Utrecht University

The Role of Gender: Gender Fairness in the Detection of Depression Symptoms on Social Media

by

Lara Gierschmann

6374875

Submitted to the Artificial Intelligence Graduate Program
in partial fulfillment of the requirements for the degree of
Master of Science

Graduate Program in Artificial Intelligence
Utrecht University
2024

The Role of Gender: Gender Fairness in the Detection of Depression Symptoms on Social Media

APPROVED BY:

dr. Heysem Kaya . . . .
(Thesis Supervisor)

Gizem Soğancıoğlu MSc. . . . .
(Thesis Co-supervisor)

dr. Dong Nguyen . . . .

DATE OF APPROVAL: 17.07.2024

ACKNOWLEDGEMENTS

Thank you to my supervisors, Heysem Kaya and Gizem Soğancıoğlu, for the helpful insights and feedback throughout this project. Your weekly comments helped shape and improve this research, and the tea was always very appreciated! Many thanks to Dong Nguyen as well, for taking the time to read this work and for providing many helpful comments on the proposal. Moreover, I want to thank the rest of the thesis group for responsible AI in mental health for the support and exchange, and my thesis study group, who helped keep me motivated and provided helpful ideas and brainstorming sessions. Finally, a special thank you to my friends, Johanna and Bela, for reviewing this thesis.

ABSTRACT

The Role of Gender: Gender Fairness in the Detection of Depression Symptoms on Social Media

AI systems for depression detection on social media have been continuously improving their performance, showing that meaningful patterns can be found in the data. While many machine learning models used to detect depression are opaque, models predicting depression symptoms can often provide more explainability. Previous research has shown that some depression detection datasets with data collected from social media exhibit gender biases, but no studies have investigated gender bias for a dataset annotated with depression symptoms yet. Therefore, this thesis aims to investigate the extent to which gender bias is present in the BDI-Sen dataset, evaluate classifier performance across different genders, and whether existing gender bias can be mitigated. A statistical analysis reveals that the dataset shows some gender bias, reflecting gender differences in depression symptoms. Analysis of mentalBERT classifiers trained on the dataset identifies several biases across the different symptoms, particularly in terms of predictive equality, with the majority of the bias favoring males. To address these biases, data augmentation strategies such as synonym replacement, back-translation, and oversampling were applied. These methods helped reduce the bias but did not remove it completely. Future research could implement different bias mitigation techniques to reduce the bias, and investigate gender bias in depression symptom detection on larger datasets or datasets annotated for symptoms from a different questionnaire, such as the PHQ-9.

Keywords: depression symptom detection, social media, mentalBERT, BDI-Sen, AI fairness, gender bias, bias mitigation

NEDERLANDSE ABSTRACT

De Rol van Geslacht: Genderrechtvaardigheid bij de Detectie van Depressiesymptomen op Sociale Media

AI-systemen voor het detecteren van depressies op sociale media zijn steeds beter geworden en laten zien dat er zinvolle patronen in de gegevens te vinden zijn. Waar veel machine learning modellen die worden gebruikt om depressie te detecteren ondoorzichtig zijn, kunnen modellen die symptomen van depressie voorspellen vaak meer transparantie verschaffen. Onderzoek heeft aangetoond dat sommige datasets voor depressiedetectie, met gegevens verzameld uit sociale media, genderbiases vertonen. Echter, er is nog geen onderzoek gedaan naar genderbias voor datasets die gericht zijn op symptomen. Deze thesis onderzoegt daarom de aanwezigheid van genderbias in de BDI-Sen dataset, de prestaties van classifiers die erop getraind zijn voor verschillene geslachten, en mogelijke manieren om deze bias te mitigeren. Een statistische analyse toont aan dat de BDI-Sen dataset enige genderbias bevat die sekseverschillen in depressiesymptomen weerspiegelt. Analyse van mentalBERT classifiers, getraind op de dataset, identificeert verschillende vertekeningen in de voorspelling van symptomen, waarbij de meeste vertekeningen in het vordeel van mannen zijn. Het toepassen van data-augmentatiestrategieën zoals synoniemvervanging, terugvertaling en oversampling helpt de bias te verminderen, maar verwijdert deze niet volledig. Toekomstig onderzoek zou verschillende technieken kunnen toepassen om bias te verminderen en de genderbias in de detectie van depressiesymptomen kunnen onderzoeken op grotere datasets of datasets die geannoteerd zijn voor symptomen uit een andere vragenlijst, zoals de PHQ-9.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

ADHD	Attention-Deficit/Hyperactivity Disorder
AI	Artificial Intelligence
AUC	Area Under the Curve
BDI-II	Beck Depression Inventory II
BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag-Of-Words
CNN	Convolutional Neural Network
CV	Cross Validation
DSM	Diagnostic and Statistical Manual of Mental Illnesses
EDA	Easy data augmentation
ELMO	Embeddings from Language Models
GABDI-CV	Gender-Annotated BDI-Sen Cross Validation
GABDI-Sen	Gender-Annotated BDI-Sen
GloVe	Global Vectors
LDA	Latent Dirichlet Allocation
LGBTQ+	Lesbian, Gay, Bisexual, Transgender, Queer
LIME	Local Interpretable Model-Agnostic Explanations
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
LSTM	Long Short-Term Memory
NLLB	No Language Left Behind
NLP	Natural Language Processing
OCD	Obsessive-Compulsive Disorder
PHQ	Patient Health Questionnaire
POS	Part Of Speech
PTSD	Post-Traumatic Stress Disorder
RNN	Recurrent Neural Network
ROC	Reject Option Classification

RSDD	Reddit Self-reported Depression Diagnosis
SMHD	Self-reported Mental Health Diagnosis
SVM	Support Vector Machine
TF-IDF	Term-Frequency Inverse-Document-Frequency
USPSTF	U.S. Preventive Services Task Force
WHO	World Health Organization

1. INTRODUCTION

Applications of artificial intelligence (AI) to the healthcare domain are becoming more and more frequent. One such application in the mental health domain is the detection of depression. AI systems have been applied to both depression detection in a clinical setting [1] and in social media [2]. Depression detection in the social media context can be especially helpful to identify individuals at risk and provide them with the help needed, as social media is used more and more for sharing personal information that allows for such detection [3]. Indeed, recent models show promising performance in depression detection for social media [2].

Mental health is a domain that requires careful consideration when applying AI to it. Therefore, the field of responsible AI is especially important. Responsible AI describes research into issues that can be caused by AI systems and aims for AI systems to work in accordance with ethical principles and human values [4]. Topics that responsible AI investigates include discrimination and bias in AI systems, a lack of transparency in black box models, and data security. Especially in the healthcare context, responsible AI is important, as an algorithm's predictions and decisions can have severe consequences. Topics here include issues regarding inconclusive evidence and the need for causal relationships, transparency, the quality and neutrality of datasets, fairness, transformative effects AI has on patient information, and who is responsible for AI decision-making [5].

While good model performance is important, many high-performing AI models in mental health require opaque machine learning methods [2]. This is an important issue in responsible AI and poses a problem for domain experts, such as therapists, who are not provided with an explanation of why a model predicts someone to be depressed or not. Especially in the mental health domain, a model's prediction can have a big impact on someone's life, and relying on a black-box model for decision-making can be problematic. To increase interpretability and create models that give experts an explanation, researchers have started building datasets that have symptoms for depression and other mental illnesses annotated per sentence [6] [7] [8]. Models trained on them can make a prediction of whether someone is depressed by detecting depression symptoms. This can not only provide higher interpretability, but also ensure that model predictions follow clinically developed and used guidelines, such as the Diagnostic and Statistical Manual of Mental Illnesses (DSM-V) or the Beck Depression Inventory II (BDI-II) questionnaire. The BDI-Sen dataset is one such dataset that allows for the prediction of BDI-II symptoms [7].

Another important aspect of responsible AI is AI fairness. As the decisions of an AI system in depression detection can influence where resources to help depressed individuals go, it is important to ensure that depression detection models are equally beneficial to different groups of individuals. However, past research in other domains has shown that this is not always the case [9] [10]. Biases such as gender and racial bias can perpetuate harmful stereotypes and lead to an unjust distribution of resources [11]. In the case of depression detection, this could lead to a systematical omission to access of healthcare facilities for specific groups, which could have severe consequences, especially for depressed individuals that experience suicidal thoughts.

Research in AI fairness for depression detection systems in social media is limited, but indicates that harmful biases are present. A study on the CLPsych2015 Twitter dataset found gender and racial biases and showed that this was largely due to imbalance in the dataset, demonstrating the influence of dataset creation in ensuring fair prediction outcomes [12]. Another study on the D-Vlog dataset with data from YouTube confirmed this, finding dataset imbalance and gender bias [13]. Moreover, the researchers showed that bias mitigation techniques can only achieve minor improvements when the pre-processing of the data is one of the reasons for the bias [13].

The limited amount of research on bias in social media datasets for depression detection and the harmful consequences these biases can have, demonstrate a need for more investigation into the topic. Specifically, no prior research has investigated gender bias in a dataset annotated for symptoms. Researching whether bias in such a dataset exists and how it can be mitigated is especially interesting, as it can provide further insights into differences in symptoms between males and females and how this impacts depression detection systems. This project aims to bring attention to this gap in research and provide further insights into the field of responsible AI for mental health by taking a closer look at the gender fairness in the BDI-Sen dataset [7]. Specifically, its goal is to investigate the extent to which gender bias is present in the BDI-Sen dataset, how a classifier trained on it performs for different genders, and whether any present gender bias can be mitigated.

This thesis begins by discussing relevant literature and background for this research, introducing topics such as depression as well as natural language processing (NLP) and gender bias, and ending with the presenting of the research questions. The chapter of background and related work is followed by a chapter on the datasets, where the two datasets used in this thesis and their creation are explained in more detail. Next, the methodology chapter provides a detailed overview of the methods used to answer the research questions. The chapter on experimental results presents an analysis of gender distributions in the BDI-Sen dataset, shows the performances of the different classifiers, and findings regarding gender bias and mitigation. These results are further discussed in the discussion and conclusion chapter, answering the research questions and discussing ethical considerations, limitations and future work, as well as providing a short overview of the research's main findings.

2. BACKGROUND AND RELATED WORK

The following chapter presents the background and related work for this project. First, the background on depression provides necessary information on the mental illness and its symptoms, followed by a discussion on gender differences in depression symptoms as researched by psychologists, and findings of depression on social media. Next, text-based depression detection is considered, providing background on natural language processing (NLP) techniques, as well as discussing datasets and models for text-based depression detection on social media. Third, a background on gender bias is provided, including explanations for different bias measures and bias mitigation techniques, and presenting prior research on gender bias in depression detection systems on social media. Finally, the research questions this project aims to answer are introduced.

2.1. Background on Depression

Before exploring depression detection systems, it is important to gain an understanding of what depression is and how it is defined. The following section provides necessary background knowledge on depression and its prevalence in the population, as well as how it is diagnosed, helping to gain a better understanding of why detecting symptoms of depression is valuable. Moreover, research on gender differences in depression symptoms is discussed, providing the background to contextualize findings of gender differences in depression detection and detection of symptoms. Finally, findings regarding the relationship between depression and social media are presented, further emphasizing why depression detection on social media is important.

2.1.1. Overview on Depression

Depression, also called major depressive disorder or clinical depression, is a common mental health disorder. Worldwide, around 3.8 percent of the population and 5 percent of adults suffer from depression [14]. Someone who is depressed experiences persistent feelings of sadness and hopelessness, and has no interest in activities they once enjoyed [15]. The DSM-V describes depression as the experience of several symptoms over a two-week period [16]. These symptoms are summarized by the SIGECAPS mnemonic: Sleep disorders, interest deficit, guilt, energy deficit, concentration deficit, appetite disorder, psychomotor retardation or agitation, and suicidality [17]. If four SIGECAPS symptoms and depressed mood or interest deficit is present in a person, it is likely they have depression [17].

For some, depression is a cause of death, as a common symptom of depression is suicidal ideation. Suicide is an important problem in society, as over 700 000 people die this way every year, and it is the fourth leading cause of death for 15 to 29-year-olds [14]. This makes depression an important cause to work on. Despite effective treatments being available, people in low- and middle-income countries often do not receive the help they need [18]. Even in high-income countries, the stigma of mental health disorders might keep people from seeking help.

2.1.2. Depression Diagnosis

Depression is diagnosed by clinicians after multiple sessions with a client, and usually according to the DSM-V criteria [17]. To screen for depression, different questionnaires are commonly used. The U.S. Preventive Services Task Force (USPSTF) recommends that every adult, regardless of risk factors, should regularly be tested for depression [19]. Some of the most prominent questionnaires are the two-item and nine-item Patient Health Questionnaires (PHQs) [17]. The PHQ-2 can be used as an initial screening tool, which should be followed up by the PHQ-9 or a clinical interview if results are positive [17]. The PHQ-9 was designed after the criteria in the DSM-IV, asking participants to fill in a nine-item questionnaire about the presence of symptoms in the last two weeks on a four-point scale [20]. Another depression metric is the Beck Depression Inventory II (BDI-II) [21]. It consists of 21 items and is based on the DSM-IV as well. Participants rate their symptoms over the past two weeks on a three-point scale. An overall score between 10 and 19 indicates mild depression, a score between

Category	**Symptoms**
Affective	Sadness, loss of pleasure, crying, indecisiveness
Behavioral	Agitation, loss of energy, irritability, tiredness or fatigue
Cognitive	Social withdrawal, concentration difficulties
Cognitive distortions	Sense of failure, guilty feelings, sense of punishment, self-dislike, self-incrimination, feelings of worthlessness
Motivational	Pessimism, suicidal ideas
Physiological or vegetative	Change of sleep, changes in appetite, loss of interest in sex

Table 2.1: Beck's symptom categorization of the 21 BDI-II items.

