contextual embeddings is Paragraph2Vec. This model associates a unique vector representation to each paragraph or document in a corpus. This vector is learned during a training phase, in which the paragraph vector is refined so that it captures the semantic meaning of the paragraph in the context of the training data. After training, the paragraph vectors can be used in various downstream tasks like text classification (Le & Mikolov, 2014).

There are multiple ways to train models for word embeddings. With supervised learning, the training data is labeled. This means that the classifier is 'told' what a certain feature of the data means, and what the desired output is based on the input, and over time the model learns. On the other side, there is unsupervised learning, in which unlabeled data is used for the training data. These algorithms discover hidden patterns, similarities, differences, and hidden groupings in the data, without the need for labels, and thus human intervention. Finally, there is the possibility of semi-supervised learning, which combines both methods. Here, the training data consists of a small amount of labeled data, and a large amount of unlabeled data (Zhu, 2005).

2.1.1 Pre-trained word embeddings

Pre-trained word embeddings are word embeddings that have been trained by the authors on large datasets. These embeddings are very useful in cases where the dataset that is to be analyzed is too small to be used as training data. However, there is a trade-off: the pre-trained embeddings might not be the best fit for the domain at hand (Rezaeinia et al., 2019).

There are two approaches for applying pre-trained language representations to downstream tasks: feature-based and fine-tuning. With the feature-based approach, task-specific architectures are used that include pre-trained representations as additional fea-

tures. For the fine-tuning approach, minimal task-specific parameters are introduced, and the training on the downstream tasks is done by simply fine-tuning all pre-trained parameters (Devlin et al., 2018).

2.1.1.1 BERT

One of the most well-known and best-performing pre-trained language models is the Bidirectional Encoder Representations from Transformers (BERT) model, developed by Google. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. There are two steps in the BERT framework, pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For the fine-tuning part, the BERT model is first initialized with pre-trained parameters. All parameters are then fine-tuned based on labeled data from the downstream tasks. Each downstream task has separate fine-tuning models, but the initialization is done with the same pre-trained parameters. However, the difference between the pre-trained architecture and the final downstream architecture of the model is minimal (Devlin et al., 2018).

The BERT model architecture uses a multi-layer bidirectional transformer encoder, based on the Transformer architecture depicted in Figure 2, which is a neural network architecture that uses self-attention mechanisms to process input sequences Vaswani et al. (2017). BERT uses a variant of the Transformer architecture that has two main components: the encoder and the output layer. The encoder generates contextualized representations of the input sequence. It is composed of a stack of N identical layers, where N is a hyperparameter that determines the depth of the model. Each layer contains two sub-layers, a multi-head self-attention mechanism, and a position-wise feedforward network. An attention function maps a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is a weighted sum of the values, where the weight that has been assigned to each value is computed by a compatibility function of the query with the corresponding key. Because the attention mechanism is multi-headed, the model can attend to different parts of the input sequence simultaneously.

In addition to this, each of the layers in the encoder and decoder contains a fully connected feed-forward network. The position-wise feed-forward network consists of two linear transformations with a rectified linear unit (ReLU) activation in between. Each layer uses different parameters. Learned embeddings are used to convert the input tokens and output tokens to vectors. The learned linear transformation and softmax function convert the decoder output to predicted next-token probabilities. Positional encodings are added to the input embeddings at the bottoms of the encoder and decoder stacks in order to have information about the position of the tokens in the sequence (Vaswani et al., 2017). BERT makes use of WordPiece embeddings. The first token is always a special classification token, [CLS]. Additionally, a separation token [SEP] is introduced. During training, BERT uses a masked language modeling (MLM) objective, where a random subset of tokens in the input sequence are masked and the model is trained to predict the original tokens. This objective encourages the model to learn contextual relationships between words in the input sequence. Additionally, BERT uses a next-sentence prediction

(NSP) task that pre-trains text-pair representations. This allows the model to capture the relationship between two sentences (Devlin et al., 2018).

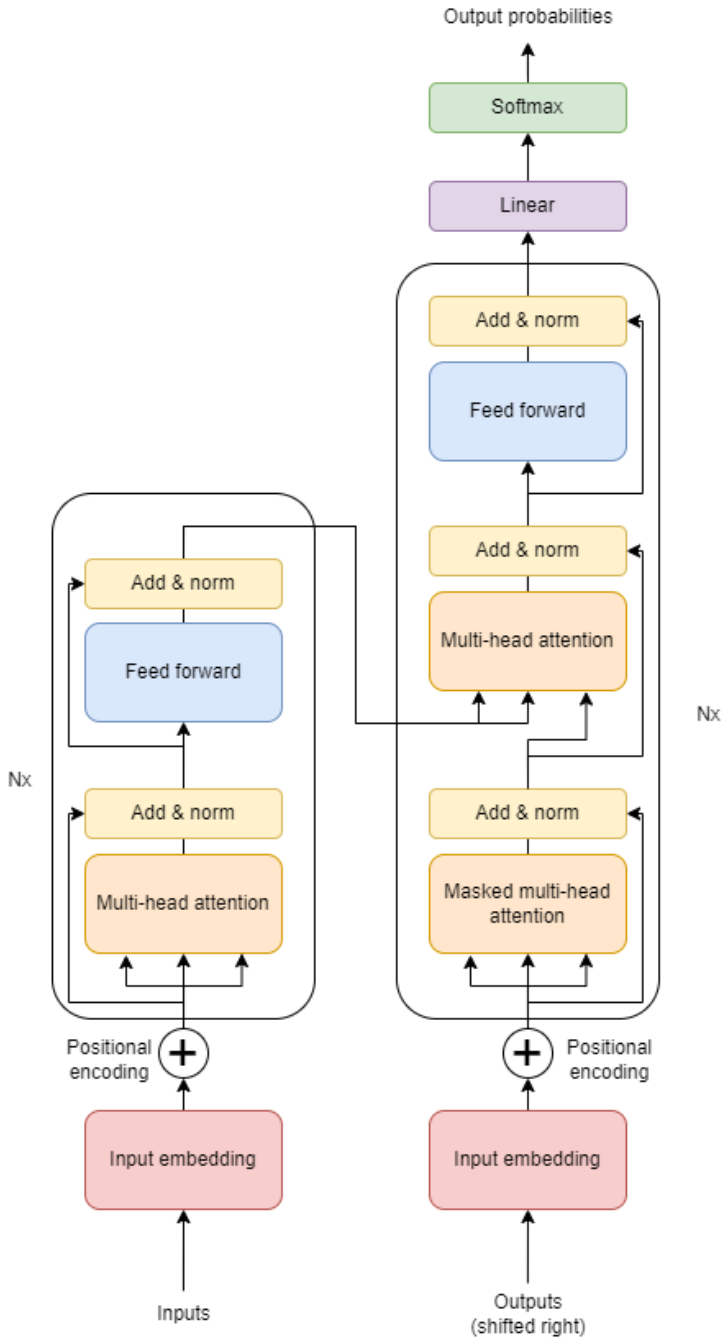

Figure 2: Transformer architecture that BERT is based on, image adapted from Vaswani et al. (2017)

It is possible to train BERT for specific domains, such as ClinicalBERT (Alsentzer et al., 2019) for clinical text, and BioBERT (Lee et al., 2020) for biomedical text. These domain-specific models yield performance improvements for a number of NLP tasks com-

pared to the standard nonspecific embeddings.

2.1.1.2 MBERT

Shortly after the release of BERT, a multilingual version of BERT (MBERT) was released. Whereas the original was only trained on the English language, this version was trained on 104 different languages. The training data consists of all the Wikipedia pages in these languages, which is a smaller dataset than the original BERT model used. Additionally, not every language has the same Wikipedia size. To counteract this, the data was weighted during pre-training. Other than that, the MBERT model works the same as the original BERT model for most languages (Devlin, 2018; Devlin et al., 2018).

2.1.1.3 BERTje

The release of the open-source BERT and MBERT models lead to the creation of several language-specific models that are based on BERT. The one of interest for this research is the Dutch version called BERTje (De Vries et al., 2019). BERTje is based on a larger and more diverse dataset than MBERT. It is trained on a collection of books, TwNC (a Multifaceted Dutch News Corpus), SoNaR-500 (a multi-genre reference corpus), almost 5 years of publications of web news from 4 different Dutch news websites, and Wikipedia, totaling approximately 2.4 billion tokens, whereas the MBERT model is only based on the Dutch Wikipedia pages. As such, BERTje is able to outperform MBERT on numerous downstream NLP tasks (De Vries et al., 2019).

2.1.1.4 MedRoBERTaNL

There is also the Dutch MedRoBERTaNL, which is trained on Dutch electronic health records. Since medical terms are often scarcely used in the training data of general models, a domain-specific model for the medical field can provide better results for certain tasks in this domain (Verkijk & Vossen, 2021). This model is based on RoBERTa, an optimized BERT pre-training approach (Liu et al., 2019).

2.1.2 Classical machine learning approaches

It is also possible to use word embeddings as input for classical machine learning approaches like Support Vector Machines (SVM). The idea here is that the learned vectors, e.g. from a Paragraph2Vec model, are used as features for an SVM model. After training the word embeddings, they are fed to an SVM model together with the labels, which the SVM uses to separate the feature vectors into different classes based on the training data (Burges, 1998; Stein et al., 2019).

2.2 Bias in NLP models

Algorithmic bias can be defined as *the outputs of an algorithm benefit or disadvantage certain individuals or groups more than others without a justified reason for such unequal*

impacts (Kordzadeh & Ghasemaghaei, 2022). This can lead to misinformed decisions and negative consequences for individuals, organizations, and society.

The recent progress that has been made with regard to NLP models has led to significant improvements for a wide variety of NLP tasks. However, success on certain criteria such as accuracy does not always mean that the model is performing satisfactorily on other important criteria such as fairness, robustness, and safety. Failures in fairness are often attributed to lack of transparency of modern AI tools. The reasoning behind this is that if biased predictions are caused by incorrect reasoning that is learned from biased data, transparency can assist to detect and understand this incorrect reasoning (Balkir et al., 2022).

Bias can be integrated into an algorithm in different stages. Fairness interventions to address these biases can be categorized in three categories: pre-processing, in-processing, and post-processing (Caton & Haas, 2020).