20 and 30 indicates moderate to severe depression, and a score of 31 or higher indicates severe depression [22]. An overview of the symptoms grouped into categories defined by Beck [23] can be found in Table 2.1. For older adults, other screening tools are the Geriatric Depression Scale and the Cornell Scale for Depression in Dementia [17]. If depressive symptoms are present, the diagnosis should be confirmed by the DSM-V criteria [17].

Unlike physical diseases, for which symptoms help to diagnose the underlying cause, mental disorders such as depression do not have a singular common pathogenic pathway [16]. Instead, the occurrence of symptoms is itself the mental disorder that is diagnosed. For depression, this means that a clinician will identify whether a sufficient number of symptoms, as defined for example in the DSM-V manual, is present. A possible explanation for the importance of symptoms in mental disorders is provided by the network approach to psychopathology [24]. This theory suggests that the spreading of activation in a symptom network causes a mental disorder. For example, the passing of a loved one can activate the symptom of depressed mood, which might further cause fatigue and social withdrawal. If the activation of symptoms is maintained over a prolonged period of time even after the initial external activating factor is gone, it can be constituted as a mental disorder [24]. Apart from diagnosis, network theory also suggests that the treatment of a mental disorder needs to be on a symptomatic level [24]. It argues that by reducing the activation of one symptom, other symptoms might also become less active. Therefore, treatments can be classified into changing the state of one or more symptoms (e.g., through prescription of medication), removing one or more causes that trigger the symptom or symptoms (e.g., planning more activities with others), or modifying symptom connections (e.g., through cognitive behavioral therapy aimed to change how to deal with a depressed mood) [24]. Rather than detecting depression itself, it therefore seems important to gain a good understanding of the symptoms causing the depression, to be able to help a person successfully and effectively.

2.1.3. Gender Differences in Depression Symptoms

Some differences in depression have been found between men and women. Depression appears to be more common in women than in men, with about 50 percent more cases of depression in women [14]. This is consistent with research that shows that internalizing disorders are more prevalent in women, while externalizing disorders are more prevalent in men [25]. Research across three German cohorts confirmed this finding, showing that women report depressed mood and symptoms of depression more often than men [26]. Another factor contributing to differences in depression between men and women are gender inequalities, such as differences in income and education [26]. Researchers showed that in countries with higher gender inequality, people have more frequent and severe symptoms of depression, with older women being most effected [27]. Individuals living in more gender equal countries, as well as young adults, show less gender differences in depression [27].

Contrary to the higher depression rates for women, researchers found that the risk of suicide is higher for men [26]. While women are more likely to try suicide, men are more likely to die from it, which is also known as the gender paradox in suicide [28]. In 2019, the World Health Organization (WHO) reports global suicide rates of 12.6 per 100 000 for men and 5.4 per 100 000 for women, showing that men are on average about twice as likely to die from suicide than women.[1] One reason for the gender differences in attempts and completion is that men have been found to use more lethal methods [28]. The higher suicide rate might also indicate that men do not have a lower depression rate than women, but that depression in men remains undetected more often than in women. Researchers suggest that expectations around masculine norms might prevent men from seeking out help [26]. Moreover, it has been shown that men show different symptoms of depression than women, many of them being more atypical. Men appear to be more likely to show external risk factors, such as irritability, anger attacks, alcohol abuse, smoking, physical inactivity and risk-taking [26] [29]. When accounting for the atypical symptoms more prevalent in males, research has shown that gender differences in depression prevalence disappears, suggesting that current screening methods might fail to adequately measure depression in men [29]. This might make it more difficult to detect depression in males, not only for experts, but also for depression detection systems. The higher suicide rate emphasizes the need for accurate and timely depression detection for men.

It is also important to consider depression in non-binary individuals. Many gender identity groups fall under the umbrella term non-binary, including gender identities between male and female identities, experiences of being male or female at separate times, or the rejection of a gender identity [30]. While research on depression in non-binaries is limited, it has been found that they suffer from mental distress more often than the general population or even binary transgender individuals, with a large survey in the US showing that 49 percent reported serious psychological distress in 2015 [31]. This is confirmed by other studies, which show that non-binary individuals have higher rates of suicide attempts than binary transgender individuals, as well as high rates of depression and anxiety [32] [33]. However, while these studies point to the importance of studying the prevalence and symptoms of depression in non-binaries, it is difficult to investigate this issue more closely, as they only make up a small proportion of the population [30].

[1] https://www.who.int/publications/i/item/9789240026643

2.1.4. Depression and Social Media

The relation between depression and social media might provide further insight into the task of depression detection on social media. Increased sharing and discussion of mental health topics on social media [3] indicates that identifying depressed individuals on social media is possible, but also raises the question if there is a connection between social media usage and experiencing symptoms of depression. A meta study analyzing 62 articles on depression in social media found that rather than time spent on social media or intensity, it is problematic social media usage that influences the connection between social media usage and depression [34]. The researchers define problematic social media use as showing symptoms of addiction and a negative impact on functioning. Moreover, they found that age, gender, year of study publication, or design characteristics did not significantly moderate the relation between depression and problematic social media use. This suggests that the problematic use of social media is equally worrying for men and women, stressing a need to detect depression in both equally well.

There are many social media platforms, with the largest ones being Facebook, YouTube, WhatsApp, Instagram, and TikTok.[2] The BDI-Sen dataset consists of Reddit data, which is another large social media platform with almost 500 million users as of 2023 [35], and popular for research [36]. Reddit is structured into topical communities, called “subreddits”, in which users can discuss and share information related to a specific topic. There are over 100 000 subreddits on Reddit, with the most popular being “r/funny”, “r/AskReddit”, “r/gaming”, and “r/aww” [35]. One can also find subreddits related to mental health, such as “r/Depression”, and “r/Anxiety”. It appears that especially these subreddits are interesting for research, as many studies on Reddit use keywords such as “mental health” and “depression”, and the subreddits “r/depression” and “r/SuicideWatch” belong to the 20 most studied subreddits, showing the social media platform’s relevance for depression research [36]. However, with two thirds of the Reddit population being male and one third female, male users are

[2] https://en.wikipedia.org/wiki/List_of_social_platforms_with_at_least_100_million_active_users

in the majority, suggesting there might be bias in Reddit datasets [35]. This makes it important to ensure that data collected from Reddit and models trained on Reddit data are fair.

2.2. Text-based Depression Detection on Social Media

The following section discusses relevant literature for text-based depression detection on social media. First, a background on natural language processing (NLP) is provided, explaining traditional language modelling, deep neural networks and word embeddings. Next, models for text-based depression detection on social media are further discussed, pointing out the usage of BERT and mentalBERT, and the need for more interpretable models. Lastly, an overview of datasets for text-based depression detection on social media is given, explaining relevant datasets in more detail.

2.2.1. Background on Natural Language Processing

Natural language processing (NLP) is a research field tasked with developing tools involving human language, such as machine translation, natural language processing and summarizing, user interfaces, multilingual and cross language information retrieval, speech recognition and expert systems [37]. Text-based depression detection is one such task, as it aims to estimate whether an individual is depressed based on text. Before discussing the different methods that have been used to detect whether someone has depression, this section aims to provide the necessary background knowledge on NLP methods. Linguistic depression detection models can be divided into two types: Traditional language modelling and deep neural networks.

2.2.1.1. Traditional Language Modelling. The first type of models is focused on word or topic-level interpretability, using basic linguistic pattern features like bag-of-words (BOW) and linguistic inquiry and word count (LIWC) [38], as well as topic modeling, emotion features, and statistics such as n-grams and Term-Frequency-Inverse-Document-Frequency (TF-IDF) [39]. Also called statistical language modelling, this type of language modelling provides word counts and probabilistic distributions of words [40]. Methods such as n-gram aim to predict the next word following a number of words given a vocabulary and a maximum likelihood estimation [40]. Often these methods are combined with more traditional, mostly supervised machine learning algorithms. They are usually more interpretable than deep neural networks, but less likely to deduct patterns and relationships across larger distances [39].

2.2.1.2. Deep Neural Networks. The second type of models has increased in popularity in recent years, and focuses on deep neural networks, enabling better temporal modeling but unable to provide explanations [6].These type of models include convolutional neural networks (CNNs), long short-term memory networks (LSTMs) and transformers. Models such as transformers are often trained on word embeddings, which provide representations of language in vectors. They are dense, distributed, fixed-length word vectors, which are based on word co-occurrence statistics [40]. Word embeddings can be divided into two categories: Prediction-based models and count-based models [40]. Prediction-based models use contextual information of a word, enabling them to encode syntactic and semantic relationships. Count-based models consider word-context occurrence globally across a corpus. Popular word embeddings include Word2Vec [41] and GloVe [42]. Most recent studies on depression detection use embedding techniques and post-feature transformation [43].

A very influential and often used word embedding is BERT [44]. BERT stands for Bidirectional Encoder Representations from Transformers. Its base uncased model consists of two layers, a hidden size of 768, twelve attention heads and 110 million parameters. The self-attention layers allow the model to consider different positions of a sequence when computing the representation of that sequence [45]. BERT is pre-trained using two unsupervised tasks, namely a masked language model and next sentence prediction. The masked language model consists of masked words that the model is required to predict, while the next sentence prediction requires the model to predict a sentence following another sentence. BERT can be fine-tuned on many downstream tasks, by supplying the necessary input and output. This makes it a very versatile word embedding model.

2.2.2. Models for Text-based Depression Detection on Social Media

With more and more people sharing their personal lives on social media, it has become interesting for AI researchers to use this data to develop new models for different tasks. Depression detection is one of these tasks, with a lot of research conducted using textual data [2]. In fact, most mental illness detection has been researched on social media data, and out of different mental health disorders, depression is the one that has received the most attention [39]. Many recent models for depression detection apply deep neural networks, with the most commonly used deep learning techniques being CNNs, LSTMs and transformers such as BERT [2]. Moreover, bidirectional LSTM networks and feed-forward deep neural networks are often used for depression detection in social media. As BERT can be fine-tuned to different tasks, in the context of depression detection, a popular model is mentalBERT [46]. MentalBERT is based on the base uncased BERT model and uses the same number of layers, attention heads and parameters. It is a representative bidirectional masked language model, pretrained on a corpus of 13 671 785 sentences collected on Reddit, specifically subreddits addressing mental health issues such as “r/depression” and “r/Anxiety”. The pretraining follows a domain-adaptive manner, starting the training from the original BERT. MentalBERT is publicly available.[3]

Multiple researchers point out that more explainable models are needed for depression detection [2] [39] [43]. Apart from more interpretable datasets such as BDI-Sen [7] and PsySym [6], some research has successfully implemented explainable models. One study applied LIME to their classifier, a combination of SVM and k-nearest-neighbor, to identify important keywords [47]. This also helped in reducing bias, by reducing the impact some keywords can have on the final outcome. Another study pointed out the importance of explainability for early depression detection, alongside the incremental classification of sequential data and support for early classification [48].

[3] https://huggingface.co/mental

Other models created for more explainable depression detection include a pattern-based classifier with a feature representation extracted from an emotion and sentiment analysis [49], a symptom-based interpretable model using prototype learning [50], and a hybrid model [51].

While much research suggests that detecting depression using social media posts can be very successful, Chancellor and De Choudhury [52] identified some issues with research on mental health status assessment. The researchers point out that studies usually do not discuss how to establish construct validity, meaning that they do not explain how to ensure the validity of the ground truth labels. Similarly, constructs such as clinical questionnaires are not validated either. Therefore, it is unclear whether the classifiers trained on these ground truth labels indeed measure the mental health status aimed for, or something else. Moreover, the researchers criticize that very few papers discuss the clinical and psychological background of the construct they aim to measure, and definitions of them often remain implicit, which further complicates the problem [52]. A possible way to improve this issue could be with a dataset that has annotations for clinically validated symptoms, or a model that provides explanations, so that experts can understand why it makes certain predictions and whether they are in line with the clinical definition of the mental disorder it is aiming to detect.

2.2.3. Datasets for Text-based Depression Detection on Social Media

Different social media datasets have been created for depression detection. While detecting a mental health status is usually conceptualized as a binary classification problem between healthy and mentally ill individuals [52], the datasets differ in terms of features, social media, and size, even though most datasets were built using Twitter or Reddit data [2]. Moreover, the methods of establishing construct validity differ between the datasets, and consequently there is no common approach to ensure that the ground truth labels are accurate [52].

Some of the most extensively used datasets for depression detection in social me-

Dataset	Modality	Classification	Size	Source	Demogr.	Symptoms
CLPsych2015 [53]	T	Depression vs. control	1746 users + tweets	Twitter	G	-
160Users [56]	T	Depression vs. control on user	160 users	Twitter	-	-
SMHD [55]	T	9 mental illnesses and control	338 358 users	Reddit	-	-
eRisk2019 [54]	T	Depression severity	20 users	Reddit	-	per user
D-Vlog [57]	A, V, T	Depression vs. control	961 videos	YouTube	G	-
PsySym [6]	T	7 mental illnesses and control	8 554 sentences	Reddit	-	per sentence
BDI-Sen [7]	T	Depression symptoms	1 845 sentences	Reddit	-	per sentence
DepreSym [8]	T	Depression symptoms	21 580 sentences	Reddit	-	per sentence

Table 2.2: Depression detection datasets considered for this project (T = textual, A = audio, V = video, G = gender).

dia are the Reddit Self-Reported Depression Diagnosis (RSDD) dataset, RSDD-Time, CLPsych2015, SAD, eRisk Lab 2017 to 2020, and the Self-reported Mental Health Diagnosis (SMHD) dataset for 9 mental health conditions [2]. Out of these datasets, three are more closely discussed due to their relevance for this research: CLPsych2015 [53], eRisk2019 [54] and SMHD [55]. CLPsych2015 has been analyzed for gender and racial fairness [12], while eRisk2019 contains symptom information per user and has been used to create the BDI-Sen dataset. Other datasets considered for this project are 160Users [56] as another example for a Twitter dataset, D-Vlog [57], a YouTube dataset that has been analyzed for bias as well [13], and three symptom annotated datasets: BDI-Sen [7], PsySym [6], and DepreSym [8]. Table 2.2 provides an overview of these datasets. The BDI-Sen dataset was selected for this research due to its symptom annotations and its source of depressed sentences being the eRisk2019 dataset, which allows gender annotation. In the following, each dataset will be described in more detail.

2.2.3.1. CLPsych2015. The CLPsych2015 Shared Task dataset consists of 1 746 Twitter users and their tweets [53]. Three tasks were part of this dataset: A binary classification of post-traumatic stress disorder (PTSD) vs. control, depression vs. control, and PTSD vs. depression. The age and gender of a user were predicted through the demographic classification tool from the World Well-Being Project [58], in order to match each user with depression or PTSD with a control user. The dataset was not annotated for the symptoms of users with depression or PTSD, and has been analyzed for gender and racial bias [12]. The researchers found a dataset imbalance, with an overrepresentation of whites and females, which led to lower performance for the minority groups, especially for women of color. Improving the representation helped to mitigate most of the bias, however error rates were still higher for people of color. These research findings are further discussed in Section 2.3.3.

2.2.3.2. 160Users. The 160Users dataset was created as part of a study on depression detection on Twitter data [56]. It is an extended dataset of the study's 60Users dataset, containing 60 users annotated for depression at tweet-level. The remaining users were annotated with an undersampled version of the dataset for fast annotation, resulting in 80 users classified as depressed and 80 control users. The researchers first trained linear SVM models to predict depression at tweet-level using the 60Users dataset, then built on these models to detect depression on a user of the 160Users dataset based on their tweets. Their findings show that informative features for the model were polarity word count, depression word counts and pronoun word counts, but also the percentage of depressed tweets and whether a Twitter user reports themselves as depressed. These features can have a large effect on model performance, as they simplify the classification task of detecting depression. The dataset does not contain any demographic information or symptom annotation, and as of this research, it appears that no studies on fairness have been conducted on it.

2.2.3.3. SMHD. The Self-reported Mental Health Diagnosis (SMHD) dataset contains Reddit posts of users suffering from a mental illness and control users [55]. It consists of 20 406 self-reported diagnosed users and 335 952 control users. Mental health related posts were removed from the dataset, to ensure that the classification is done on posts that do not contain mental health terms. Control users were chosen based on similarity to diagnosed users in terms of subreddits they posted on and amount of posts. SMHD does not provide information on user demographics or prevalent symptoms. Moreover, it appears that no research has been conducted on fairness for the dataset.

2.2.3.4. eRisk2019. The aim of eRisk datasets is the evaluation of early risk detection systems [54]. The eRisk2019 Shared Task dataset consists of three tasks. The first two tasks address early detection of signs of anorexia and self-harm, while the third is to predict a social media user's depression level based on their posts. This depression severity shared task dataset contains the Reddit posts and comments of 20 users and their symptom annotations in the form of the BDI-II questionnaire the users filled in. The symptom annotations are taken from this questionnaire, resulting in 21 symptoms, such as sadness, pessimism, and loss of energy. The data is based on a dataset collection from 2016 [59] and does not include annotated demographic information. A related work search resulted in no studies evaluating fairness on this dataset.

2.2.3.5. D-Vlog. D-Vlog is a dataset consisting of vlogs of both depressed and non-depressed individuals collected on YouTube [57]. It has 961 videos with 816 different individuals, 555 of which show depressed individuals, and 406 non-depressed. Using videos allows for multi-modal fusion to detect depression, combining audio and visual features. Moreover, it is possible to add a textual modality. Four college students annotated the data for depression. The researchers created a multi-modal transformer model to identify depressed individuals, taking acoustic and visual inputs and showing good results. Given the additional modalities, demographic information such as gender is easier to extract. While symptoms are not annotated, research on gender fairness for the dataset exists [13]. The study points out an uneven distribution in the dataset, with roughly double the number of females compared to males. However, even though the number of females is higher, the researchers identified a bias in favor of males.

Using different mitigation techniques, they found that none of them could improve the bias, with many fairness measures showing poor performance and contradicting each other. Their findings are further discussed in Section 2.3.3.

2.2.3.6. PsySym. PsySym is an annotated symptom identification corpus of multiple psychiatric disorders, providing a knowledge graph of 38 symptom classes and seven mental diseases [6]. The symptoms were found using the DSM-V, and help to explain predictions made by machine learning models. By using a knowledge graph, it is possible to clearly see which symptoms are shared by multiple diseases. The mental illnesses included are: depression, anxiety, attention deficit hyperactivity disorder (ADHD), bipolar disorder, obsessive-compulsive disorder (OCD), PTSD, and eating disorder. These symptoms were manually annotated for 8 554 Reddit post sentences from mental health related subreddits. The annotation was split into relevance judgement and status inference. The relevance judgement describes how relevant the sentence is for the given symptoms. Status inference describes whether the symptom(s) deemed as relevant are present in the sentence. A model trained on the status inference provides a predicted probability that the symptoms are present, becoming the status feature. Moreover, the researchers include a subject feature to account for whether the symptoms mentioned in the post are experienced by the poster themselves, looking at the number of third-person pronouns against first-person pronouns. These three features are multiplied to reach a final score. The dataset does not provide demographic information and has not been analyzed for gender bias.

2.2.3.7. BDI-Sen. BDI-Sen is a dataset with sentences taken from Reddit posts of the eRisk2019 depression severity collection [54] and corresponding annotations of 21 depression symptoms based on the BDI-II questionnaire [7]. The dataset consists of 550 sentences with depressive symptoms and 1 295 sentences without depressive symptoms. Annotations of the text were made by a psychologist, a speech therapist and a PhD student with knowledge in the field, and symptoms were separated into four severity levels, corresponding to the BDI-II. Sentences were annotated for relevance and severity label. A dataset analysis showed that there are linguistic and emotional differences between the symptoms. Moreover, the researchers trained models for both symptom detection and symptom severity classification. The dataset does not provide demographic information and no fairness analysis has been conducted for it.

2.2.3.8. DepreSym. Similar to BDI-Sen, DepreSym consists of sentences labeled for the BDI-II questionnaire symptoms [8]. The sentences were taken from the eRisk 2023 shared-data ranking task [60], which asked contributors to identify whether a sentence contains symptoms of depression. Three experts annotated the sentences for each of the 21 symptoms. The results were compared to the annotations of two large language models (LLMs), ChatGPT and GPT-4, showing that while LLMs still make errors, they can be used to filter out non-relevant sentences. The dataset consists of 21 580 sentences and does not include demographics. No fairness research has been conducted on it.

2.3. Gender Bias

Investigating gender bias is part of the research field of responsible AI. Different bias metrics have been developed to measure bias, and many mitigation strategies aim to reduce harmful biases, as they can often be amplified by machines.

The harm of bias can be separated into two categories: Representational harm and allocational harm [11]. Representational harm describes the harm of reinforcing certain stereotypes related to identity, such as race, class and gender. Allocational harm describes the harm of not providing the same amount of resources or opportunities to a group. Gender bias related to depression detection in social media causes allocational harm rather than representational harm, as it can lead to one group of depressed individuals receiving on average less help for their mental health than the other group (e.g., if men are less likely to be detected as depressed, they are less likely to receive resources to improve their condition).

Another difference can be made between statistical bias and social bias. Statistical bias occurs when a sample is non-representative [61]. For example, if a dataset for depression detection is imbalanced and consists of more data for depressed women than for depressed men, there can be a statistical bias. Statistical bias also occurs when there is a measurement error. For example, it might be that the same proportion of men and women suffer from depression, but if a system detects more depressed women than men, one might conclude that more women have depression than men, which would be a statistical bias. This can also lead to a feedback loop, as researchers might consequently develop more strategies to identify depressed women, believing that they are more at risk and should be focused on, which would further increase the proportion of depressed women being detected. Social bias describes a bias in society, regardless of statistical data [61]. One type of social bias are stereotypes.

2.3.1. Background on Bias Measures

Bias in AI models can be measured in different ways. Some of the most commonly used measures are statistical parity, equalized odds, predictive equality, equal opportunity, and predictive parity. For each of these measures, both rates of a sensitive attribute are calculated (such as the true positive rate for males and females), and one is either subtracted from the other, or divided by the other. A recent study showed that taking the ratio measure is usually a better indicator for bias, as it is more sensitive to differences and captures more biases [62]. Therefore, ratio measurements are used in this research. A score of 1 is considered perfectly fair, as it indicates equal scores for both sensitive attributes. To follow the disparate impact measure, studies usually regard scores outside a range of 0.8 to 1.2 as biased [13] [63]. For depression detection, studies often compare F1 scores and measure for equalized odds, which can be divided into predictive equality and equal opportunity [12] [13]. In the following, these bias measures are explained in more detail.

2.3.1.1. Equal Opportunity. Equal opportunity considers the true positive rates of two groups, considering the predicted and the actual outcome when measuring bias [64]. For a binary classifier, it represents the notion that everyone who falls under the positive class should receive the same treatment, regardless of which type of sensitive attribute they are a member of. For example, this would mean that women who are depressed should be just as likely to be predicted as depressed as men who are depressed. Equal opportunity can be satisfied if all groups of a sensitive attribute receive equal treatment when their actual outcome is positive, resulting in the following equation, where $\hat{Y}$ represents the predicted outcome or decision, Y the actual outcome, and A the sensitive attribute:

$$Equ.opp. = \frac{\mathbb{P}\{\hat{Y} = 1 | Y = 1, A = a\}}{\mathbb{P}\{\hat{Y} = 1 | Y = 1, A = b\}} \tag{2.1}$$

2.3.1.2. Predictive Equality. Predictive equality requires individuals with a predicted positive outcome but actual negative outcome to be treated the same regardless of sensitive attribute, and uses false positive rates [64]. In the case of depression detection, this means that women who are not depressed should be just as likely to be predicted as depressed as men who are not depressed. Therefore, it requires this equation to be satisfied:

$$Pred.equ. = \frac{\mathbb{P}\{\hat{Y} = 1 | Y = 0, A = a\}}{\mathbb{P}\{\hat{Y} = 1 | Y = 0, A = b\}} \tag{2.2}$$

2.3.1.3. Equalized Odds. Equalized odds requires the satisfaction of both equal opportunity and predictive equality [64]. Often, this means that the worse-performing measure is reported. However, looking at equal opportunity and predictive equality separately helps to understand where exactly the bias is and allows for a more thorough fairness analysis.

2.3.2. Background on Bias Mitigation Techniques

Bias mitigation techniques aim to reduce or eliminate the bias that can be found in the data or the classifier. They can be categorized into pre-processing, in-processing, and post-processing mitigation techniques.

2.3.2.1. Pre-processing. As the name suggests, pre-processing strategies aim to alter the data before training the classifier, meaning there are no changes made to the classifier itself [64]. Examples for pre-processing are data augmentation [65] and reweighing [66]. Data augmentation aims to increase training data by creating synthetic data. For example, in [65], the researchers created additional training data by replacing male entities with female entities and vice versa (e.g. "he" being replaced with "she"). Reweighing is a mitigation method that assigns weights to training instances [66]. Based on the assumption that the predicted outcome is independent of a sensitive attribute, one calculates an estimated probability and divides that by the observed probability, assigning each instance with the same sensitive attribute and outcome this weight.

2.3.2.2. In-processing. In-processing strategies focus on both the classifier and the data, manipulating them to make the outcome more fair. An example for an in-processing technique is adversarial debiasing [67]. Adversarial debiasing is the idea of using a model to try to predict the sensitive attribute based on the outcome of the classifier [67]. The classifier is then trained to decrease the adversarial model's accuracy while increasing its own performance.

2.3.2.3. Post-processing. Post-processing techniques change the outcome after training the classifier, meaning there is direct control on the outcome distribution. An example for this technique is equalized odds, which is related to the same bias measure and aims to adjust precision and recall values so that they are the same regardless of the value of the sensitive attribute [68]. This is done by using a predictor that depends only on the predicted outcome and the sensitive attribute. While training it, it aims to find the parameter for the probabilities by minimizing the loss of the predicted outcome and true outcome.

2.3.3. Gender Bias in Depression Detection on Social Media

Not much research has been conducted to investigate gender bias in depression detection systems on social media. From the datasets discussed previously, only the CLPsych2015 dataset and the D-Vlog dataset have been analyzed for gender bias, as shown in Table 2.3. All research has focused on binary gender bias, disregarding non-binary individuals. One study found that a gender and racial bias exist in the CLPsych2015 Shared Task dataset [53] and the MULTITASK dataset [69], and that this is caused by a dataset imbalance [12]. They showed that both gender and race are not equally represented in the datasets, with an overrepresentation of whites and females. To mitigate this bias, they created a balanced dataset using data augmentation, which improved performance for the underrepresented groups. However, the researchers found that even with a balanced dataset, performance is lower for people of color. As LIWC correlations with depression are different for different demographic groups, they suggest that this technique might introduce bias.

Other studies' findings suggest a potential gender bias by investigating the connection between gender and depression rate. Another study on the CLPsych2015 dataset shows that age and gender are predictive for depression and PTSD, especially in combination with each other, suggesting they contain complementary information [70]. This is confirmed by another study, which finds differences in depression and PTSD rates among different demographic groups as well [71]. The researchers created a digital cohort to detect depression and PTSD. The cohort approach is a well-established technique in the medical field, and uses a pre-selected group to investigate disease causes and links between risk factors and health outcomes. Similar to [70], the researchers found a correlation between age and PTSD rates, as well as a higher likelihood for women, blacks, and Hispanics to be depressed than men and whites. They suggest that this could either be due to a bias in the data that the classifiers were trained on, or it shows that different demographic groups use social media differently. Their finding is further confirmed by other research, which used demographic information and information about sentiment to identify depression on social media [72]. By adding this information, the researchers aimed to account for the context in which a post is written, which could be important for depression detection. They found that information about a user's gender is more useful than information about a user's age, supporting the idea that there might be a difference in expression of depression between men and women.