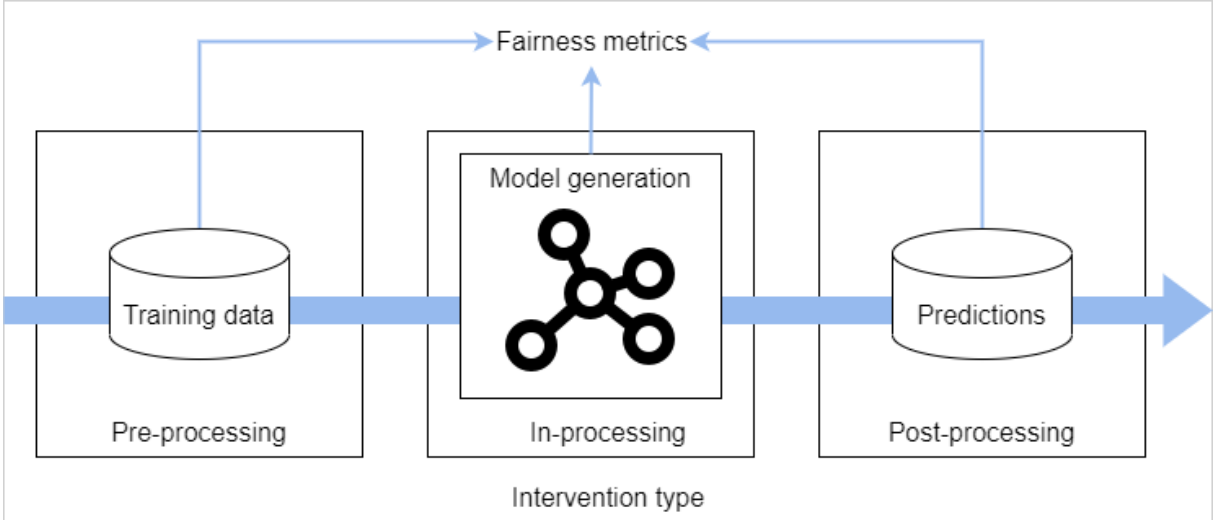

Figure 3: High-level illustration of fairness intervention stages, adapted from Caton & Haas (2020)

2.2.1 Pre-processing

Pre-processing bias stems from problems with the training data. An example of this is that if the training dataset is contaminated by the social biases that are present in organizations or society, those biased might be reflected in the output of the algorithm. For example, if mortgage applications have historically been approved more often for men than women, while important factors such as debt and income were similar, the algorithm might conclude that gender is an important criterion for risk assessment and approval of an application. If unaddressed, this algorithm will then reinforce the gender bias that was present in the training data. Another source of bias in the pre-processing stage is non-representative data. This can be caused by poor data selection, where data is not representative of the diversity of the population and certain groups are underrepresented (Favaretto et al., 2019). Pre-processing problems can also occur in word embeddings. For

example, word embeddings that have been trained on co-occurrence in bodies of text can serve as a dictionary for NLP tasks. This makes it possible to infer the meaning of words. Given the question 'man is to king as woman is to X', simple arithmetic of the embedding vectors finds the word 'queen' for X. However, using the same question structure for the question 'man is to computer programmer as woman is to X' gives the word 'homemaker' for X. This shows a tremendous gender bias towards certain words, which is problematic (Bolukbasi et al., 2016a). One way to reduce pre-processing bias is by using strategic sampling for the training data. This approach makes the data look balanced with regard to the class-domain frequencies. Rare examples are sampled more often during training. However, there are also drawbacks to this method. With exact copies of the same example in the training data, overfitting becomes more likely. Additionally, learning time is increased, because additional training examples are introduced that do not contain new information (Wang et al., 2020).

Research has shown that pre-trained word embeddings inherit the biases that are present in the corpora on which they are trained. Even the embeddings that use a vast amount of training data from multiple sources suffer from the societal and cultural stereotypes that are present in the training data (Caliskan et al., 2017).

2.2.2 In-processing

In-processing interventions, also known as enforcing fairness during model training, concern direct changes to the model. As such, it is highly effective. However, it has many practical limitations. This method requires training a model from scratch. With the release of powerful and publicly available pre-trained models such as BERT and its derivatives (e.g. MBERT or BERTje), and given the computational resources that are required, creating a model from scratch has become less common (Petersen et al., 2021).

2.2.3 Post-processing

Post-processing approaches recognize that the output of an ML model may be unfair towards one or more protected variables, and as such tend to apply transformations to the output of the model to improve the fairness of predictions. This approach only needs access to the predictions and sensitive attribute information, and not to the actual algorithms and models, and as such is one of the most flexible approaches (Caton & Haas, 2020).

2.3 Bias measures

Bias metrics in NLP models are used to measure and quantify the presence of bias in the model's outputs. Bias in NLP models can manifest in different ways, such as underrepresentation or over-representation of certain groups or demographics, leading to unfair or discriminatory results. Bias metrics can help to identify and mitigate these issues.

Most of these metrics can be categorized in the following categories(Caton & Haas, 2020):

- Statistical parity: each group should receive an equal fraction of the possible decision outcome.
- Demographic parity: This metric measures the extent to which the model's predictions are consistent across different demographic groups, such as gender or race, or if there is a disparate impact. If the model's predictions show significant differences between these groups, that could indicate that the model is biased (Feldman et al., 2015).
- Equal opportunity: A model is considered to be fair and under equal opportunity if the model's true positive rate (TPR) is consistent across different demographic groups. If there is a significant difference in the TPR across these groups, it could indicate bias in the model (Hardt et al., 2016).
- Calibration: False positive rates across groups should be similar. If a certain group has significantly more false positive hits, this could indicate bias.
- Counterfactual fairness: This metric measures the extent to which the model's predictions for an individual would change if certain sensitive variables were changed. If the predictions change significantly, it could indicate bias in the model with regards to these variables (Chiappa, 2019).

Specific metrics will be further discussed in Section 3.

2.4 Violence risk assessment

In the domain of acute psychiatric care, violence from patients directed at staff is prevalent in almost any treatment facility. In a systematic review Iozzino et al. (2015) found that 17% of patients committed at least one act of violence. This can have many adverse effects, not only for potential injuries to patients and staff, but also because of the counter-therapeutic effects of both violence and the measures that are taken to prevent violence. As such, it is important to be able to assess the risk of a patient becoming violent. Violence risk assessment is the process of evaluating the likelihood that an individual will engage in violent behavior in the future, and is a well studied topic (Douglas & Skeem, 2005; Skeem & Monahan, 2011). This type of assessment is typically conducted by mental health professionals, law enforcement officials, or other professionals who have been trained in assessing and managing risk. The goal of a violence risk assessment is to determine the level of risk that an individual poses to themselves or others and to develop a plan to manage that risk. This involves evaluating a range of factors, including the individual's history of violence, mental health status, substance abuse, and social and environmental factors that may contribute to violent behavior (Douglas & Skeem, 2005). There are different types of violence risk assessment tools and approaches used, including clinical interviews, structured risk assessment instruments, and actuarial risk assessment methods. These methods typically involve gathering information from multiple sources, including interviews with the individual, family members, and other professionals involved in their care. The results of these assessments can be used to guide decisions about the individual's treatment, management, and supervision. This may involve developing a safety plan,

recommending appropriate treatment and medication, and monitoring the individual's behavior over time to assess their ongoing risk of violence. A common approach is the Brøset Violence Checklist (BVC) (Almvik et al., 2000). This is a questionnaire used by nurses and psychiatrists to evaluate the risk of a patient becoming involved in a violent incident. However, filling out this form is a time-consuming process, and moreover, it is highly subjective. It is important to note that violence risk assessment is not a perfect science, and there are limitations and potential biases associated with these assessments. However, when conducted by trained professionals using evidence-based methods, violence risk assessment can be a useful tool for reducing the risk of violent behavior and improving public safety (Skeem & Monahan, 2011).

3 Related work

This chapter delves into the related works in areas that are relevant for our research. First, we discuss other works on fairness of NLP models in mental healthcare, followed by works on specific bias measurement methods, finishing with specific bias mitigation methods that proved successful in other works.

3.1 Fairness of NLP models in mental healthcare

Adamou et al. (2018) describe the use of NLP on medical notes for suicide risk assessment. They use the Just-Add-Data (JAD) Bio tool for classification analyses, which is an automated multivariate statistical analysis pipeline comprising of a complete set of learning steps that lead to the production of the final predictive model. JAD Bio performs the pre-processing of the data, feature selection, training of predictive models, automated selection of the best configuration (fine-tuning), construction of the final predictive model, and performance estimation. Even though the dataset contains numerous demographic variables (including date of birth, gender, marital status, ethnicity, religion, and postcode), the only type of bias reduction that is mentioned is a feature of JAD that uses a bootstrap-based method called Bootstrap Bias Corrected Cross Validation (BBC-CV) (Tsamardinos et al., 2018). The main idea is to bootstrap the whole process of selecting the best-performing configuration on the out-of-sample predictions of each configuration, without additional training of models, which corrects for bias.

Chen et al. (2018) describe the use of NLP on Twitter posts to identify depression. The data that is used is collected from public Twitter posts. The system parses the text and classifies part-of-speech tags through a NLP pipeline, which is not further specified. The influences of certain attributes, such as personality, age, and gender on the disclosure of mental health problems were also examined. This was then used to create a control group to reduce the bias for the analysis. Additionally, the authors claim that the applied feature set provides reliable, unbiased measures. Similarly, Yazdavar et al. (2017) analyzes Twitter posts to monitor clinical depressive symptoms. They use a semi-supervised approach. First, a random selection of user profiles was manually examined. Then, an unsupervised text mining method called Latent Dirichlet Allocation (LDA) was used to extract topics. However, these topics were not specific enough to correspond to depressive symptoms. So, supervision was added to the LDA in order to arrive at specific topics. They mention that their sample might be biased because of the nature of the data, e.g. more severely depressed people being more likely to express their depression, but this is listed as a limitation so no solutions are discussed.

Gaur et al. (2018) discuss contextualized classification of Reddit posts in order to map a certain subreddit or community within the platform to the best matching DSM-5 (Diagnostic and Statistical Manual of Mental Disorders - 5th edition) category. The baseline approach resulted in the model being biased towards certain DSM-5 categories, since these were over-represented in the training and testing data. An oversampling procedure was tested, which reduced the false accurate rate, but also reduced the true positive rate. Additionally, an existing implementation bootstraps the process of oversampling by randomly drawing samples from the majority class. This resulted in a significant reduction

of the false positive rate.

Park et al. (2018) employ text mining on three online mental health communities from Reddit, namely anxiety, depression, and PTSD. The mined data was pre-processed and then represented in vector spaces. The results were then clustered based on topic similarity. The authors mention that the dataset is representative of the user base of the platform, which consists predominantly of younger males. As such, the dataset shows bias toward a younger, male audience. There is no mention of a solution or correction for this, only a direction for future work.

Nobles et al. (2018) describe the use of NLP on text messages to identify imminent suicide risk among young adults. They tried to identify unique patterns of communication that occur in advance of a suicide attempt, by using a deep neural network classifier.

Ray et al. (2019) use a multi-level attention network to predict depression, using text, audio, and video data. They used the pre-trained Universal Sentence Encoder (Cer et al., 2018) for the text classification. As a direction for future work, they mention getting rid of inductive bias from classes of data with more training samples and using few-shot learning techniques that can help with compensating for classes with fewer data.

Nikfarjam et al. (2015) describe the use of NLP techniques to extract mentions of adverse drug reactions (ADR) from informal text in social media. To form an exhaustive list of ADR concepts, a semi-supervised learning approach was used. First, a subset of user posts was annotated, which was then used for supervised learning. Then, the remaining set of user posts was used for unsupervised learning. To represent the individual tokens, they used the Word2Vec tool, which uses contextualized word embeddings. These word embeddings were then clustered, which was used to define features. However, this work does not mention bias or fairness.

Similarly, L. Ma et al. (2017) use Word2Vec on a dataset consisting of posts of numerous social media sites in order to extract depression symptoms. First, the word frequency is calculated. Then, using Word2Vec, the relationships among words in the documents are learned. The resulting vectors are then used to cluster the word embeddings. To reduce bias, the data is pre-processed.

3.2 Measuring bias

Caliskan et al. (2017) introduce the Word-Embedding Association Test (WEAT). This is based on the Implicit Association Test (IAT), which documents known human biases. WEAT uses the distance between a pair of vectors. In principle, the idea is that there are two sets of target words (e.g. programmer, engineer, scientists; and nurse, teacher, librarian) and two sets of attribute words (e.g. man, male; and woman, female). The null hypothesis is that there is no difference between the two sets of target groups with regard to the relative similarity to the two sets of attribute words. Then, permutations are tested to measure the likelihood of the null hypothesis by computing the probability that a random permutation of the attribute words would produce a difference in sample means. However, WEAT is only tested on non-contextual word embeddings. May et al. (2019) propose a simple generalization of WEAT, the Sentence Encoder Association Test (SEAT). SEAT is applied to sentences generated by inserting individual words from the tests in Caliskan et al. (2017) into simple templates such as 'This is a [word]. In addition

to being tested on non-contextual word embeddings, SEAT was also tested on BERT.

Bartl et al. (2020) propose a technique to measure and mitigate BERT's gender bias. They do this by first applying Name-Based Counterfactual Data Substitution (CDS), in which the gender of words denoting persons in a training corpus is swapped in place in order to counterbalance bias. Then, BERT is fine-tuned on the GAP corpus, which was developed as a benchmark for measuring gender bias in co-reference resolution systems. GAP contains 8908 ambiguous pronoun-name pairs in 4454 contexts sampled from Wikipedia.

Zhang et al. (2020) quantify bias in BERT-based embeddings by evaluating classifiers on three definitions of fairness:

- Demographic parity, defined as:
  $P(\hat{Y} = y) = P(\hat{Y} = \hat{y}|Z = z), \forall z \in Z$
- Equality of opportunity for the positive class, defined as:
  $P(\hat{Y} = 1|Y = 1) = P(\hat{Y} = 1|Y = 1, Z = z), \forall z \in Z$
- Equality of opportunity for the negative class, defined as:
  $P(\hat{Y} = 0|Y = 0) = P(\hat{Y} = 0|Y = 0, Z = z), \forall z \in Z$

3.3 Bias mitigation methods

P. Ma et al. (2020) propose MT-NLP, a framework to identify and mitigate unfair predictions of NLP models. This framework is based on a well-established software testing scheme named metamorphic testing (MT). MT first uses arbitrary inputs to generate a set of mutations, and then checks whether a given metaphoric relation still holds with relation to the mutated inputs. MT-NLP is designed as a pipeline. First, advanced NLP techniques are used on arbitrary natural language sentences to systematically explore sub-populations within sensitive attributes. Sentences are then perturbed regarding the sensitive attributes that were identified, and MT is performed on NLP models to detect fairness violations. Finally, model fairness is formalized and unfair predictions are mitigated accordingly. Since this framework does not concern any implementation details of a specific model, the authors claim it can be used on any NLP task. As such, it can be classified as a post-processing bias mitigation method.

Bolukbasi et al. (2016a) describe a method for specifically debiasing word embeddings. The first step is called Identify gender subspace. Here, a direction, or subspace, is identified that captures the bias of the embedding. For the second step, there are two options: Neutralize and Equalize, or Soften. Neutralize ensures that gender-neutral words are zero in the gender subspace. Equalize then equalizes sets of words outside the subspace, thus enforcing the property that any neutral word has the same distance to all words in each equality set. This can be utilized for applications where a word pair should not display any bias with respect to neutral words. However, the disadvantage is that Equalize removes certain distinctions that are valuable for certain applications. The Soften algorithm reduces the differences between sets while maintaining as much similarity to the original embedding as possible. This trade-off is controlled by a parameter.

Brunet et al. (2019) propose a method to eliminate bias in word embeddings at the source: the training data. Given a word embedding that has been trained on a certain corpus and a bias evaluation metric, the proposed method approximates how removing a small part of the training corpus would affect the resulting bias. The method provides a highly efficient way of understanding the impact that every document in the training dataset has on the overall bias of a word embedding. As such, it becomes possible to identify the most bias-inducing documents in the training corpus. The test data can then be pruned selectively to manipulate the word embedding's bias, or the results can be used to identify particularly biased subsets of the training data.

Bolukbasi et al. (2016b) propose two methods to systematically quantify the gender bias of word embeddings. First, it is quantified how words corresponding to professions are distributed along the direction between embeddings of he and she. Second, the authors have designed an algorithm that generates analogy pairs from an embedding given two seed words. Crowdworkers were used to quantify whether the embedding analogies reflect stereotypes. Using this, the authors proposed an approach that can reduce the amount of bias that is present in an embedding, given an embedding and only a handful of words, without a significant performance reduction.

Dev et al. (2020) argue that existing bias discovery methods often suffer from two problems. First, there is a mismatch between what is measured, e.g. vector distances or similarities, and how the embeddings are actually utilized for the downstream tasks. Second, many tests are designed for word types, and not word tokens. Given that contextualized word embeddings that are currently performing the best, e.g. BERT, use word tokens, this is also a mismatch. As such, the authors have designed probes that measure the effect of specific biases, and defined aggregate measures that quantify bias effects over a large number of predictions. Furthermore, the authors show that simple mechanisms for removing bias on static word embeddings work. For contextualized word representations, these mechanisms only work on the non-contextual part of the representation and can reduce gender bias, but not bias related to religion or nationality.

4 Methodology

This chapter will detail the methodology that is used during the research.

4.1 Data analysis

First, the dataset is analyzed to gain relevant insights. We analyze the gender distribution and class balance, and class distribution per gender, to get a better understanding of the domain. The gender distribution analysis identifies the ratio of male to female patients, while the class balance examination delves into the distribution of violent incidents and non-incidents within the dataset. Additionally, we also look at the length of the texts. Lastly, we make a list of all the gendered language that is used in the text. The insights gained from the data analysis will be used to select relevant pre-processing steps and evaluation metrics. For instance, in a highly imbalanced dataset, a metric like overall accuracy does not provide valuable insight into performance. In this case, unweighted F1 and AUC provide a more robust insight into model performance.

4.2 Violence risk assessment models

This section will explain the application of the VRA models. First, the necessary pre-processing steps are discussed. Then, the training process for the BERT-based and SVM models is detailed. Lastly, we will discuss the evaluation and model analysis steps.

4.2.1 Pre-processing

To standardize and simplify the text for processing, the input text is cleaned by removing single quotes and replacing sequences of non-alphanumeric characters with a single space. Since BERT-based models provide contextual embeddings over the entire input text, this pre-processing method alters the text as little as possible. For Doc2Vec, additional pre-processing steps are applied. First, the text is converted to lowercase. Then, diacritics (e.g. accents, umlauts, etc.), special characters, obsolete periods, whitespace, and tabs are removed. After cleaning, the text is tokenized into a list of words. Finally, stopwords are filtered, words are stemmed, and periods are filtered out.

After cleaning the text, the dataset is split into training, testing, and validation sets. The training data has the same class distribution of the outcome variable for males and females. This is done to ensure that differences in class distribution do not cause any potential gender bias. Then, the remaining data is evenly divided between the validation and test set.

A known issue with most transformer-based models is that the maximum input size is 512 tokens, which is roughly 500 words. Longer sequences are disproportionally expensive because the necessary computational resources scale quadratically with the sequence length (Devlin et al., 2018). There are a few ways to handle this problem. One option that works in some cases is to just use the first or last 512 tokens and discard the rest of the text (Sun et al., 2019). The drawback of this method for longer sequences is that a lot of potentially important information is lost. Another option is to split each sequence

into multiple chunks of max 512 tokens each, and then recombine the results to get a prediction for the entire sample (Ivgi et al., 2023). The drawback here is that not every chunk contains relevant information for the prediction. There are also some models available that can handle larger input sizes, like Longformer (Beltagy et al., 2020). However, these models are less complex and as such, the contextual representations might not be as good. In summary, there are different approaches to handle the maximum input size, but there is no perfect solution. Therefore, all the methods listed above will be tried initially to see what works best, and then we will continue with the best-performing method.

4.2.2 Training

After pre-processing, the training data is used to train four different models: BERTje, MedRoBERTaNL, MBERT, and an SVM model using Doc2Vec embeddings. BERTje, MedRoBERTaNL, and MBERT are pre-trained models, which need to be fine-tuned on a task-specific dataset. In this case, the objective is text classification. As such, the fine-tuning we do on these models initializes the classification layer of the models. As for the SVM model, we first train a paragraph-to-vector model called Doc2Vec on the entire dataset to get the document embeddings. This Doc2Vec model is then used to create the embeddings of the training data, which is used as input for the SVM model. Each different model is trained 5 times using different train/test/validation splits as cross-validation to reduce the effect of variability. The validation sets are used to tune hyper-parameters to get the best results.