Multi-modal depression detection is another research direction for social media data. A study has shown that the D-Vlog dataset is biased, achieving better performance for depression detection on males than females [13]. Different mitigation techniques were used to mitigate this bias, including data augmentation, reweighing, and reject option classification (ROC). ROC re-classifies predictions of the minority group if they are within a certain decision threshold region, a technique suggested by [73]. However, these methods did not successfully reduce the bias, with many fairness measures showing poor performance and contradicting each other. The researchers suggest this might be due to the pre-processing method of the data. As pre-processing entailed the truncation of vlogs to ensure the same video length for all data entries, which affected females more than males, more information of female vlogs might have been lost.

Dataset	**Fairness**	**Bias**	**Mitigation techniques**
CLPsych2015 [53]	G, R	Dataset imbalance [12]	pre [12]
160Users [56]	-	-	-
SMHD [55]	-	-	-
eRisk2019 [54]	-	-	-
D-Vlog [57]	G	Dataset imbalance [13]	pre, in, post [13]
PsySym [6]	-	-	-
BDI-Sen [7]	-	-	-
DepreSym [8]	-	-	-

Table 2.3: Fairness studies for the depression detection datasets considered for this project (G = gender, R = race).

2.3.3.1. Gender Inference on Social Media Datasets. In order to investigate gender bias or research gender differences on text-based social media posts, it is often necessary to infer the gender of a user. The gender of a user is sensitive data and often not provided, but different methods have been developed to estimate it. One of the first methods to infer the gender of a social media user was developed by Burger et al. [74]. Their method was used to analyze gender and racial bias in the CLPsych2015 Shared Task dataset [12]. The dataset used consists of 184 000 Twitter users, their gender labels derived by connecting gender information found on the users' blogs to their Twitter account. Another technique developed to identify gender is the Demographer [75]. This technique uses the name of a Twitter user to infer their gender, by applying a list of names with associated genders and a classifier that estimates a user's gender based on certain features in their name. The combination of both methods with no accuracy threshold reaches an accuracy of 90.42 percent. More recent research has found new ways to infer gender and could be helpful for research in gender bias using Twitter datasets [76] [77] [78].

Research has investigated gender inference on Reddit as well, with findings indicating privacy issues and gender differences in usage of the platform. One study pointed out a potential privacy issue on Reddit by building models to classify gender and citizenship [79]. They built a test set by deriving gender annotations on Reddit using lexical rules. The regular expressions for gender included words such as “my” in combination with “husband” or “girlfriend”. A weighted soft-margin SVM and supervised latent Dirichlet allocation (LDA) were used to classify users according to their gender and citizenship. Another study on Reddit data investigated gender differences in interests [80]. The researchers created a classifier to infer gender based on usernames. As Reddit users often have usernames that do not contain their first name, this method led to the majority of Reddit comments of the dataset being discarded, as their authors’ gender could not be inferred. Analyzing different subreddits for differences in male and female participation, the researchers concluded that there is an indication of gender differences in subreddit commenters between topics. Finally, a third study created a dataset called RedDust, which contains annotations for Reddit users regarding profession, hobby, family status, age, and gender [81]. The dataset consists of over 300 000 Reddit users. To obtain information regarding gender, the researchers manually created a set of patterns that indicate gender. This consisted of patterns such as “I am” followed for example by “a mother”. Moreover, bracket patterns were used (e.g. “I [30f]...”). Lastly, flairs gave an indication for gender. Flairs are subreddit-specific and provide users with more information. In subreddits such as “r/AskWomen” or “r/AskMen”, users often use flairs to indicate their gender.

Inferring gender is a sensitive issue. Not only are there privacy concerns with predicting the gender of a person when they explicitly do not provide that information, but it is also possible that this amplifies gender stereotypes [82]. Foschvillaronga et al. [82] criticize automatic gender inference, pointing out the dangers of misgendering, the reinforcement of gender stereotypes, and the simplifications used to turn gender inference into an algorithmic task. Moreover, these gender inferring methods usually only infer binary gender, disregarding non-binary gender identities. However, even when gender information is not provided, an algorithm might use proxy variables and still be influenced by gender, without the intention of the researchers. It can therefore be helpful to create explicit labels for gender and analyze whether such bias exists to avoid unintended biased outcomes.

2.4. Research Questions

Considering the lack of research regarding gender bias on a social media dataset annotated for symptoms of depression, this research aims to answer the following research questions:

RQ1: How does a classifier trained on the BDI-Sen dataset perform for different genders?

(i) Are the observed gender distributions in the BDI-Sen dataset per symptom clinically validated or specific to the dataset?
(ii) Measuring F1 score, predictive equality, and equal opportunity, what is the level of gender bias in the classifier?
(iii) Are there symptoms for which the classifier's performance is lower compared to other symptoms?
(iv) Are symptoms with lower performance more prevalent in men or women in the dataset?

RQ2: How can any present gender bias in terms of dataset imbalance, F1 score, predictive equality, or equal opportunity be mitigated?

(i) How do different data augmentation techniques impact gender bias as measured by F1 score, predictive equality, and equal opportunity?
(ii) How do these mitigation techniques impact the classifier's performance for different symptoms?

3. DATASETS

In the following chapter, the two datasets used for this project are presented. First, the properties of the BDI-Sen dataset are provided and the method to allocate sentences with depression symptoms to the users of the eRisk2019 dataset is explained. Next, the GABDI-Sen dataset is presented, explaining how the separate set of sentences without depression symptoms was created, and presenting an overview of the basic statistics of the dataset, as well as the distribution of symptoms over the GABDI-Sen test set, which was used to measure gender fairness.

3.1. BDI-Sen

As mentioned in Section 2.2.3, the BDI-Sen dataset [7] contains sentences from Reddit posts taken from the eRisk2019 depression severity collection [54]. Each sentence is annotated for the 21 depression symptoms of the BDI-II questionnaire, indicating for each symptom whether it is present in the sentence or not. The number of sentences with depression symptoms constitutes approximately 30 percent of the dataset, with 550 sentences containing at least one depressive symptom and 1 295 sentences without depressive symptoms. Sentences with depression symptoms will also be referred to as positive instances, while sentences without depression symptoms will be referred to as negative instances. A description of the BDI-II symptoms and an overview of the categories they are grouped into can be found in Section 2.1.2.

Apart from symptom annotation, the BDI-Sen dataset also has severity annotations. These severity annotations refer to the severity of the symptom(s) present in the sentence and not overall depression, and were taken from the answers users from the eRisk2019 dataset gave to the BDI-II questionnaire. They range from 0 to 4, with the lowest severity being indicated by 0, the highest by 3, and severity 4 indicating no severity. Looking at the severity distribution in Figure 3.1 provides closer insight into how severe the annotated symptoms for each sentence are. The severity level with the largest number of sentences apart from level 4, which matches the number of sentences with no symptoms, is level 1, with 237 sentences. This suggests that the users' experience of most symptoms is not very severe.

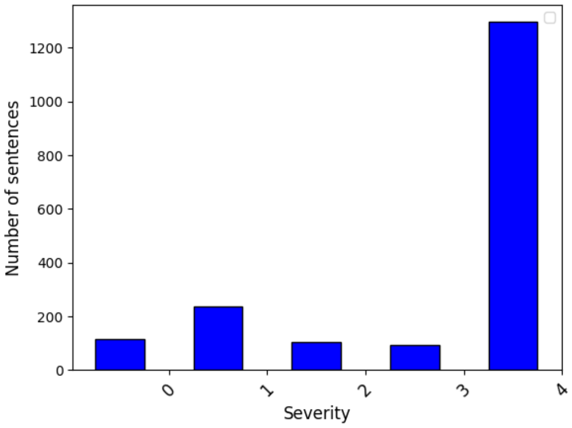


Figure 3.1: Distribution of sentences over severity levels. Severity is annotated based on the symptom(s) present in the sentence.

Figure 3.2 provides an overview of the distribution of positive sentences across the symptoms. It shows that sadness is the most common symptom in the dataset, with 154 sentences. Loss of pleasure is the second most common symptom with 141 sentences, with a steep decline before the other symptoms follow. This matches with the definition of depression according to the DSM-V, which defines that a depressed person will have at least depressed mood or diminished interest or pleasure among other symptoms that can vary for each individual [15]. Loss of interest in sex is the least common symptom in the dataset, with only three sentences. This could indicate that sex is not frequently talked about, even anonymously.

3.1.1. Sentences per User

The BDI-Sen dataset [7] is not annotated for gender or user ID. Therefore, using the eRisk2019 depression severity dataset [54], the positive sentences were allocated to the 20 users in the eRisk2019 dataset by searching for each sentence, and each sentence was annotated for the user accordingly. 549 of the 550 positive sentences were found. None of the sentences appeared in more than one user's posts. The collection

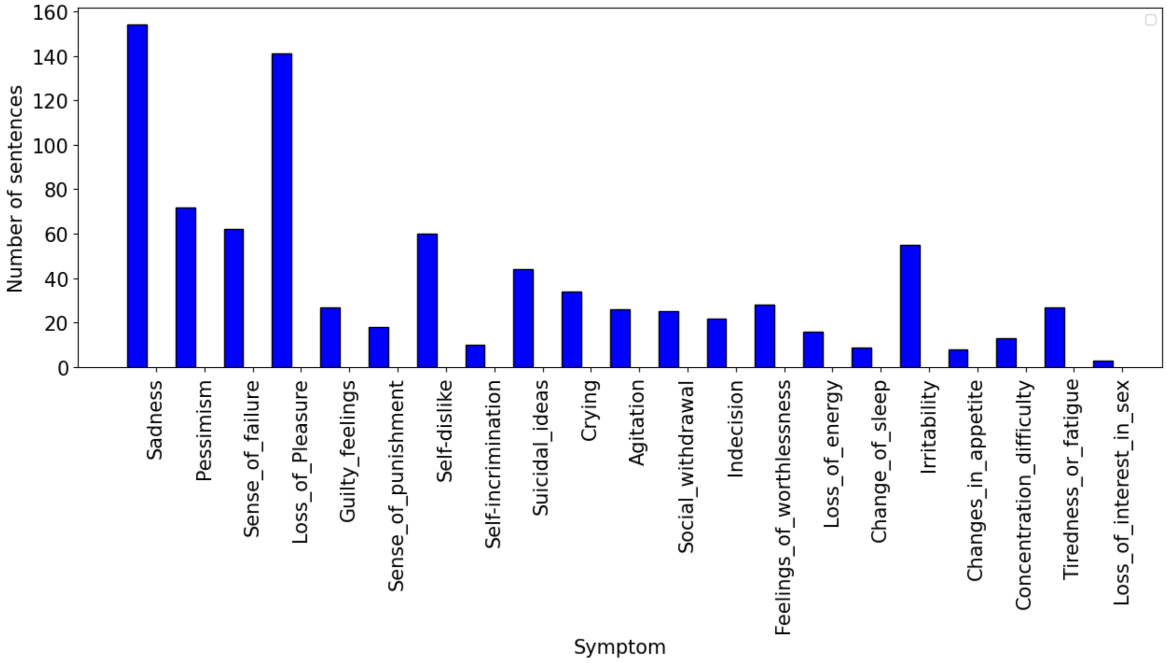


Figure 3.2: Distribution of sentences over symptoms.

of the sentences appears to have been very random, with large variation in the number of sentences appearing in the BDI-Sen dataset per user. A factor that might have influenced this is the variation in the number of posts and comments for each user. Figure 3.3 visualizes the number of sentences collected from each user.

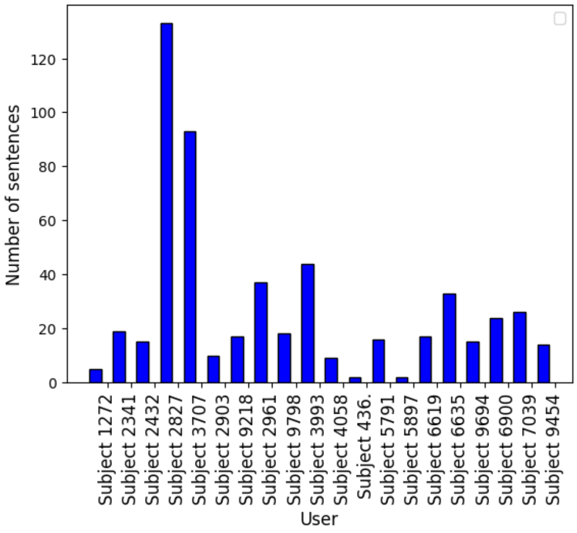


Figure 3.3: Distribution of sentences labelled with symptom(s) over users.

Due to the small size of the dataset, the gender of each user could be found manually. Searching the posts and comments in the eRisk2019 dataset for gendered

Female	“woman”, “female”, “girl”, “lady”, “wife”, “mother”, “sister”
Male	“man”, “male”, “boy”, “husband”, “father”, “brother”

Table 3.1: Terms used to retrieve users’ genders. Terms are used in combination with: “I am a”, “I’m a”, “I am” or “I’m”, and taken from [81].

terms such as “man”, “woman”, “girl”, and “guy”, each user’s gender was inferred. Each positive sentence in the BDI-Sen dataset was then annotated for gender according to its user annotation. The positive sentences of BDI-Sen are authored by 11 males and 9 females, with 212 sentences from male users, and 337 sentences from females. The negative sentences of BDI-Sen were randomly collected from Reddit, making gender inference impossible. Therefore, a new dataset for gender bias analysis was created, which will be referred to as GABDI-Sen (gender-annotated BDI-Sen).