4.2.3 Evaluation

After training, each model is evaluated on multiple metrics. We look at the performance and fairness of the original test set as well as the augmented test set, which is the original test set augmented with the same data but with all gendered words swapped. This is done to reduce the effect of the gender of individual examples, by feeding the model the exact same example with a different gender. After evaluating the different models we compare the performance and fairness scores of the different models to see how model behavior differs. We try to gain insights into model behavior, understand decision boundaries, and look for potential areas of improvement. This helps us interpret the results better and shows the strengths and weaknesses of each model.

4.2.3.1 Performance measures

The unweighted F1 score and unweighted AUC score are the most important for performance metrics. The unweighted score is the unweighted average of the scores for the positive class and the negative class, which is useful for imbalanced datasets since poor performance for the minority class is reflected better. Furthermore, overall accuracy is reported, and precision, recall, and F1-score are reported separately for the positive and negative classes.

4.2.3.2 Fairness measures

For fairness measures, we use the true positive rate ratio (the ratio of male true positives compared to female true positives), false positive rate ratio (the ratio of male false positives compared to female false positives), AUC score ratio (AUC score for male compared to AUC score for female), and mismatch ratio (the number of examples where the prediction changes if the gendered words are swapped).

4.3 Bias mitigation

For bias mitigation, we use two different methods. The first method is a pre-processing method called data augmentation. Like with the augmented test data, the training data will be augmented with the same data with the gendered words swapped, e.g. 'she' becomes 'he' or 'woman' becomes 'man'. The full list of gendered words from female to male is:

- 'zij': 'hij'
- 'ze': 'hij'
- 'haar': 'hem'
- 'mw': 'dhr'
- 'vrouw': 'man'
- 'patiente': 'patient'
- 'mevrouw': 'meneer'
- 'mevr': 'mr'
- 'meisje': 'jongen'
- 'dame': 'heer'

For male to female, this is the reverse. This way, the gender bias that is introduced during fine-tuning/model training is reduced and counterfactual fairness is improved. Furthermore, it serves to improve the model's ability to generalize across different gendered examples.

Furthermore, a post-processing method called Reject Option based Classification (ROC) is used, which uses the low confidence region of a classifier in order to reduce discrimination (Kamiran et al., 2012). By adjusting decision thresholds in the low confidence region, the ROC method contributes to enhancing equal opportunity fairness in model predictions. ROC allows for the selection of different fairness metrics to improve. Since we are looking to improve both the true positive rate and false positive rate between both genders, we will use 'Average Odds Difference'. This metric computes the average difference in these rates across all possible classification thresholds. It is possible to specify multiple parameters, like the upper and lower classification bounds and the number of

thresholds that will be tried. This influences the outcome and computing time, so multiple configurations will be tried on the validation set to find the best parameters. These will then be used on the test set to get the new results. Furthermore, the ROC needs a protected attribute, in our case gender, and a privileged and unprivileged group. This depends on the results of the models. The method also needs a favorable label and an unfavorable label. Since the task is violence risk assessment, the favorable label is debatable. We decided that positive would be the favorable label since ultimately the goal is to provide the proper care for a patient. They will not face any negative consequences for being labeled as positive, but might not get the proper care if falsely labeled negative.

Additionally, all models are trained on a neutralized dataset, where all gendered words are replaced by gender-neutral terms e.g. 'she' becomes 'they' or 'man' becomes 'person'. This allows us to gain insights into bias that is introduced during pre-training of the BERT-based models, since the models will not make any new gender associations. The full list of gender-neutral replacement terms is:

- 'hij': 'diegene'
- 'hem': 'diegene'
- 'zijn': 'diens'
- 'dhr': 'persoon'
- 'man': 'persoon'
- 'meneer': 'persoon'
- 'mr': 'persoon'
- 'jongen': 'persoon'
- 'heer': 'persoon'
- 'zij': 'diegene'
- 'ze': 'diegene'
- 'haar': 'diens'
- 'mw': 'persoon'
- 'vrouw': 'persoon'
- 'patiente': 'patient'
- 'mevrouw': 'persoon'
- 'mevr': 'persoon'
- 'meisje': 'persoon'
- 'dame': 'persoon'
- 'hr': 'persoon'

5 University Medical Center Utrecht Psychiatry department dataset

This chapter elaborates on the details of the dataset that we use to make our predictions. First, an overview of the source of the dataset and the general contents of the texts is provided to facilitate a better understanding of the texts that need to be classified. After that, we will provide an analysis of the demographics of the dataset with regard to class balance and gender distribution. Finally, an overview of the training, testing, and validation split is shown.

5.1 Dataset source

The dataset for predicting violence incidents was acquired from the Psychiatry Department at the University Medical Center Utrecht (UMCU) in the Netherlands. This department comprises six inpatient units that admit patients with diverse medical backgrounds, each focusing on specific patient populations, diagnoses, and treatments. The department provides secondary care to patients with severe but general psychiatric symptoms, and tertiary care for patients with complex symptomatology or comorbidities, which ensures a diverse population. The psychiatry units mandated the reporting of violence incidents by healthcare professionals, typically involving verbal and physical aggression from patients directed at staff or other patients. The prediction objective is to anticipate violence incidents between the first and 28th day, based on clinical texts written up to and including the first day of admission. The prediction task excludes violence incidents on the admission day due to insufficient data available at that point, particularly in acute admissions. This interval ensures specificity in prediction, with a high inclusion rate of incidents (Menger et al., 2018; Mosteiro et al., 2022).

Most of the clinically relevant information was entered into the Electronic Health Record (EHR) in free text format by psychiatrists and nurses, with entries typically containing between 100 and 500 words. These are respectively referred to as 'doctor notes' and 'nurse notes'. Doctor notes mainly contain information about patient history, current treatment details (e.g. types of medication and therapy), and changes therein. Nurse notes typically contain details about the patient's current well-being and activities. These notes are written three times a day by trained nurses. All notes were de-identified using the De-identification Method for Dutch Medical Text (DEDUCE) before any other processing took place (Menger et al., 2018).

A patient may undergo multiple admissions to the psychiatry ward, and during each admission, they might spend time in different sub-wards. These are referred to as admission periods, and are the data points for this study. All notes collected between 28 days before and 1 day after the beginning of the admission period are concatenated and considered a single period note for each admission period. The outcome variable is determined based on the occurrence of a violence incident within 1 to 28 days after the start of the admission period. If such an incident occurs, the outcome is recorded as violent (positive), otherwise it is noted as non-violent (negative) (Mosteiro et al., 2022).

To summarize, the relevant columns in the dataset are:

- Text: this column contains the period notes.
- Outcome: no violent incident (0) or violent incident (1) within 1 to 28 days after the start of the admission period.
- Gender: female (0) or male (1).

5.2 Demographics

The dataset is highly imbalanced with regard to the outcome variable. Of the total 4280 datapoints, 3855 (90%) are negative, where there was no violent incident, and 425 (10%) are positive, where there was a violent incident. This is depicted in Figure 4.

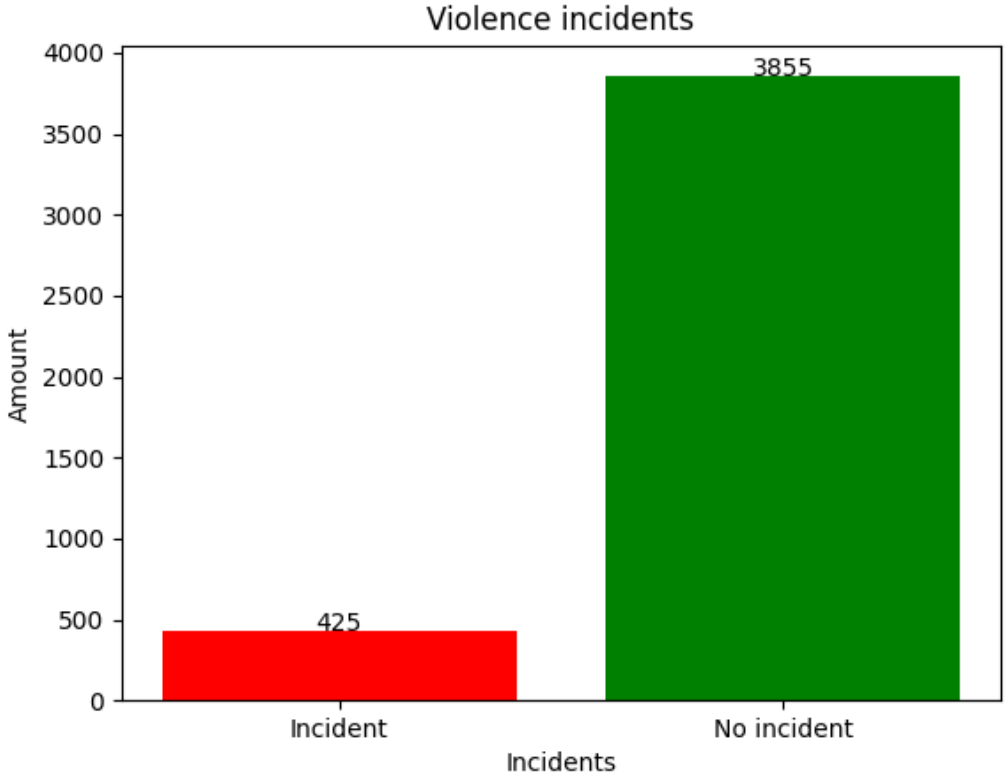

Figure 4: Amount of violent incidents

The gender distribution is better balanced. There are 2199 (51%) female patients and 2081 (49%) male patients, as depicted in Figure 5.

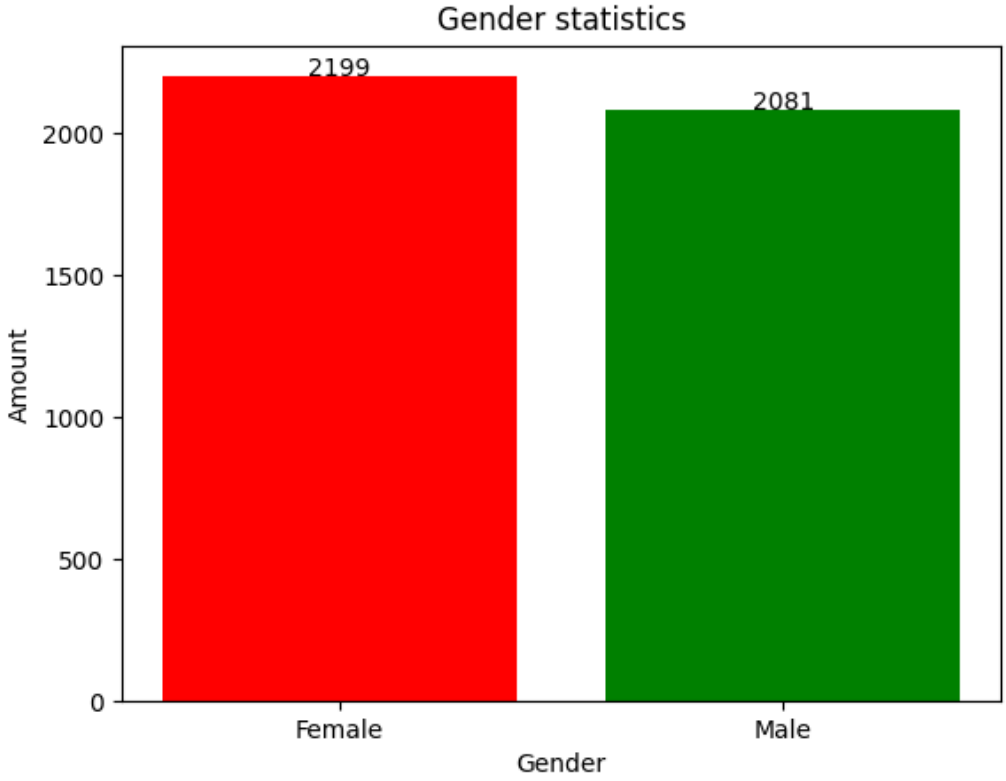

Figure 5: Gender distribution

There is also an imbalance with regard to the distribution of violent incidents per gender. For females, there are 166 violence incidents out of 2199 patients, which is 7.55%. For males, there are 259 violence incidents out of 2081 patients, which is 12.45%. This is shown in 6.

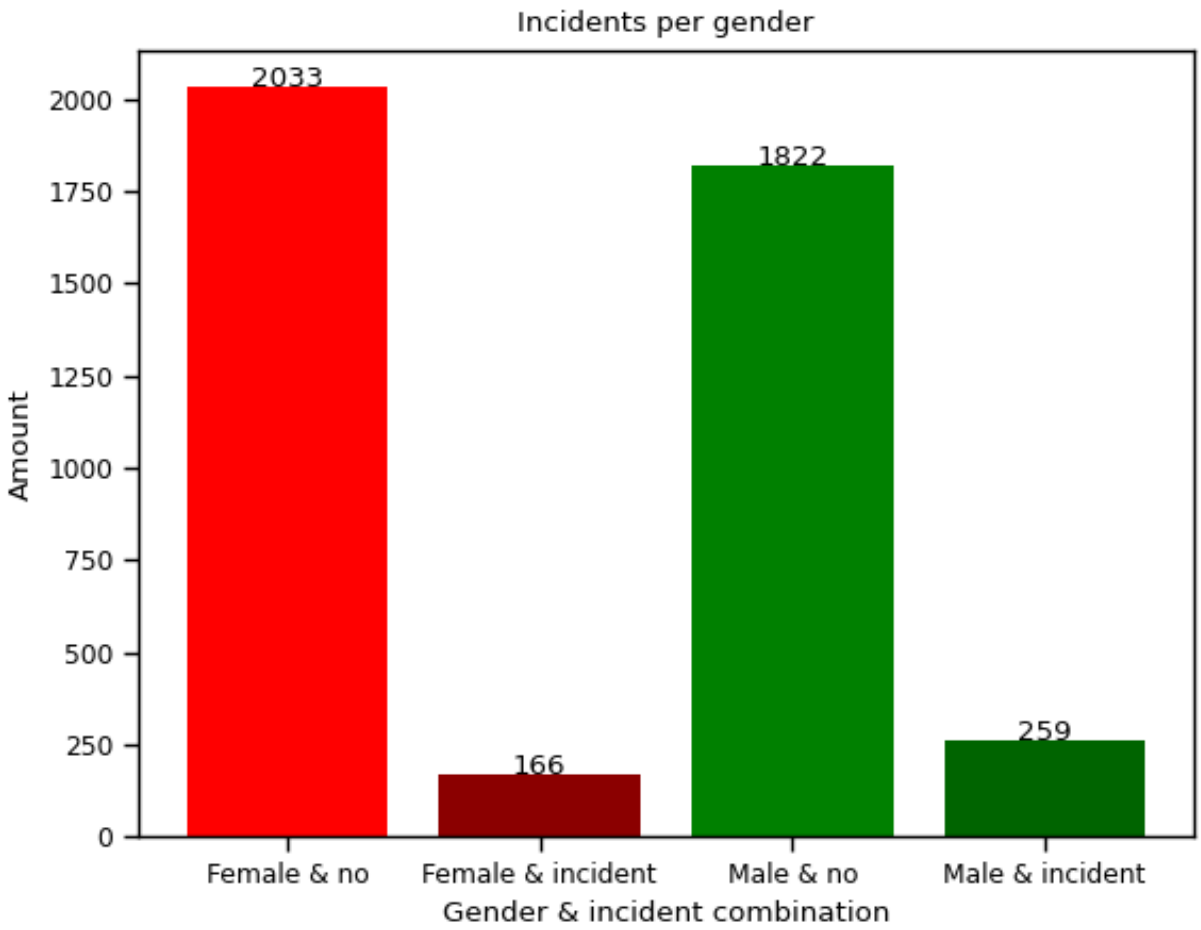

Figure 6: Incident and gender distribution

5.3 Training, testing, and validation splits

Table 1 shows the distribution of the data between the training, testing, and validation sets. To minimize the gender bias that is introduced during training, the ratio of positive and negative examples in the training set is the same for males and females. The remaining data is then evenly split between the testing and validation sets.

	Train			Test			Val			Total
	0	1	T	0	1	T	0	1	T	
M	1260	140	1400	281	59	340	281	59	340	2080
F	1260	140	1400	386	13	399	386	13	399	2198
T	2520	280	2800	667	72	739	667	72	739	4278

Table 1: Train, test, and validation split per gender, 0 is no incident, 1 is violence incident

6 Experimental results

This chapter delves into the experimental results of our research. We show the most important findings regarding counterfactual fairness and performance first, which have been obtained by testing on the augmented test set. Then, we visualize the results regarding predictive parity and performance, which we obtained from testing on the original test set. Both the counterfactual fairness and predictive parity results are shown in a single table, to allow for easier comparison. Finally, we show the overall scores regarding the performance and fairness of all models in a complete overview per model.

6.1 Counterfactual fairness evaluation

Table 2 shows the results concerning counterfactual fairness and the performance of the different models on the augmented test set. The augmented set is the original test set combined with the same test set with all gendered words swapped. As such, the model classifies each test text twice, but with different genders, which can be used to gain insights into counterfactual fairness. Furthermore, each model is trained in three different ways: on the original training set, the neutralized training set, and the augmented training set. Additionally, the ROC method was applied to the results of each model from the original setting. This is shown in the 'setting' column. The mismatch ratio shows how often the model has a different outcome for a pair of gender-swapped examples, where a lower score means a fairer performance. The true positive rate ratio (TPRR) is the ratio of true positives for males compared to females. The bias direction is shown by the F (female) or M (male) next to the number, meaning that if the TPR for one gender is higher, it is biased towards this gender. Here, a higher score means a fairer performance. This is the same for the false positive rate ratio (FPRR) and area under curve ratio (AUC-ratio). The last two columns show the F1-score and overall AUC respectively, where a higher score means a better performance. Scores that show a significant increase compared to the original setting are highlighted in dark green, while scores that show a significant decrease are highlighted in dark red, to provide a clearer overview of the biggest changes. Scores that show a small improvement are highlighted in light green, while scores that show a small decrease are highlighted in light red.

6.1.1 Original setting

BERTje and MedRoBERTaNL perform similarly in the original setting, with BERTje scoring slightly better on counterfactual fairness. The bias direction of both models is towards females. MBERT scores significantly worse performance-wise but does better for counterfactual fairness, with a slight bias towards males. The SVM model does significantly better regarding performance than all other models, and shows slightly worse counterfactual fairness than MBERT but better than BERTje and MedRoBERTaNL, with the bias direction being mainly towards males.

6.1.2 Neutralized setting

For the neutralized setting, BERTje shows a minor performance increase and a significant counterfactual fairness increase with a lower mismatch ratio and higher TPRR and FPRR. MedRoBERTaNL shows a slight performance decrease, but a significant counterfactual fairness increase as well. MBERT shows a slight performance increase, but a slight counterfactual fairness decrease because of a lower FPRR. SVM scores slightly worse on performance, but shows a significant decrease in fairness with a lower TPRR and FPRR as well as a slightly higher mismatch ratio.

6.1.3 Bias mitigation: data augmentation

Compared to the original setting, for the augmented setting BERTje shows a worse performance, but a significant increase in counterfactual fairness with improvements on all metrics. MedRoBERTaNL also performs worse but shows significantly increased counterfactual fairness with improvements on all metrics. MBERT shows a large performance improvement, with a minor counterfactual fairness decrease. For the SVM model, the performance improves marginally, and the counterfactual fairness improves mainly on FPRR.

6.1.4 Bias mitigation: Reject Option Classification

Compared to the original setting, BERTje shows very small changes after applying the ROC method. MedRoBERTaNL shows an improvement in both performance and counterfactual fairness, though the mismatch ratio is higher. MBERT shows much better performance after the ROC is applied, and fairness increases in TPRR but decreases in both mismatch ratio and FPRR. Finally, the SVM model shows significant improvements in both performance and fairness, except for the higher mismatch ratio.