3.2. GABDI-Sen

The negative sentences of BDI-Sen were randomly collected from Reddit. These sentences were used for replicating a model that was used by the BDI-Sen researchers. However, to measure gender bias, a different set of negative sentences with gender annotations was created. The Reddit API was used to collect 50 authors from the four most popular subreddits (“r/funny”, “r/AskReddit”, “r/gaming”, and “r/aww”). In order to find the authors’ genders, the methodology to collect gendered users for RedDust [81] was implemented, searching for language patterns such as “I am a mother” or “I’m a man”, as well as patterns of brackets revealing age and gender of a person, such as “35m”. The terms used to retrieve users’ genders can be found in Table 3.1. The posts and comments of users whose gender was inferred with the use of language patterns were manually checked. Example posts to identify a user’s gender are given in Table 3.2.

A total of 50 gendered authors was collected from the ten most popular posts of each of the four largest subreddits on either March 6 or March 7, 2024. In order to create a sample of sentences from these authors, 45 sentences of each author were

"I'm a male so there are more opportunities."	**Male**
"I was 16 and quite interested (for context, I'm a girl)."	**Female**

Table 3.2: Example posts to retrieve user with known gender. Posts were paraphrased and shortened.

randomly pulled from their posts and comments. After removing duplicates, keywords for depression symptoms described in the DSM-V and BDI-II such as "fatigue" and "loss of pleasure", as well as some synonyms for them collected from an online thesaurus were searched for in the remaining sentences (for a list of keywords used, see Appendix A). While this could not guarantee that every user with depression symptoms was excluded from the dataset, it was used to lower the chance of sentences with depression symptoms being part of the set of sentences without symptoms. Sentences containing these keywords were analyzed manually to determine whether depression symptoms are discussed, and users who describe depression symptoms were removed. This revealed two users talking about their depression symptoms, and consequently their sentences were removed from the dataset, resulting in 1803 sentences from 48 users. As the usernames of the users from the eRisk2019 dataset are not known, it could not be ensured that none of these users authored some of the positive sentences from the BDI-Sen dataset. However, given that the users were selected from vastly different subreddits (e.g., subreddits such as "r/depression" for depressed users and subreddits such as "r/AskReddit" for non-depressed users), and at very different times (2016 and 2024), chances that the 20 users of each collection overlap are low.

Out of the resulting sentences the same number of female and male users as in the set of positive sentences, as well as the same number of female and male sentences, were randomly selected, resulting in 9 female users with a total of 337 sentences, and 11 male users with a total of 212 sentences. Using this new set of negative sentences and the original set of positive sentences, a new dataset was created. This dataset is called GABDI-Sen (gender-annotated BDI-Sen). Since the dataset split in the original BDI-Sen dataset is not user-independent, a new split for the positive sentences was made. Data was split into training, validation, and test sets using a 7:1:2 split, which

Subset	**Total**	**Positive**	**Negative**	**Female sentences(users)**	**Male sentences(users)**
Train	824	438	386	489(12)	302(16)
Val	99	45	54	60(2)	38(2)
Test	208	99	109	124(4)	84(4)
Total	1098	549	549	651(18)	424(22)

Table 3.3: Basic statistics of the gendered dataset.

was the approach taken for the original BDI-Sen dataset. The data was split so that for each gender, both the number of users and the number of sentences matches this split. The basic statistics for the dataset can be found in Table 3.3. The training set, as well as the entire dataset, were used to train classifiers, and the test set was used for the gender fairness analysis, as it contains gender annotations for both sentences with depression symptoms and without depression symptoms. An overview of the distribution of the positive sentences over the symptoms for the GABDI-Sen test set can be found in Figure 3.4.

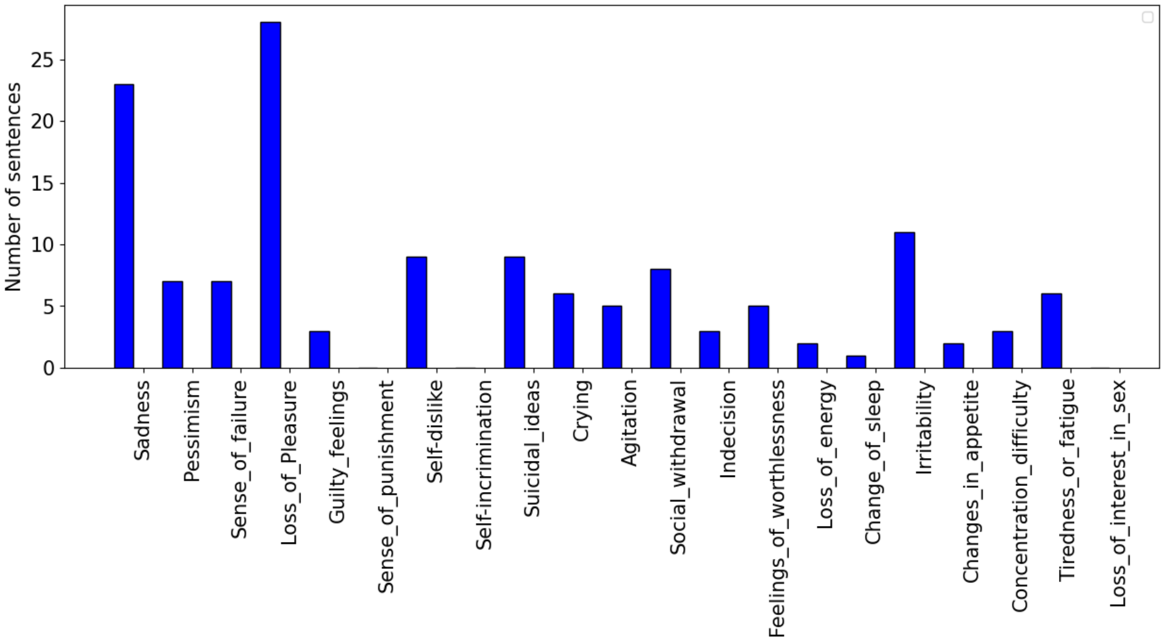


Figure 3.4: Distribution of sentences labelled with symptom(s) over symptoms for the GABDI-Sen test set.

4. METHODOLOGY

The following chapter describes the methods used to answer the research questions. Five steps had to be taken. First, the BDI-Sen and GABDI-Sen datasets were prepared and analyzed for gender differences. Next, three classifiers were set up, trained on either the BDI-Sen dataset or the GABDI-Sen dataset, and either predicting symptoms or symptom categories. Finally, gender bias was measured using ratios of F1 score, predictive equality and equal opportunity, and mitigated using different data augmentation techniques. An overview of these steps can be found in Figure 4.1.

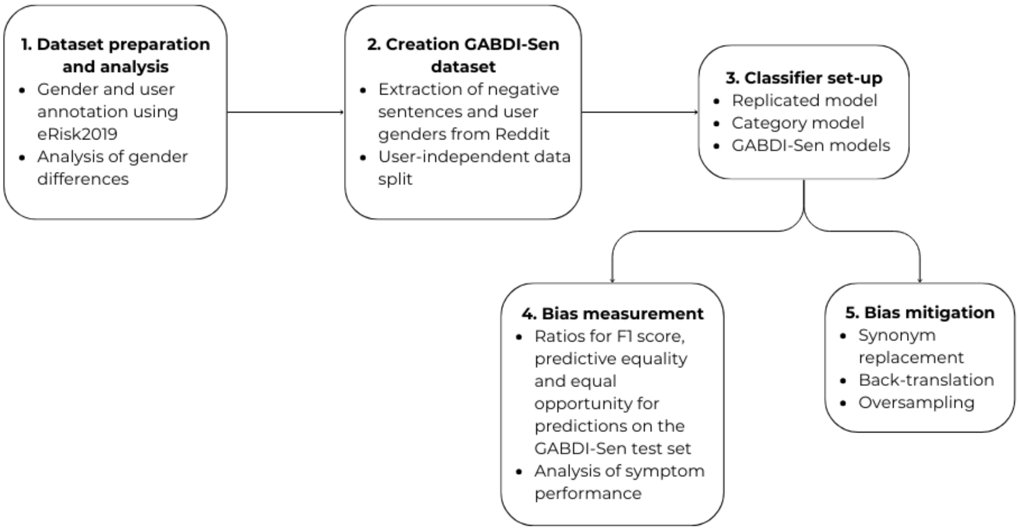

1. Dataset preparation and analysis
• Gender and user annotation using eRisk2019
• Analysis of gender differences
2. Creation GABDI-Sen dataset
• Extraction of negative sentences and user genders from Reddit
• User-independent data split
3. Classifier set-up
• Replicated model
• Category model
• GABDI-Sen models
4. Bias measurement
• Ratios for F1 score, predictive equality and equal opportunity for predictions on the GABDI-Sen test set
• Analysis of symptom performance
5. Bias mitigation
• Synonym replacement
• Back-translation
• Oversampling

Figure 4.1: Overview of the methodology.

4.1. Analysis of Gender Differences in the Data

After gender annotation, the positive sentences were analyzed for gender differences in symptom prevalence. Distributions of male and female sentences per symptom and male and female users per symptom were analyzed. Moreover, distributions of male and female sentences per symptom category were investigated. The significance of the differences in gender distributions per label was measured with chi-square tests. The significance analysis was done only on the sentence-level distributions as models are trained on sentences, making results on the sentence-level more important than results on the user-level. These steps aimed to answer the first sub-question of research question 1, demonstrating whether different gender distributions are present and whether they are clinically validated.

4.2. Classifier Set-up

Three classifiers were set up to compare for fairness. First, a classifier implemented by the BDI-Sen authors was replicated. Next, a classifier trained on the original BDI-Sen dataset and predicting symptom categories was used. Finally, classifiers were trained on the GABDI-Sen dataset. For all three types of classifiers, mentalBERT was used, as this was the best-performing model tested by the BDI-Sen dataset authors [7]. Moreover, it has been pre-trained on Reddit data, which fits well for the BDI-Sen data, and is available on Huggingface, making it easily accessible and therefore interesting to test for gender fairness. As transformers have become popular for depression detection in recent years, it is important to ensure fairness in them. The hyperparameters used to fine-tune mentalBERT were taken from the BDI-Sen study as well, specifically a learning rate of $2e^{-5}$, a maximum sequence length of 128, 20 epochs, and a batch size of 32. As the classification head was not reported, a dropout layer and a linear layer were used.

4.2.1. Replicated Model

The replicated model is the classifier that was implemented by the BDI-Sen authors. It was fine-tuned on the original BDI-Sen training set using the provided hyperparameters, and predicts the presence of each of the 21 depression symptoms of the BDI-II questionnaire in a sentence. Performance is reported on the BDI-Sen test set, which can confirm whether its performance is the same as reported in the BDI-Sen study. Moreover, performance is also reported on the GABDI-Sen test set, making fairness analysis possible. As the dataset split of the positive sentences in the GABDI-

Sen dataset is different from the BDI-Sen dataset split, some sentences the model is trained on are likely part of the GABDI-Sen test set. This could have had an impact on its performance for the GABDI-Sen test set.

4.2.2. Category Model

Apart from replicating the original BDI-Sen mentalBERT model, a model was trained to predict the presence of the categories each symptom falls into, according to Beck [21]. These categories include: Affective, behavioral, cognitive, cognitive distortions, motivational, and physiological. The input for this classifier was the BDI-Sen training set with category annotations instead of symptom annotations for each sentence, increasing the number of instances per label and making results more meaningful. Again, results are reported on both the BDI-Sen test set and the GABDI-Sen test set, and similar to the replicated model, the different dataset split might have had an impact on performance for the GABDI-Sen test set.

4.2.3. GABDI-Sen Models

Finally, models were trained on the GABDI-Sen dataset. Training was done at both the symptom and category level to compare results with the replicated model and the category model. First, a model was trained on the GABDI-Sen training set. Moreover, due to the small number of instances in the GABDI-Sen dataset, models were trained using 9-fold stratified cross-validation (CV) over the entire GADBI-Sen dataset to measure gender fairness. Nine folds were chosen to ensure that each validation fold contained sentences from one depressed female, as only nine depressed females are present in the dataset. Results for each validation split were collected and averaged. Hyperparameters were kept the same as in the replicated and the category model.

4.3. Bias Measurement

Once the models were trained, they were analyzed for gender bias. Similar to the gender bias analysis on the CLPsych2015 dataset [12], multiple bias measures were used, including F1 score ratio and equalized odds, which is split into the predictive equality and equal opportunity measures. All scores were measured for overall performance, as well as per label. Laplace smoothing was used for the true positive and false positive rates when measuring predictive equality and equal opportunity, ensuring that values of 0 were not encountered by adding 1 to both the numerator and the denominator. For each measure, ratios were reported, with the minority class in the numerator and the majority class in the denominator. As males were in the minority (see Section 3.1), male scores were divided by female scores:

$$F1Score = \frac{F1_m}{F1_f} \tag{4.1}$$

$$Pred.equ. = \frac{\mathbb{P}\{\hat{Y} = 1|Y = 0, A = m\}}{\mathbb{P}\{\hat{Y} = 1|Y = 0, A = f\}} \tag{4.2}$$

$$Equ.opp. = \frac{\mathbb{P}\{\hat{Y} = 1|Y = 1, A = m\}}{\mathbb{P}\{\hat{Y} = 1|Y = 1, A = f\}} \tag{4.3}$$

With these measures, differences in performance and fairness per label were identified and compared to the results of significance tests of the gender distributions for the dataset. Looking at both predictive equality and equal opportunity allows for a better understanding of the source of bias because it provides a more complete picture than one measure for equalized odds. Moreover, for the 9-fold CV model, t-tests were conducted to test for a significant difference in distributions of true positive rates and false positive rates across gender for each label. The gender bias analysis on the dataset answers research question 1 and its remaining sub-questions, analyzing the impact of gender on the classifiers.