	Setting	Counterfactual Fairness				Performance	
		Mismatch Ratio%↓	TPRR ↑	FPRR↑	AUC-Ratio↑	F1%↑	AUC%↑
BERTje	orig.	2.62	0.90_F	0.79_F	0.98_F	63.78	64.08
	neutr.	1.78	0.99_F	1.00	0.99_M	63.99	64.59
	BM: aug.	1.41	0.95_M	0.98_M	0.99_M	62.14	60.95
	BM: ROC	2.95	0.91_F	0.79_F	0.99_F	62.83	63.60
MedRoBERTaNL	orig.	2.86	0.79_F	0.70_F	0.96_F	64.01	63.47
	neutr.	0.92	0.94_F	0.97_M	0.98_F	62.59	62.04
	BM: aug.	0.54	1.00	0.98_M	1.00	62.03	60.83
	BM: ROC	3.92	0.97_F	0.85_F	0.97_F	64.19	66.35
MBERT	orig.	0.46	0.86_M	1.00	0.99_M	55.47	54.31
	neutr.	0.62	0.88_M	0.83_M	0.99_M	58.53	56.62
	BM: aug.	1.43	0.96_M	0.94_M	0.99_M	62.06	60.51
	BM: ROC	2.66	0.99_F	0.93_F	1.00	61.86	64.95
Doc2Vec SVM	orig.	0.46	0.98_F	0.78_M	0.99_F	78.02	72.05
	neutr.	0.70	0.89_M	0.67_M	0.97_M	75.79	69.64
	BM: aug.	0.27	0.99_M	0.96_M	1.00	78.22	72.59
	BM: ROC	2.19	0.98_F	0.95_M	0.98_M	78.82	86.82

Table 2: Counterfactual fairness in violence risk assessment from clinical notes: comparison of fairness measures of different models that are evaluated on counterfactual pairs. BM: bias mitigation, orig.: original, aug.: augmentation, neutr.: neutralization, ROC: Reject Option Classification, M: bias towards males (higher score for males), F: biased towards females, ↑: higher score is better, ↓: lower score is better, colors indicate scores compared to orig setting, dark green: significant improvement, light green: small improvement, light red: small decrease, dark red: significant decrease.

6.2 Equalized odds evaluation

Table 3 shows the predictive parity and performance results on the original test dataset without augmentation. This table has the same columns as 2, except that the mismatch ratio is not included since this cannot be calculated on just the original set. Because this is only on the original set, this table shows the predictive parity.

6.2.1 Original setting

As with the augmented test results, MBERT scores the worst performance-wise. However, MBERT also scores very low on predictive parity with a very low TPRR. BERTje has slightly higher performance scores than MedRoBERTaNL, but scores worse on predictive parity. Lastly, the SVM model has the best performance scores, but slightly worse predictive parity results across all metrics than MedRoBERTaNL.

6.2.2 Neutralized setting

In the neutralized setting, BERTje shows a small performance improvement, but a significant predictive parity increase. MedRoBERTaNL shows a minor performance decrease and a decrease in TPRR, but good improvements in FPRR. MBERT performs slightly better, and shows a slightly higher TPRR. Finally, the SVM model shows fairness improvements in all categories at the cost of a slight performance decrease.

6.2.3 Bias mitigation: data augmentation

BERTje performs slightly worse in the augmented setting compared to the original setting regarding performance, but the overall fairness stays roughly the same with a decrease in TPRR but an increase in FPRR. MedRoBERTaNL shows a small increase in fairness at the cost of a small decrease in performance, but no significant changes. MBERT shows significant improvements in both performance and fairness, scoring similarly to BERTje and MedRoBERTaNL performance-wise, and sharply increases in both TPRR and FPRR. Again, the SVM model performs best but shows no significant fairness increase and a marginal performance increase.

6.2.4 Bias mitigation: Reject Option Classification

After applying the ROC, BERTje scores largely the same as in the original setting. MedRoBERTaNL shows a performance increase and a small decrease in FPRR. MBERT shows a large improvement in both performance and fairness with a much higher TPRR. Finally, the SVM model performs best and shows an increase in fairness because of a higher FPRR and a sizeable performance increase in AUC.

	Setting	Fairness - Predictive parity			Performance	
		TPRR ↑	FPRR↑	AUC-Ratio↑	F1%↑	AUC%↑
BERTje	orig	0.64_M	0.80_F	0.89_M	62.77	63.11
	neutr.	0.69_M	0.95_M	0.91_M	63.46	64.02
	BM:aug.	0.45_M	0.98_F	0.86_M	62.41	61.27
	BM:ROC	0.58_M	0.83_F	0.87_M	62.09	62.94
MedRoBERTaNL	orig	0.97_F	0.77_F	0.99_M	62.51	61.82
	neutr.	0.83_M	1.00	0.96_M	62.00	61.34
	BM:aug.	0.97_F	0.84_F	1.00	62.03	60.83
	BM:ROC	0.97_F	0.72_F	0.99_M	62.74	64.66
MBERT	orig	0.27_M	0.73_F	0.92_M	56.03	54.65
	neutr.	0.35_M	0.76_M	0.90_M	58.76	56.95
	BM:aug.	0.79_M	0.99_F	0.95_M	62.49	60.84
	BM:ROC	0.77_M	0.77_F	0.90_M	61.77	64.98
Doc2Vec SVM	orig	0.89_F	0.69_F	0.96_F	77.96	72.05
	neutr.	0.94_M	0.83_M	0.98_M	76.23	70.01
	BM:aug.	0.86_F	0.73_F	0.95_F	78.28	72.66
	BM:ROC	0.90_F	0.85_F	0.96_F	78.15	86.55

Table 3: fairness in violence risk assessment from clinical notes: comparison of fairness measures of different models that are evaluated on the original test set. BM: bias mitigation, orig. : original, aug. : augmentation, neutr. : neutralization, M: bias towards males (higher score for males), F: biased towards females, ↑: higher score is better, ↓: lower score is better, colors indicate scores compared to orig setting, dark green: significant improvement, light green: small improvement, light red: small decrease, dark red: significant decrease.

6.3 Complete performance overview

6.3.1 BERTje

Table 4 shows the average performance of BERTje in all settings, on both the augmented and original test sets. Precision, recall, and F1 are split between the positive class and the negative class. AUC, TPR, and FPR are split between males and females. Lastly, the number of mismatches is shown. Overall, the model performs significantly worse in correctly classifying positive examples compared to negative examples. In the neutralized setting, there are slight performance improvements, and the differences between males and females become slightly smaller. The amount of mismatches decreases as well. For the augmented setting, there is a performance decrease overall, but the performance for males and females becomes almost the same, and the amount of mismatches decreases. The impact of the ROC is quite insignificant, with a small decrease in performance overall, no significant improvement in performance differences between males and females, and a slightly higher number of mismatches. Overall, the model performs slightly better on the augmented test set. The differences between males and females are also larger on the

original test set in all settings.

Setting	Orig.		Neutr.		Aug.		ROC	
Test set	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.
Accuracy	0.87	0.87	0.87	0.87	0.88	0.88	0.86	0.86
Precision pos.	0.34	0.33	0.34	0.34	0.36	0.36	0.32	0.30
Recall pos.	0.36	0.34	0.37	0.36	0.27	0.28	0.36	0.34
F1 score pos.	0.35	0.33	0.35	0.34	0.31	0.31	0.33	0.32
Precision neg.	0.93	0.93	0.93	0.93	0.92	0.92	0.93	0.93
Recall neg.	0.93	0.92	0.92	0.92	0.95	0.95	0.92	0.91
F1 score neg.	0.93	0.93	0.93	0.93	0.94	0.94	0.92	0.92
F1 unweighted	0.64	0.63	0.64	0.63	0.62	0.62	0.63	0.62
AUC	0.64	0.63	0.64	0.64	0.61	0.61	0.64	0.63
AUC male	0.64	0.65	0.64	0.65	0.61	0.63	0.63	0.65
AUC female	0.65	0.57	0.64	0.59	0.61	0.54	0.64	0.56
TPR male	0.34	0.36	0.36	0.38	0.28	0.31	0.34	0.37
TPR female	0.38	0.23	0.37	0.26	0.26	0.14	0.37	0.22
FPR male	0.07	0.07	0.08	0.08	0.05	0.05	0.07	0.08
FPR female	0.08	0.08	0.08	0.08	0.05	0.05	0.09	0.09
Mismatches	19.4	-	13.2	-	10.4	-	21.8	-

Table 4: Average performance of BERTje in all settings on the augmented and original test sets, colors indicate scores compared to the original setting, green: significant improvement, red: significant decrease.

6.3.2 MedRoBERTaNL

Table 5 shows the average performance of MedRoBERTaNL on both the augmented and original test sets. Overall, the performance is in line with BERTje. This model also performs significantly worse regarding the positive class. In the original setting, there are also differences in performance between males and females, with better performance for females, as well as overall more positive predictions for females as shown with the higher TPR and FPR. These differences become smaller in the other settings. The biggest change for the neutralized and augmented settings is the lower number of mismatches and the smaller differences between male and female performance. Applying the ROC method increases the recall for the positive class significantly, at the cost of a small decrease in precision for the positive class. Also, the TPR for both genders increases significantly, at the cost of a smaller increase in FPR for both genders. However, the amount of mismatches also increases.

Setting	Orig.		Neutr.		Aug.		ROC	
Test set	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.
Accuracy	0.88	0.87	0.87	0.87	0.88	0.88	0.86	0.85
Precision pos.	0.37	0.35	0.35	0.34	0.36	0.36	0.33	0.31
Recall pos.	0.33	0.30	0.30	0.29	0.27	0.27	0.42	0.39
F1 score pos.	0.35	0.32	0.32	0.31	0.31	0.31	0.36	0.34
Precision neg.	0.93	0.93	0.93	0.92	0.92	0.92	0.94	0.93
Recall neg.	0.94	0.94	0.94	0.94	0.95	0.95	0.91	0.90
F1 score neg.	0.93	0.93	0.93	0.93	0.94	0.94	0.92	0.92
F1 unweighted	0.64	0.63	0.63	0.62	0.62	0.62	0.64	0.63
AUC	0.63	0.62	0.62	0.61	0.61	0.61	0.66	0.65
AUC male	0.62	0.62	0.61	0.62	0.61	0.61	0.65	0.65
AUC female	0.65	0.62	0.63	0.59	0.61	0.61	0.67	0.65
TPR male	0.29	0.30	0.29	0.30	0.27	0.27	0.39	0.39
TPR female	0.37	0.31	0.31	0.25	0.27	0.28	0.46	0.40
FPR male	0.05	0.05	0.06	0.06	0.05	0.05	0.08	0.08
FPR female	0.08	0.07	0.06	0.06	0.05	0.06	0.11	0.11
Mismatches	21.2	-	6.8	-	4.0	-	29.0	-

Table 5: Average performance of MedRoBERTaNL in all settings on the augmented and original test sets, colors indicate scores compared to the original setting, green: significant improvement, red: significant decrease.