4.4. Bias Mitigation

To mitigate the gender bias, different data augmentation techniques were used. This addresses research question 2 and its sub-questions, answering how data augmentation techniques impact gender bias and whether the mitigation techniques impact the performance of the classifier for the different symptoms. Reviewing relevant background on bias mitigation allowed for the selection of appropriate data augmentation techniques to mitigate bias. Considering the small size of the dataset, the type of mitigation techniques that can be applied to reduce bias in the datasets or classifiers is limited. However, data augmentation might help decrease bias by providing the model with more training data and producing a dataset with a more balanced representation of each gender. Data augmentation is often also used to improve performance and can be achieved in different ways. While many augmentation methods exist, not all of them are label-preserving [83], which is necessary for the task of depression symptom prediction. Some label-preserving techniques include: Synonym replacement, back-translation, and oversampling [83]. These three methods were used to compare different augmentation techniques and their effect on fairness.

4.4.1. Implementation

Augmentation was implemented such that for the replicated model and the category model, sentences with depression symptoms or depression categories were balanced for gender (e.g., for each label, augmented data was added to achieve the same amount of data for each gender). As gender is not annotated for the sentences without depression symptoms, only positive sentences were augmented. For the GABDI-CV model, augmentation was achieved on the negative sentences as well. Using an augmentation technique that aims to balance data by equalizing the data for both the discriminated and privileged group has been shown to achieve the best gender fairness, and was therefore the approach for this research [84]. Since sentences can have multiple symptoms or belong to multiple categories of symptoms, no exact gender balance was achieved, but the approach ensured that the minority class had at least as many instances as the majority class for each symptom or category, which might have led to a higher representation for some labels. In the following, each augmentation technique is explained.

4.4.1.1. Synonym Replacement. An often used technique for synonym replacement is EDA, short for easy data augmentation techniques for boosting performance on text classification tasks [85]. EDA allows for synonym replacement, random insertion of synonyms, random deletion of a word and a random swap of words in the sentence. As random word deletions and random swaps are often not label preserving, only synonym replacement and random insertion were used. Synonym replacement describes the replacing of a word with a synonym. Random insertion describes the adding of a random synonym of one of the words in the sentence. The code for EDA is publicly available.[4] The recommended augmentation parameter for small datasets, 0.05, was used to determine how much of a sentence should be augmented, and the number of augmentations for each symptom or category was chosen individually in such a way as to produce a sufficient number of sentences to balance the dataset. An example of a synonym replacement can be found in Table 4.1.

Sentence	I feel the future is hopeless and that things cannot improve.
Synonym repl.	I sense the future is hopeless and that things cannot improve.

Table 4.1: Example for synonym replacement. The sentence is taken from the BDI-II questionnaire and indicates pessimism.

4.4.1.2. Back-translation. Another label-preserving method is back-translation [83]. This method involves translating a sentence into one or multiple other languages and then translating it back into its original language. This can often lead to different

[4] https://github.com/jasonwei20/eda_nlp

phrasing and formulations that have the same semantic meaning as the original sentence. While many back-translation methods use the Google API, due to the sensitive nature of depression data, a local translation system, NLLB [5] , was used. This translation tool provides translations for over 200 languages, and with its distilled versions and high performance, it was suitable to create translations for this research [86]. In order to increase the likelihood of a back-translation sentence differing from the original sentence in syntax or usage of words, the back-translation process was implemented for four different languages: German, Chinese, Arabic, and Spanish. These languages were chosen because previous studies have successfully implemented back-translation with them [83] and because of their linguistic differences, increasing the likelihood of new sentences, as even if for one language the translation remains the same, another language might provide a different outcome. This process resulted in a number of augmented sentences, which were subsampled to achieve the correct number of sentences for gender balance. Table 4.2 provides an example for back-translation.

Sentence	I feel the future is hopeless and that things cannot improve.
Back-translation	I believe the future is hopeless and things cannot get better.

Table 4.2: Example for back-translation. The sentence is taken from the BDI-II questionnaire and indicates pessimism.

4.4.1.3. Oversampling. A popular data augmentation method within fairness research is random oversampling. This method aims to balance data by increasing the data of the minority group through sampling with replacement, adding already existing instances of the group to the dataset [84]. This third method was chosen as it is well-established and assures label preservation. Moreover, it is interesting to compare oversampling to two augmentation methods producing new sentences, as differences in results could indicate whether synonym replacement and back-translation can capture the meaning and nuance of the sentences correctly, which is important for detecting symptoms of depression.

[5] https://huggingface.co/docs/transformers/v4.42.0/model_doc/nllb

5. EXPERIMENTAL RESULTS

This chapter presents the results of the significance testing regarding gender distributions of the positive sentences, performances of the classifiers, and the fairness and mitigation results. All code used for this project is freely available on GitHub.[6]

5.1. Gender Differences in the Data

A closer analysis of the gender distribution per symptom and per category for the positive sentences was conducted. This aimed to answer research question 1.1.: Are the observed gender distributions in the BDI-Sen dataset per symptom clinically validated or specific to the dataset? A distribution of male and female sentences and users per symptom can be found in Figure 5.1 and Figure 5.2. Due to the imbalance in number of sentences per gender, the proportions were compared, and the graphs show the proportion of males with a specific symptom compared to the proportion of females with a specific symptom. All severity levels were considered. The graphs show that for many symptoms, gender prevalence varies. These differences appear to be larger in the comparison of sentences, and are sometimes reversed between sentences and users.

Since the model is trained to detect symptoms per sentence, significance tests were only applied for sentence-level gender distributions. Chi-square tests were performed to determine whether there are significant differences between males and females in the number of sentences for a symptom. Table 5.1 shows the resulting p-values. Even though some symptoms have a low number of sentences (see Table 3.2), chi-square tests were chosen over the Fisher's exact test, as less than 20 percent of all symptoms have less than five instances. While this might not be ideal for specific symptoms, it makes the results more comparable. When running Fisher's exact test, similar p-values could be found, resulting in the same symptoms passing the p <0.05 threshold.

[6] https://github.com/larag11/Gender-roles-BDI-Sen

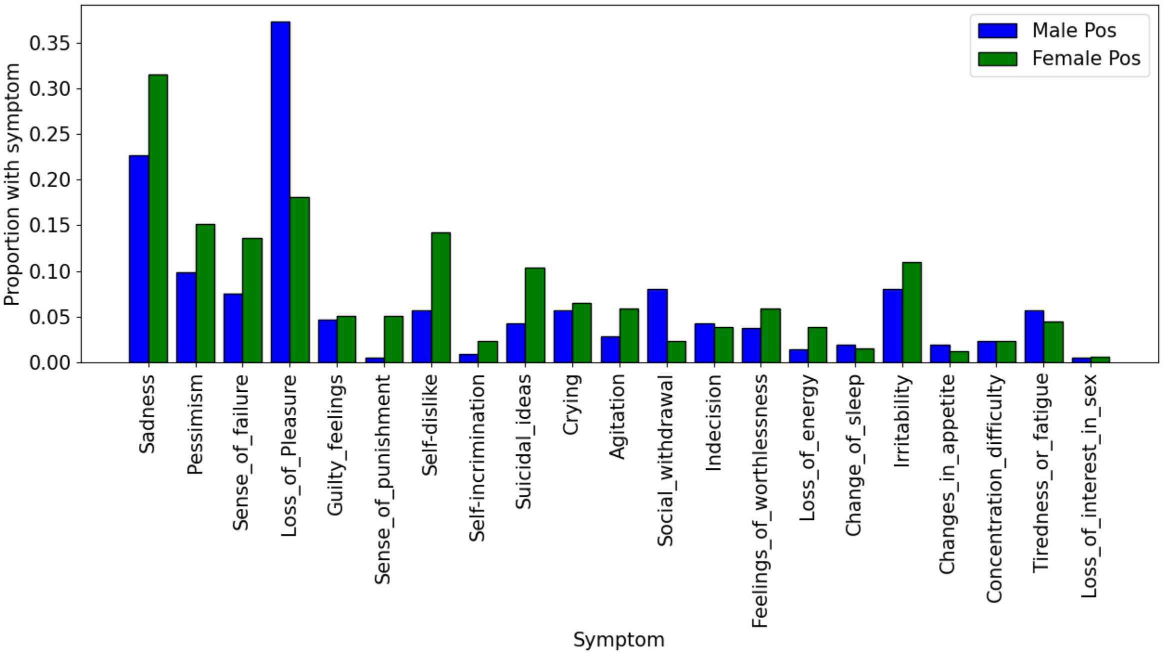


Figure 5.1: Gender distribution per symptom (sentence level). Comparing the proportion of male sentences labelled with the symptom to the proportion of female sentences labelled with the symptom.

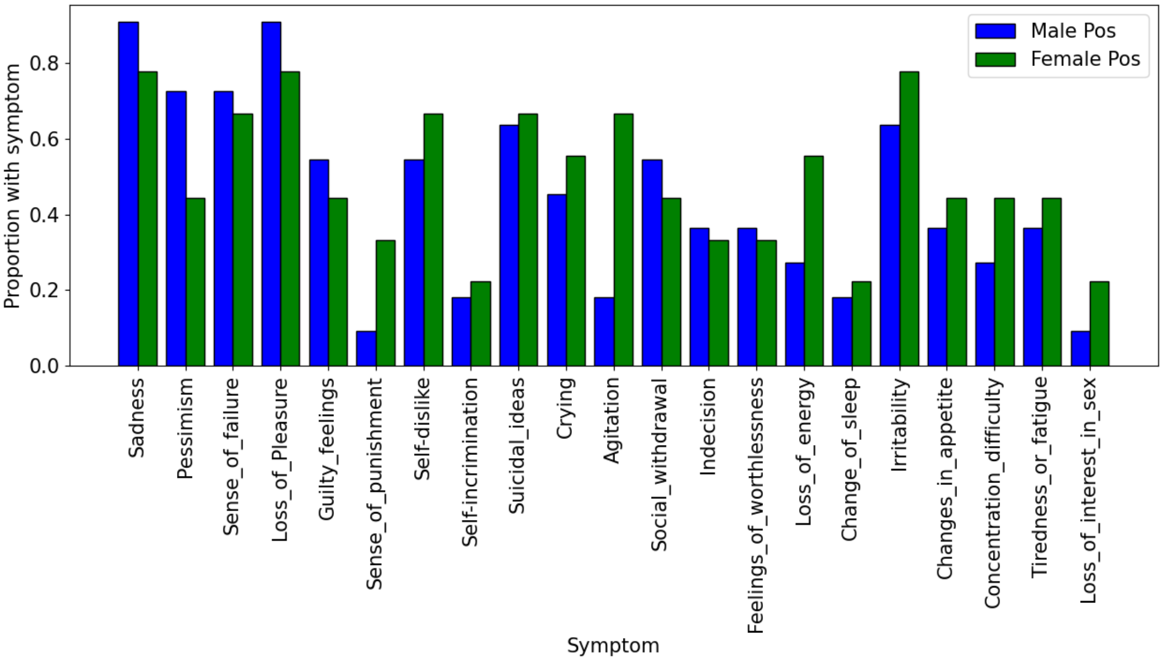


Figure 5.2: Gender distribution per symptom (user level). Comparing the proportion of male users labelled with the symptom to the proportion of female users labelled with the symptom.

Symptom	P-value
Sadness	0.032*
Pessimism	0.102
Sense of failure	0.039*
Loss of pleasure	8.882e-07*
Guilty feelings	1.0
Sense of punishment	0.007*
Self-dislike	0.003*
Self-incrimination	0.372
Suicidal ideas	0.016*
Crying	0.819
Agitation	0.144

(a) P-values symptoms 1-11.

Symptom	P-value
Social withdrawal	0.004*
Indecision	0.998
Feelings of worthlessness	0.357
Loss of energy	0.163
Change of sleep	0.986
Irritability	0.324
Changes in appetite	0.764
Concentration difficulty	1.0
Tiredness or fatigue	0.663
Loss of interest in sex	1.0

(b) P-values symptoms 12-21.

Table 5.1: P-values comparing male and female proportions per symptom. P-values are rounded to three decimal places. P-values passing the significance threshold of $p<0.05$ are indicated with a *.

For seven of the 21 symptoms, the p-value is below 0.05, suggesting a significant difference. These symptoms are: Sadness, sense of failure, loss of pleasure, sense of punishment, self-dislike, suicidal ideas, and social withdrawal. Out of these, loss of pleasure and social withdrawal have a significantly larger proportion of male sentences than female sentences, while sadness, sense of failure, sense of punishment, self-dislike and suicidal ideas contain a significantly larger number of female sentences. Table 5.2 provides an overview of the expected and observed proportions of the symptoms with significant difference, as well as their standard deviations.

Significance was also determined for the sentences when grouped into the categories. Table 5.3 presents the results, showing that the motivational and cognitive category have significantly different gender distributions. While for the motivational category, the proportion of female sentences is significantly larger, for the cognitive category, the proportion of male sentences is larger, as can be seen in Table 5.2.

Label	**Gender**	**Exp. proportion**	**Obs. proportion**	**SD**
Sadness	Male	0.281	0.226	0.449
	Female		0.315	
Sense of failure	Male	0.113	0.076	0.317
	Female		0.137	
Loss of pleasure	Male	0.255	0.373	0.436
	Female		0.181	
Sense of punishment	Male	0.033	0.005	0.178
	Female		0.05	
Self-dislike	Male	0.109	0.057	0.312
	Female		0.142	
Suicidal ideas	Male	0.080	0.043	0.271
	Female		0.104	
Social withdrawal	Male	0.046	0.08	0.208
	Female		0.024	
Motivational	Male	0.204	0.137	0.403
	Female		0.246	
Cognitive	Male	0.064	0.094	0.244
	Female		0.045	

Table 5.2: Expected and observed proportions, and standard deviations (SD) of sentences labelled with the labels that show significant differences between male and female proportions. Results are shown for symptoms with significant differences and categories with significant differences. Numbers are rounded to three decimal places.