6.3.3 MBERT

MBERT performs the worst of all models in the original setting, especially for the positive class. For the augmented test set, the performance difference between both genders is very small. In the neutralized setting, performance improves slightly, but the overall performance is still low. The augmented setting shows a large improvement in performance in most metrics, except for a decrease in precision for the positive class and a larger amount of mismatches. Applying the ROC increases the recall and F1-score for the positive class, at the cost of precision for the positive class. Overall accuracy and recall for the negative class also decreased. There is a sharp increase in AUC and TPR for both genders and unweighted F1, but also in FPR and mismatches.

Setting	Orig.		Neutr.		Aug.		ROC	
Test set	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.
Accuracy	0.91	0.91	0.90	0.90	0.89	0.89	0.84	0.84
Precision pos.	0.62	0.63	0.50	0.51	0.39	0.41	0.28	0.28
Recall pos.	0.09	0.10	0.15	0.16	0.26	0.26	0.42	0.42
F1 score pos.	0.16	0.17	0.22	0.23	0.30	0.31	0.33	0.33
Precision neg.	0.91	0.91	0.91	0.92	0.92	0.92	0.93	0.93
Recall neg.	0.99	0.99	0.99	0.98	0.95	0.96	0.88	0.88
F1 score neg.	0.95	0.95	0.95	0.95	0.94	0.94	0.91	0.91
F1 unweighted	0.55	0.56	0.59	0.59	0.62	0.62	0.62	0.62
AUC	0.54	0.55	0.57	0.57	0.61	0.61	0.65	0.65
AUC male	0.55	0.55	0.57	0.58	0.61	0.61	0.65	0.67
AUC female	0.54	0.51	0.56	0.52	0.60	0.59	0.65	0.60
TPR male	0.10	0.12	0.16	0.18	0.26	0.27	0.41	0.44
TPR female	0.09	0.03	0.14	0.06	0.25	0.22	0.42	0.34
FPR male	0.01	0.01	0.02	0.02	0.05	0.04	0.11	0.12
FPR female	0.01	0.01	0.01	0.01	0.04	0.04	0.10	0.13
Mismatches	3.4	-	4.6	-	10.6	-	19.8	-

Table 6: Average performance of MBERT in all settings on the augmented and original test sets, colors indicate scores compared to the original setting, green: significant improvement, red: significant decrease.

6.3.4 SVM

The SVM model performs best in all settings compared to the other models. Neutralization has a rather small impact. Overall, the performance is slightly worse, mainly with a decrease in TPR for females and an increase in mismatches. Augmentation also has limited effects, with the biggest change being a lower number of mismatches. The biggest changes are seen after applying ROC. For the positive class, recall is increased at the cost of a precision decrease. For the negative class, this is the opposite. A large increase in AUC for both genders can be seen as well. Furthermore, the TPR for both genders shows large increases, at the cost of a higher FPR. Finally, the amount of mismatches increases significantly.

Setting	Orig.		Neutr.		Aug.		ROC	
Test set	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.	Aug.	Orig.
Accuracy	0.94	0.94	0.94	0.94	0.94	0.94	0.90	0.90
Precision pos.	0.89	0.89	0.91	0.91	0.87	0.87	0.51	0.50
Recall pos.	0.45	0.45	0.40	0.41	0.46	0.46	0.83	0.82
F1 score pos.	0.59	0.59	0.55	0.56	0.60	0.60	0.63	0.62
Precision neg.	0.94	0.94	0.94	0.94	0.94	0.94	0.98	0.98
Recall neg.	0.99	0.99	1.00	1.00	0.99	0.99	0.85	0.91
F1 score neg.	0.97	0.97	0.97	0.97	0.97	0.97	0.95	0.94
F1 unweighted	0.78	0.78	0.76	0.76	0.78	0.78	0.79	0.78
AUC	0.72	0.72	0.70	0.70	0.73	0.73	0.87	0.87
AUC male	0.72	0.72	0.71	0.70	0.73	0.72	0.89	0.86
AUC female	0.72	0.74	0.69	0.69	0.73	0.76	0.87	0.90
TPR male	0.44	0.44	0.42	0.41	0.46	0.45	0.82	0.81
TPR female	0.45	0.49	0.38	0.38	0.46	0.52	0.84	0.89
FPR male	0.01	0.01	0.01	0.01	0.01	0.01	0.09	0.09
FPR female	0.01	0.01	0.00	0.00	0.01	0.01	0.08	0.10
Mismatches	3.4	-	5.2	-	2.0	-	16.2	-

Table 7: Average performance of SVM model in all settings on the augmented and original test sets, colors indicate scores compared to the original setting, green: significant improvement, red: significant decrease.

7 Discussion

This chapter will analyze the results that are shown in the results section. We will further explore model performance, fairness considerations, bias mitigation techniques, as well as limitations.

7.1 Model performance

In evaluating model performance, notable distinctions emerge among the monolingual BERTje and MedRoBERTaNL, multilingual MBERT, and the SVM model. BERTje and MedRoBERTaNL performed much better than MBERT in both the original and neutralized settings, but MBERT showed significant improvement in the augmented setting. It seems that MBERT benefits from a larger fine-tuning dataset size. This holds true for both the original as well as the augmented test results. That said, the SVM model performed significantly better than the BERT-based models, likely due to its capability to handle large input sizes without significant trade-offs. For the BERT-based models, the only technique that showed promising results was chunking the texts to comply with the maximum input size. Using only the first or last part of the texts proved to lose too much relevant information, and using Longformer which allowed for longer input sizes gave a fine-tuned model with extremely low predictive power. However, the trade-off with chunking is that the models treated certain parts of the texts that are potentially irrelevant to the outcome the same as critical sections, which can lead to unwanted model behavior. These nuances show the complexity of the task and emphasize the need for tailored strategies to balance input size constraints with the preservation of essential information. Further exploration of techniques or model architectures may offer insights into optimizing the performance of BERT-based models for this specific domain.

7.2 Fairness

BERTje and MedRoBERTaNL show similar counterfactual fairness in the original setting, including the bias direction. In the neutralized setting, BERTje shows almost no bias regarding counterfactual fairness, while MedRoBERTaNL shows a slight bias. The neutralized setting is used to see if the pre-training introduces bias. For BERTje, this does not seem to be the case. For MedRoBERTaNL, a slight amount of bias can be attributed to pre-training. As for MBERT, there is also counterfactual fairness bias in the neutralized setting, which can be attributed to pre-training, though it is difficult to make strong conclusions since the model performance is bad overall. Since the original settings for the BERTje and MedRoBERTaNL show more counterfactual fairness bias than the neutralized settings, it can be concluded that bias is also introduced during model training, with the models being more likely to predict a positive outcome for females than for males. As for the SVM model, the input here is a Doc2Vec model that is trained on the entire dataset. As such, counterfactual fairness bias in the neutralized setting can be attributed to bias that is present in the dataset. Interestingly, this model is more likely to predict a positive outcome for males. This makes sense, given that in the entire dataset, the violent incident rate for males is higher than for females. However, both

BERTje and MedRoBERTaNL have a higher propensity to predict a positive outcome for females, while for MBERT it is the opposite, though slightly. The counterfactual fairness becomes much better in both the neutralized and augmented settings for BERTje and MedRoBERTaNL. As such, it seems that the bias regarding counterfactual fairness is introduced during fine-tuning.

As for predictive parity, BERTje shows slightly more bias in the original setting than in the neutralized setting, though the bias in the neutralized setting is quite high. So, it can be concluded that predictive parity bias is mostly introduced during pre-training for BERTje, with a slight bias introduced during fine-tuning. Interestingly, the bias direction is mainly towards males. For MedRoBERTaNL, the bias direction in the original setting is toward females, while in the neutralized setting it is toward males. As such, it can be concluded that the bias towards males is introduced by the pre-training and the bias towards females during fine-tuning. As for MBERT, the bias towards males is mainly caused by the pre-training, and a bias towards females is introduced during fine-tuning. As for the SVM model, a bias towards males can be seen in the neutralized setting, which can be attributed to the Word2Vec embeddings. However, a bias towards females can be seen in the original setting. This seems to be caused during the SVM model training. Combining these findings from all models, the UMCU dataset seems to have a predictive parity bias towards females, whereas pre-training bias is mostly towards males.

7.2.1 Source of bias and bias magnitude

Overall, to answer sub-research question one 'What is the main source of bias', it can be concluded that there is no single source of bias. Bias is introduced during pre-training and fine-tuning for the BERT-based models, and during the Doc2Vec as well as SVM training. However, counterfactual fairness bias towards females is mostly introduced during fine-tuning, whereas predictive parity bias towards males is introduced during pre-training and towards females during fine-tuning.

Considering sub-research question 3, 'Is there a difference in terms of the magnitude of bias that pre-trained embeddings have', the multilingual MBERT seems to have less pre-training counterfactual fairness bias than both monolingual models BERTje and MedRoBERTaNL. For predictive parity, this is the other way around, with MBERT showing significantly more bias. Comparing the general BERTje with the domain-specific pre-trained MedRoBERTaNL, we observe a slightly higher amount of counterfactual fairness bias. However, BERTje shows more predictive parity bias. Overall, all the pre-trained models show a form of bias, but it differs per model how severe this bias is, and also what kind of bias is most prevalent.

7.2.2 Impact on downstream task

To answer sub-research question two 'How does this bias impact the downstream task of violence risk assessment', the results on the original test set reflect the real-world situation the best. BERTje and MBERT both predict males to be involved in a violent incident more often, whereas MedRoBERTaNL and SVM predict females to be involved more often.

7.3 Bias mitigation methods

Overall, data augmentation worked well in mitigating counterfactual fairness bias for both BERTje and MedRoBERTaNL. Both models performed almost the same for both genders, with BERTje being slightly less fair. However, it did mean a slight trade-off regarding performance, with both models showing a decrease in F1-score and AUC. As for predictive parity, data augmentation for BERTje led to mixed results, with overall fairness staying roughly the same, while for MedRoBERTaNL there was a slight predictive parity increase. For MBERT, data augmentation showed mixed results regarding counterfactual fairness. Certain aspects improved while others decreased. However, the performance increased quite significantly. This might be attributable to the larger amount of data that was used for fine-tuning. Since MBERT is a larger model than BERTje and MedRoBERTaNL, it is possible that it needs more data before it can generalize well enough to make good predictions. Predictive parity also improved significantly. As for the SVM model, data augmentation led to an increase in counterfactual fairness, while also showing a minor performance increase. However, predictive parity stayed largely the same.