Category	Affective	Motivational	Cognitive	Cognitive distortions	Behavioral	Physiological
P-value	0.051	0.005*	0.035*	0.051	0.262	0.580

Table 5.3: P-values comparing male and female proportions per category. P-values are rounded to three decimal places. P-values passing the significance threshold of $p<0.05$ are indicated with a *.

5.2. Classifier Performance

This section provides the performance results of the different classifiers. First, the mentalBERT model used by the researchers of the BDI-Sen dataset [7] was replicated and results were reported on the BDI-Sen and the GABDI-Sen test sets. Next, another model was trained on the BDI-Sen training set to predict symptom categories. Finally, different models were trained on the GABDI-Sen dataset, including models predicting symptoms and categories, and models trained with cross validation (CV). Their performance was compared to the models trained on BDI-Sen and each other.

Testset	Model	F1 Score	F1 Score (w.)	Precision	Recall	AUC
BDI-Sen	Replicated	0.821	0.808	0.831	0.810	0.903
	Category	0.853	0.818	0.807	0.905	0.946
	Symp-GABDI-Sen	0.682	0.622	0.832	0.577	0.788
GABDI-Sen	Replicated	0.800	0.771	0.835	0.768	0.882
	Category	0.865	0.865	0.828	0.906	0.941
	Symp-GABDI-Sen	0.545	0.452	0.773	0.420	0.708

Table 5.4: Performance scores on the BDI-Sen and GABDI-Sen test set. All scores are rounded to three decimal places and are micro averaged, apart from the weighted macro F1 score (F1 score (w.)).

5.2.1. Replicated Model

The mentalBERT model was replicated by fine-tuning it on the BDI-Sen training set using the hyperparameters provided by the BDI-Sen authors [7]. With a classification head consisting of a dropout layer and a linear layer, similar results to the original model could be reported. Results on both the BDI-Sen test set and GABDI-Sen test set are documented in Table 5.4. As micro averaged measures were used to report the original model performance in the BDI-Sen study, these measures are also used here for ease of comparison. For the fairness measures, macro weighted measures are used, and therefore the macro weighted F1 score is also provided.

5.2.2. Category Model

A model was also trained on the BDI-Sen dataset labelled with the categories that each symptom falls into. Due to the low number of physiological symptoms in the test sets, this category was excluded. The averaged performance of the model can be seen in Table 5.4. With fewer labels and more instances per label, performance slightly improves.

5.2.3. GABDI-Sen Models

Model	F1 Score (w.)
Symp-GABDI-Sen	0.495
User-ind.+BDI-Sen	0.792
Replicated	0.771
5-CV-Symp-GABDI-Sen	0.376
5-CV User-ind.+BDI-Sen	0.375
9-CV-Symp-GABDI-Sen	0.366
9-CV-Cat-GABDI-Sen	0.609

Table 5.5: Performance comparison of different models trained on the GABDI-Sen or BDI-Sen dataset. Performance is reported as a weighted macro score on the GABDI-Sen test set, or as a weighted macro averaged score for the CV models.

Using the same procedure as for the replicated model to train a model on the GABDI-Sen training set resulted in lower performance scores, as can be seen in Table 5.4 (the model is referred to as Symp-GABDI-Sen). However, a model trained on a user-independent split of the positive instances and with the negative sentences from BDI-Sen shows similar performance to the replicated model, as seen in Table 5.5. This suggests that the performance of the replicated model is not due to it learning user-specific information that it could not have learned in a user-independent split. It appears that expressions of symptoms of depression on a sentence-level are not very specific to the user and might be similar across different individuals. This could mean that the limited number of users might not impact generalizability too much.

Looking at the misclassifications of the Symp-GABDI-Sen model reveals that most misclassifications are positive sentences not classified as containing any symptom of depression, which shows that while it is not the different set of negative sentences that is misclassified, the lower number of negative sentences might have an impact on the classification. A comparison of stratified 5-fold cross validation for both models shows that performance is poor for both of them (see Table 5.5), suggesting that the better performance of the model trained on the user-independent split with the negative instances from the BDI-Sen dataset might be due to a larger number of training instances and more variety, as there is a larger number of negative sentences in BDI-Sen than GABDI-Sen, and presumably they have been written by more than 20 different users.

Due to the low performance of the Symp-GABDI-Sen and the 5-fold CV models, a different approach to train a model on GABDI-Sen was taken. Instead of splitting the dataset into training, validation and test set, the entire dataset was used for a stratified 9-fold cross validation. Stratification was implemented across positive and negative sentences, gender, and users. The number of folds was chosen to be nine as there are nine female users with depression symptoms in the dataset, allowing for one female to be in each validation split. While performance when trained to predict all 21 symptoms was still poor, when trained to predict the six categories, it improved (see Table 5.5). An analysis of the misclassifications per category shows that most misclassifications are positive sentences not correctly classified with their category (see Appendix B). These results are similar to the Symp-GABDI-Sen model, suggesting that even with fewer labels, and more instances per label, the model assigns most sentences as negative, which is the majority class. As the 9-CV-Cat-GABDI-Sen model achieves the highest performance with the GABDI-Sen dataset, it is the GABDI-Sen model used for fairness analysis and mitigation. The model will be referred to as the GABDI-CV model.

5.3. Fairness Measures

The models were analyzed for bias, both by averaging results and per label. Fairness is measured by F1 score ratios and ratios of predictive equality and equal opportunity. The sense of punishment, self-incrimination, and loss of interest in sex symptoms have been excluded from the analysis of the replicated model and category model as they do not appear in the GABDI-Sen test set. Since the minority class is male, male scores are in the numerator and female scores in the denominator of the ratio measures, meaning that results larger than 1.2 indicate a bias favoring males, and results smaller than 0.8 indicate a bias favoring females.

		Replicated	**Category**	**GABDI-CV**
F1 Score	Male	0.816	0.915	0.627
	Female	0.713	0.818	0.604
	Total	0.771	0.87	0.613
Fairness measures	F1 score ratio	1.144	1.118	1.074
	Pred. equ.	**0.648**	**1.253**	**2.704**
	Equal opp.	0.948	0.930	1.074

Table 5.6: Comparison of macro weighted fairness measures for the different classifiers. Scores indicating gender bias are highlighted in bold.

5.3.1. Replicated Model

As can be seen in Table 5.6, the replicated model shows a higher F1 score performance for males than females, dropping from 0.816 to 0.713 for females, despite the larger number of female sentences. This suggests that the model either predicts a higher number of female sentences not containing a symptom as containing the symptom, or does not find as many of the sentences containing a symptom as with males. The F1 score ratio shows that while there is a bias towards males, it does not pass the 1.2 threshold and is therefore acceptable. However, the low predictive equality score suggests that there is a large bias towards females regarding false positive predictions.

Symptom	**F1 Score**	**Precision**	**Recall**	**AUC**	**Equalized odds**	
					Pred. equ.	**Equal opp.**
Sadness*	0.792	0.760	**0.826**	0.897	**0.626**	0.955
Pessimism	0.933	0.875	**1**	0.998	**1.328**	1
Sense of Failure*	0.857	0.857	**0.857**	0.926	**1.328**	**0.750**
Loss of Pl.*	0.862	0.833	**0.893**	0.933	**0.538**	0.850
Guilty Feelings	0	0	0	0.500	**0.697**	**0.250**
Self-dislike*	0.667	0.833	0.556	0.775	**0.324**	**1.313**
Suicidal ideas*	0.941	1	**0.889**	0.944	**0.689**	0.857
Crying	0.800	1	0.667	0.833	**0.645**	**1.500**
Agitation	0.889	1	**0.800**	0.900	**0.653**	**1.250**
Social withdr.*	0.933	1	**0.875**	0.938	**0.642**	**0.667**
Indecision	0	0	0	0.500	**0.697**	**0.250**
Feel. of worthl.	1	1	**1**	1	**0.653**	1
Loss of energy	0.500	0.500	0.500	0.748	**0.339**	**0.500**
Change of sleep	0	0	0	0.500	**0.672**	**2**
Irritability	0.783	0.750	**0.818**	0.901	**0.467**	1.185
Changes in app.	0	0	0	0.500	**0.677**	1
Concentration dif.	0.500	1	0.333	0.667	**0.683**	**0.333**
Tiredness/fatigue	0.714	0.625	**0.833**	0.909	**0.430**	**1.200**

Table 5.7: Results of the replicated model per symptom. Recalls in bold show high performance, predictive equality and equal opportunity in bold indicate gender bias. Symptoms with significantly different gender distributions in the dataset are marked with a *.

Table 5.7 shows the model performance and fairness measures per symptom. Out of the 21 symptoms, 10 symptoms have a recall that is higher than 0.8, indicating that the model correctly identifies a large majority of instances for them. For all of these symptoms, the predictive equality measure shows unfairness, suggesting that the false positive rates differ between gender. Most of the measures show results below 0.8, indicating again a favor towards females regarding false positives. Four of the symptoms with a recall rate larger than 0.8 also show an equal opportunity ratio outside the 0.8 to 1.2 range, indicating a large difference. These symptoms include: Sense of failure, agitation, social withdrawal, and tiredness or fatigue.

5.3.2. Category Model

While the category model has overall better performance than the replicated model, as can be seen in Table 5.6, the model performs better for males than for females, though the averaged F1 score ratio is below 1.2. Contrary to the replicated model, the category model shows a bias favoring males regarding the averaged predictive equality score, but otherwise averaged scores suggest a fair model. A closer look at the categories reveals that the model is particularly unfair for cognitive symptoms and symptoms of cognitive distortions, as the equal opportunity measure indicates a gender bias for them (see Table 5.8). While the bias for the predictive equality shows that for some categories, the male group is favored, and for others the female, for both categories with bias in equal opportunity females appear to be favored, which is in contrast to the higher F1 scores for males.

5.3.3. GABDI-CV Model

Looking at the averaged F1 scores for the GABDI-CV model, a minor bias towards males can be found, as shown in Table 5.6. Again, the F1 score ratio suggests that this bias is not large. However, the predictive equality result shows an even higher bias towards males. Moreover, looking at the categories reveals that bias for predictive equality is not only very high on an averaged level, but also for each label, with larger results than for the category model or the replicated model, while equal opportunity results are more fair, with the only biased category being physiological (see Table 5.9). Another difference to the other two models is that all results, apart from the behavioral F1 score ratio, indicate a bias in favor of male instances. This is further underlined by the F1 scores per category, which show that model performance is better for males regarding all categories, apart from the behavioral category.

Category	**F1 Score**			**Equalized odds**	
	Male	Female	Ratio	Pred. equ.	Equal opp.
Affective	0.880	0.831	1.059	**1.421**	1.021
Behavioral	0.870	0.870	1	**0.493**	0.917
Cognitive*	**1**	0.667	**1.500**	**0.653**	**0.600**
Cog. distortions	**1**	0.750	**1.333**	**1.974**	**0.778**
Motivational*	**0.933**	0.842	1.108	**1.333**	1

Table 5.8: Fairness results of the category model per category. Results indicating gender bias are marked in bold. Categories with significantly different gender distributions in the dataset are marked with a *.

Category	**F1 Score**			**Equalized odds**	
	Male	Female	Ratio	Pred. equ.	Equal opp.
Affective	0.7181	0.506	**1.420**	**3.481**	0.915
Behavioral	0.254	**0.608**	**0.417**	**2.462**	0.972
Cognitive*	0.139	0.114	**1.218**	**2.798**	1.045
Cog. distortions	0.583	0.441	**1.321**	**1.594**	1.195
Motivational*	0.614	0.451	**1.360**	**2.211**	1.293
Physiological	0.157	0.059	**2.667**	**2.375**	**2.206**

Table 5.9: Fairness results of the GABDI-CV model per category. Results indicating gender bias are marked in bold. Categories with significantly different gender distributions in the dataset are marked with a *.

5.4. Mitigation Techniques

After identifying gender biases in the classifiers, three data augmentation techniques were used to attempt to mitigate the bias: Synonym replacement, back-translation, and oversampling. These mitigation techniques were applied to the replicated model, the category model, and the GABDI-CV model, enabling a comparison not only across data augmentation methods, but also across different classifiers and different datasets.

5.4.1. Replicated Model

Before augmenting data to balance it for gender, it was tested if increasing the instances of a symptom could have an impact on performance and fairness. Using the synonym replacement and random insertion from EDA, the training instances for the irritability symptom were augmented to a total of 232 instances, making it the symptom with the largest number of sentences. While the model trained on this dataset showed a slightly improved performance for irritability, overall, the fairness measures remained unequal. This suggests that the small number of training instances might not have a large impact on the gender bias. Given slight changes in performance for other symptoms, it appears that the model also learns relations between symptoms. For instance, the performance for agitation improved, showing that it learned more about this particular symptom as well, which could indicate a connection to irritability.