As for the Reject Option Classification method, for BERTje the effects were minimal. Predictive parity got slightly worse, but the other changes are negligible. MedRoBERTaNL saw an increase in counterfactual fairness, but a slight decrease in predictive parity, with the performance improving slightly. As for MBERT, the counterfactual fairness decreased, but the predictive parity increased. Also, the performance scores improved significantly. For the SVM model, counterfactual fairness and predictive parity improved slightly because of an FPRR increase, but the mismatch ratio also got significantly higher. Additionally, the AUC score increased significantly.

It is interesting to note that for all models, the mismatch ratio increased after applying ROC, and mostly decreased because of data augmentation. The results show differences per model, where certain models react differently to the bias mitigation methods. Furthermore, a fairness improvement in one category can lead to a decrease in another. This shows that it is important to carefully consider which metrics to focus on.

7.4 Limitations

While the results show promising insights into gender bias and gender bias mitigation, there are also some limitations to this research that should be considered when interpreting the findings.

7.4.1 Generalizability

First, it is difficult to generalize the results. What proves efficient for psychiatric notes and violence risk assessment might not work for other domains or datasets. From our experimental results, we see that the different models already have different outcomes on the same task. As such, it would be even more difficult to generalize these results for other domains. Furthermore, dataset characteristics should also be taken into account. For example, our dataset has a large class imbalance, which means certain techniques might work better than others. Finally, violence risk assessment is a complex task. What

might work for the psychiatric domain, might not work for other domains where violence risk assessment is used, such as law enforcement.

7.4.2 Domain-specific model limitations

Furthermore, the use of the domain-specific MedRoBERTaNL model, pre-trained on a broader medical dataset, introduces a nuanced limitation. While the goal was to leverage its enhanced understanding of medical terms, the difference between the training domain (general medical data) and the target domain (psychiatry) poses challenges. Given that the training domain and the target domain are not exactly the same, it is difficult to make strong conclusions about domain-specific models. However, the training domain of MedRoBERTaNL was the closest to our target domain for pre-trained Dutch language models. That said, an interesting part of our findings is that MedRoBERTaNL is pre-trained on a significantly smaller dataset than BERTje, yet it performs largely the same for our task.

7.4.3 Manual selection of gendered terms

For both the augmentation and neutralization, we used a list of gendered terms. This list was constructed by carefully going over the individual text entries. However, this was a manual task, and therefore it is possible that certain terms were missed. This means that it is possible that the augmented and neutralized datasets contain wrongly gendered words, which in turn can affect gender bias.

7.5 Ethical considerations

Given that we used a classified dataset with sensitive patient data, and that the classification objective is violence risk assessment, some ethical considerations should be taken into account. First off, data security and anonymization are important. The dataset that we used had been anonymized already before we did any data processing in order to minimize personally identifiable data. Furthermore, ethical approval had been gained for the overarching project this research fits into. Finally, if the classification we discussed in this research were to actually be implemented for real-world usage, further considerations should be made. There is potential stigmatization associated with being labeled as likely to commit a violent incident, though the overall goal is to provide better care. Transparency is an important requirement if this solution were to be implemented.

8 Conclusion

In concluding this research, we reflect on an exploration into gender bias and bias mitigation in the context of violence risk assessment models. Our research shows the complex dynamics involved in the use of NLP models in the mental healthcare domain. The most important findings of this research are that monolingual BERT-based models perform better than multilingual ones, and that domain-specific pre-trained models perform on par with general pre-trained models. Our research also concludes that the classical machine learning model SVM is more suited for the task at hand, which corroborates earlier works on this topic. This is caused by the fact that the SVM model can handle larger input sizes, whereas the BERT-based models need an additional step where the input is chunked, which compromises the model performance significantly. However, all these models showed bias in some form. For the BERT-based models, this was partly introduced during pre-training and further increased during fine-tuning on our dataset. For the SVM model, the bias was entirely caused by the dataset. In conclusion, both pre-training, fine-tuning, and model training introduce bias. However, there are methods to deal with these biases. Data augmentation works well to reduce counterfactual fairness bias and to a lesser extent predictive parity bias. However, for certain models, this comes at the cost of performance, but for others, it increases performance. Furthermore, the post-processing technique Reject Option Classification can also be used for some models to improve counterfactual fairness and predictive parity, but for others, it will decrease this. The same holds for performance measures; in certain situations, these improve, while in others they decrease.

In summary, bias mitigation is a delicate issue. It is important to decide what metrics to optimize for since improvements in certain areas often come with trade-offs in others. This is domain-specific, so there is no one-size-fits-all solution. Careful consideration should be put into deciding what is most important. While the study's findings highlight the results for violence risk assessment in psychiatric care, we caution against broad generalizations to other domains or datasets. The task-specific nature of the study necessitates careful consideration of the context in interpreting and applying the results.

8.1 Future work

As for future research, looking into fairness and bias mitigation for other classical machine learning approaches for this domain is an interesting topic. Since the SVM model outperformed the BERT-based models, and it is possible to reduce its bias as well, it would be interesting to see how other methods would perform. Furthermore, similar research could be done on other domains to look into the generalizability of the findings. Furthermore, while MedRoBERTaNL's training domain is not the exact same as our target domain, it did perform on par with BERTje. This highlights the potential of domain-specific models, and their usage could be of interest in future research, even if the domain is not an exact fit.

Additionally, an interesting approach would feature extraction. In this approach, the word embeddings, or features, are extracted from the BERT-based models, and used as input for another machine learning approach, e.g. SVM. This can either be done directly

after pre-training or after first fine-tuning the embeddings on the task-specific dataset, but fine-tuning on the task-specific data is the most promising method. This would work similarly to the Doc2Vec embeddings we use as input for the SVM model. However, for long sequences, the outputs of different chunks should be pooled together to get the representation of the entire text.

Furthermore, given the sensitive subject, transparency is an important factor that should be considered. It would be interesting to research what the key factors are for a positive or negative classification. We did touch on this subject with the data augmentation, but a further analysis could provide more insights.

References

Adamou, M., Antoniou, G., Greasidou, E., Lagani, V., Charonyktakis, P., & Tsamardinos, I. (2018). Mining free-text medical notes for suicide risk assessment. In *Proceedings of the 10th hellenic conference on artificial intelligence* (pp. 1–8).

Almvik, R., Woods, P., & Rasmussen, K. (2000). The brøset violence checklist: sensitivity, specificity, and interrater reliability. *Journal of interpersonal violence*, *15*(12), 1284–1296.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Balkir, E., Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2022). Challenges in applying explainability methods to improve the fairness of nlp models. *arXiv preprint arXiv:2206.03945*.

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016b). Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016a). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. In *International conference on machine learning* (pp. 803–811).

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121–167.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., ... others (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chen, X., Sykora, M. D., Jackson, T. W., & Elayan, S. (2018). What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion proceedings of the the web conference 2018* (pp. 1653–1660).

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 7801–7808).

Dev, S., Li, T., Phillips, J. M., & Srikumar, V. (2020). On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 7659–7666).

Devlin, J. (2018, Nov). *Bert/multilingual.md at a9ba4b8d7704c1ae18d1b28c56c0430d41407eb1 · google-research/bert.* Retrieved from `https://github.com/google-research/bert/blob/a9ba4b8d7704c1ae18d1b28c56c0430d41407eb1/multilingual.md`

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: getting specific about being dynamic. *Psychology, Public Policy, and Law*, *11*(3), 347.

Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big data and discrimination: perils, promises and solutions. a systematic review. *Journal of Big Data*, *6*(1), 1–27.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 259–268).

Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (2018). " let me tell you about your mental health!" contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 753–762).

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hatton, C. M., Paton, L. W., McMillan, D., Cussens, J., Gilbody, S., & Tiffin, P. A. (2019). Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. *Journal of affective disorders*, *246*, 857–860.

Iozzino, L., Ferrari, C., Large, M., Nielssen, O., & De Girolamo, G. (2015). Prevalence and risk factors of violence by psychiatric acute inpatients: a systematic review and meta-analysis. *PloS one*, *10*(6), e0128536.

Ivgi, M., Shaham, U., & Berant, J. (2023). Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, *11*, 284–299.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 ieee 12th international conference on data mining* (pp. 924–929).

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196).

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, L., Wang, Z., & Zhang, Y. (2017). Extracting depression symptoms from social networks and web blogs via text mining. In *Bioinformatics research and applications: 13th international symposium, isbra 2017, honolulu, hi, usa, may 29–june 2, 2017, proceedings 13* (pp. 325–330).

Ma, P., Wang, S., & Liu, J. (2020). Metamorphic testing and certified mitigation of fairness violations in nlp models. In *Ijcai* (pp. 458–465).

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Menger, V., Scheepers, F., & Spruit, M. (2018). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, *8*(6), 981.

Miaschi, A., & Dell'Orletta, F. (2020). Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th workshop on representation learning for nlp* (pp. 110–119).

Mosteiro, P., Rijcken, E., Zervanou, K., Kaymak, U., Scheepers, F., & Spruit, M. (2022). Machine learning for violence risk assessment using dutch clinical notes. *arXiv preprint arXiv:2204.13535*.

Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681.

Nobles, A. L., Glenn, J. J., Kowsari, K., Teachman, B. A., & Barnes, L. E. (2018). Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–11).

Park, A., Conway, M., & Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. *Computers in human behavior*, *78*, 98–112.

Petersen, F., Mukherjee, D., Sun, Y., & Yurochkin, M. (2021). Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, *34*, 25944–25955.

Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019). Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop* (pp. 81–88).

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, *117*, 139–147.

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current directions in psychological science*, *20*(1), 38–42.

Sogancioglu, G., Mijsters, F., van Uden, A., & Peperzak, J. (2022). Gender bias in (non)-contextual clinical word embeddings for stereotypical medical categories. *arXiv preprint arXiv:2208.01341*.

Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, *471*, 216–232.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th china national conference, ccl 2019, kunming, china, october 18–20, 2019, proceedings 18* (pp. 194–206).

Tsamardinos, I., Greasidou, E., & Borboudakis, G. (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine learning*, *107*, 1895–1922.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Verkijk, S., & Vossen, P. (2021). Medroberta. nl: a language model for dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, *11*, 141–159.

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8919–8928).

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., ... Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 ieee/acm international conference on advances in social networks analysis and mining 2017* (pp. 1191–1198).

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the acm conference on health, inference, and learning* (pp. 110–120).

Zhu, X. J. (2005). Semi-supervised learning literature survey.

A Appendix A: Code

All the code that has been used for this project is uploaded to a GitHub repository: https://github.com/JoppeKo/VRAThesisJoppe