To take a closer look at whether biases can be mitigated through achieving a better data balance, data augmentation methods were applied to the symptoms sense of failure and social withdrawal. Both symptoms showed high unfairness not only for predictive equality, but also for equal opportunity, performing worse for females in terms of equal opportunity, with one false positive and one false negative female for sense of failure, and one false negative instance for social withdrawal. Synonym replacement and random insertion by EDA, back-translation, and oversampling methods were used to augment the data and balance it across gender, raising the male instances for sense of failure by 30, and the female instances for social withdrawal by 9. The results can

		Replicated	**Synonyms**	**Back Transl.**	**Oversampling**
Total F1 Score	Male	0.816	**0.838**	0.816	0.825
	Female	**0.713**	**0.713**	0.651	0.658
	Ratio	1.144	1.175	**1.253**	**1.254**
Sense of Failure	F1 Score Ratio	**1.500**	1.111	1.111	1.111
	Pred. equ.	**1.328**	**0.332**	**0.332**	**0.332**
	Equal opp.	**0.750**	**0.750**	**0.750**	**0.750**
Social withdrawal	F1 Score Ratio	**1.500**	**1.500**	**1.500**	**1.500**
	Pred. equ.	**0.642**	**0.642**	**0.642**	**0.642**
	Equal opp.	**0.667**	**0.667**	**0.667**	**0.667**

Table 5.10: Comparison of fairness measures across different data augmentation techniques for the replicated model. Total F1 scores are weighted macro scores averaged across all symptoms. Highest performing scores and scores indicating gender bias are highlighted in bold.

be found in Table 5.10. As can be seen, all data augmentation methods helped remove the F1 score bias for the sense of failure symptom. Moreover, the increase in data led to a slight improvement in overall performance, but reverted the gender bias for the predictive equality of the sense of failure symptom and could not mitigate any other gender bias. Since increasing the instances of a symptom by a large amount of data, as well as balancing gender distributions with additional data, did not decrease gender bias, the small number of instances per symptom in the test set might hinder the decrease of bias. Therefore, the bias mitigation techniques were tested on the category model next.

5.4.2. Category Model

Due to large predictive equality unfairness, all categories of the category model were augmented, such that the number of instances for each category were roughly the same for males and females. For all but the cognitive category, this led to an increase in male instances (see Appendix C for the number of added sentences per category). Table 5.11 shows the averaged results. As can be seen, the averaged predictive equality score, which indicated unfairness in the original model, improved. However, performance

		Original	**Synonyms**	**Back-Translation**	**Oversampling**
F1 Score	Male	**0.915**	0.863	0.860	0.883
	Female	0.818	0.792	0.800	**0.822**
	Total	**0.870**	0.839	0.829	0.852
Fairness measures	F1 Score Ratio	1.118	1.089	1.075	1.075
	Pred. equ.	**1.253**	1.091	0.840	0.991
	Equal opp.	0.930	0.899	0.930	0.930

Table 5.11: Comparison of macro weighted fairness measures across different data augmentation techniques for the categories model. Highest performing scores and scores indicating gender bias are highlighted in bold.

for the different categories slightly decreased for all augmented models, and for most categories, no increase in F1 scores was found (see Appendix D). Moreover, Table 5.13 shows that most of the bias remains in the new models, even though it often appears to be reduced and sometimes reversed. Some augmented models achieve more fairness, with the synonyms model eliminating predictive equality bias for the affective category, the back-translation eliminating predictive equality bias for the motivational and cognitive category, and the oversampling method eliminating predictive equality bias for the affective and the cognitive distortions category. None of the models succeeds in removing equal opportunity bias or bias for the behavioral category.

5.4.3. GABDI-CV Model

Augmentation methods were also tested on the GABDI-CV model. Appendix C provides the number of added sentences per category. As the macro weighted fairness measures for the overall results indicate, bias could be slightly reduced, but not removed (see Table 5.12). To analyze the gender distributions of F1 scores, true positive rates, and false positive rates, t-tests were conducted on each model, finding five distributions with significant differences, as indicated in Table 5.14. Moreover, t-tests between the predictive equality of the original and augmented models, as well as the equal opportunity between original and augmented models were conducted, showing that one distribution of the back-translation model was significantly different, with sig-

		Original	**Synonyms**	**Back-Translation**	**Oversampling**
F1 Score	Male	**0.627**	0.491	0.617	0.579
	Female	0.604	0.585	**0.619**	0.585
	Total	0.613	0.531	**0.618**	0.582
Fairness measures	F1 Score Ratio	1.038	0.839	0.997	0.990
	Pred. equ.	**2.704**	**2.551**	**2.037**	**1.530**
	Equal opp.	1.074	1.032	1.067	1.077

Table 5.12: Comparison of macro weighted fairness measures across different data augmentation techniques for the GABDI-CV model. Highest performing scores and scores indicating gender bias are highlighted in bold.

nificantly better fairness results for the predictive equality measure of the behavioral category. The results of the t-tests can be found in Appendix E. Overall, augmentation appears to increase performance, often for both males and females (see Appendix D). Nonetheless, performance for the categories remains low, with F1 scores for the cognitive and physiological categories below 0.5. While predictive equality biases could not be mitigated, they were often reduced. However, some augmented models introduce equal opportunity bias, suggesting that the true positive rates increased much more for one gender than the other, usually in favor of females. Moreover, some models reverse the biases, showing how sensitive biases can be to the number of sentences per gender.

Category	**Fairness measure**	**Original**	**Synonyms**	**Back-Translation**	**Oversampling**
Affective	F1 score ratio	1.059	1.045	1.009	1.029
	Pred equ.	**1.421**	1.011	**0.737**	0.884
	Equal opp.	1.021	0.985	0.985	0.985
Behavioral	F1 score ratio	1	0.900	0.900	0.900
	Pred equ.	**0.493**	**0.395**	**0.395**	**0.395**
	Equal opp.	0.917	1.008	1.008	1.008
Cognitive*	F1 score ratio	**1.500**	**1.385**	**1.846**	**1.385**
	Pred equ.	**0.653**	**0.326**	0.979	**0.326**
	Equal opp.	**0.600**	**0.600**	**0.600**	**0.600**
Cognitive dist.	F1 score ratio	**1.333**	**1.529**	**1.412**	**1.255**
	Pred equ.	**1.974**	**0.658**	**1.645**	0.987
	Equal opp.	**0.778**	**0.556**	**0.778**	**0.778**
Motivational*	F1 score ratio	1.108	**1.250**	1.039	1.188
	Pred equ.	**1.333**	**3.333**	0.889	**2.667**
	Equal opp.	1	1	1	1

Table 5.13: Comparison of fairness measures across different augmentation techniques for the categories model. Scores indicating gender bias are highlighted in bold. Categories with significantly different gender distributions in the dataset are marked with a *.

Category	**Fairness measure**	**Original**	**Synonyms**	**Back-Translation**	**Oversampling**
Affective	F1 Score Ratio	**1.420**	**1.218**	**1.274**	**1.427**
	Pred. equ.	**3.481**	**3.030**	**2.110**	**1.812**
	Equal opp.	0.915	**0.779* (F)**	0.869* (F)	0.840
Behavioral	F1 Score Ratio	**0.417**	**0.672**	1.085	0.985
	Pred. equ.	**2.462**	**1.502**	**0.636***	0.937
	Equal opp.	0.972	**1.228**	**1.438**	1.103
Cognitive*	F1 Score Ratio	**1.218**	**0.381**	**1.952**	**1.137**
	Pred. equ.	**2.798**	**1.742**	0.971	0.874
	Equal opp.	1.045	0.925	**1.248**	**1.310**
Cognitive dist.	F1 Score Ratio	**1.321**	**2.401**	**1.803**	**1.809**
	Pred. equ.	**1.594**	**3.061**	**1.919**	1.023
	Equal opp.	1.195	**1.505 (M)***	**1.360**	**1.308**
Motivational*	F1 Score Ratio	**1.360**	**1.442**	**0.682**	**1.683**
	Pred. equ.	**2.211**	**2.495**	**3.713**	**2.320**
	Equal opp.	**1.293**	0.890 (F)*	**0.782* (F)**	**1.237**
Physiological	F1 Score ratio	**2.667**	-	**0.732**	**2.133**
	Pred. equ.	**2.375**	**0.669**	0.978	**0.474**
	Equal opp.	**2.206**	**1.717**	**1.551**	**1.250**

Table 5.14: Comparison of fairness measures across different data augmentation techniques for the GABDI-CV model per category. Scores indicating gender bias are highlighted in bold. Statistically significantly different gender distributions are marked with a *. A green color indicates that the fairness measure distribution is significantly improved in the augmented model, a red color indicates significant decline. An F or M in parentheses indicates a significantly higher distribution of females (F) or males (M). For the synonyms model, no F1 score for the physiological category is given, as the female F1 score was 0.

6. DISCUSSION AND CONCLUSION

This chapter discusses the findings of this research, and places them into context with the literature. The first section addresses the first research question, discussing how a classifier trained on the BDI-Sen dataset performs for different genders and what kind of gender bias could be found. The next section addresses the second research question, asking how gender bias could be mitigated. The next two sections present ethical considerations, and limitations and future research directions. Finally, a conclusion is given.

6.1. RQ1: Gender Bias

The first research question was: How does a classifier trained on the BDI-Sen dataset perform for different genders? Different sub-questions provided grounds for a closer analysis of the dataset and classifiers. First, the gender distributions per symptom in the BDI-Sen dataset were analyzed. It needs to be noted that while the gender distribution of users is relatively even (with 11 males and nine females), the distribution of sentences appears to be imbalanced, with 212 sentences by male users, and 337 sentences by female users. As research has shown depression to be more prevalent in women than men, this could be a reflection of this finding [14] [26]. However, the slightly higher number of male users in the dataset suggests otherwise. The gender difference in number of sentences might be due to men often showing atypical symptoms of depression, or being less likely to express it [26].

Seven symptoms were found to have significantly different gender distributions: Sadness, sense of failure, loss of pleasure, sense of punishment, self-dislike, suicidal ideas, and social withdrawal. Out of these, loss of pleasure and social withdrawal appear to have a significantly larger proportion of male sentences than female sentences, while sadness, sense of failure, sense of punishment, self-dislike and suicidal ideas contain a significantly larger proportion of female sentences. This is partly in line with research findings showing that internalizing disorders are more prevalent in women [25], as symptoms such as sadness and sense of failure appear to be more internally focused. Loss of pleasure and especially social withdrawal have an external component to them, which provides confirmation of research showing that men are more likely to show external risk factors [26]. Moreover, for five out of the seven symptoms with significant difference, there are proportionally more sentences written by female users than male users, which seems to confirm the tendency of women to more often report depressed mood and symptoms of depression than men [26]. Significance on the category-level seems to align with clinical research as well, with two categories with significant differences: Motivational and cognitive. A larger female distribution is present for the motivational category, containing the symptoms pessimism and suicidal ideation, and a larger male distribution is present for the cognitive category, containing the symptoms social withdrawal and concentration difficulties. The motivational category containing largely internalized symptoms and the cognitive category containing the social withdrawal symptom seem to indicate that these biases reflect research findings.

However, other symptoms do not show a significant difference between the number of male-authored sentences and female-authored sentences. In particular, the symptoms agitation and irritability, but also loss of energy and tiredness or fatigue, which all fall into the behavioral category, do not exhibit significant differences between males and females, despite research suggesting that men are more likely to experience these symptoms [26]. This might simply indicate that female users talk proportionally more often about these symptoms when they occur than males. Moreover, for three out of these four symptoms, the proportion of male sentences containing them is larger than the proportion of female sentences containing them, and for all of these symptoms proportionally more male users are in the dataset. Therefore, the lack of a significant difference might be due to the small size of the dataset, and in particular for this category of symptoms. Overall, it can be concluded that for most symptoms, the dataset does not show significant gender differences in proportion of sentences. For the symptoms with a significant difference in proportion of female sentences and male sentences, it appears the differences can be validated by clinical research.

Second, gender fairness was analyzed, measuring F1 score, predictive equality and equal opportunity. For all three models, predictive equality showed gender biases, suggesting that false positive rates between gender differ a lot. Whereas for the replicated model trained on symptoms, predictive equality is largely in favor of females, for both the category model and the GABDI-CV model, predictive equality is mostly in favor of males, with the only exceptions being the cognitive and behavioral categories for the category model. This suggests that whether the model is trained on symptoms or on categories can impact gender fairness, with category models favoring males and symptom models favoring females. The presence of bias towards males in a dataset with more female sentences resembles findings for the D-Vlog dataset [13]. However, while authors argued that this bias was partially due to the preprocessing of the data, which might have caused the loss of information for female instances, for the BDI-Sen dataset, no sentences were cut. Therefore, there might be other factors leading to this bias.

Regarding equal opportunity, all classifiers showed more fairness. For the replicated model, the symptoms with high recall rate and unequal opportunity were: Sense of failure, agitation, social withdrawal and tiredness or fatigue. While sense of failure and social withdrawal favored females, agitation and tiredness or fatigue favored males. This is interesting considering that out of these four, only for sense of failure significantly more female instances are present, while for social withdrawal significantly more male instances are present. The symptoms with significantly different distributions not aligning with the biased symptoms might be due to the small number of instances in the test set, making a single misclassification more important.

Containing a larger number of instances, the category model shows a favoring of females in equal opportunity, for the cognitive and cognitive distortions categories. In contrast, the GABDI-CV model shows gender bias in equal opportunity for the physiological category in favor of males, suggesting that the unfairness is influenced by the data the model is trained and tested on, even if the sentences do not contain any symptoms of depression. Regarding the physiological changes, the gender bias appears to be in line with literature, which notes physical inactivity to be a symptom more prevalent in males [26] [29], and might therefore be easier to spot in them. As only two of the seven symptoms with significantly different gender distributions in the dataset show both good recall and an unfair equal opportunity score in the replicated model, and only one of the two categories with significant differences shows gender bias for equal opportunity in one of the models, the small number of instances per symptom might play a larger role in fairness measures than the distribution of symptoms. This emphasizes the need for more data in order to ensure gender fairness, especially for multi-label classification.

Third, a closer look at the performance of the classifiers for the different symptoms was taken. Comparing the F1 scores of the symptoms with significantly different distributions shows that sadness and self-dislike have a lower performance for the replicated model than the other symptoms. Both of these symptoms appear with a significantly higher proportion of female sentences than male sentences in the dataset, indicating that this might be a possible reason for the overall lower performance for females. However, performance for the category model shows no particularly low performing labels, which could suggest that a low number of instances per label might be the cause of the low performance, but with sadness being the most prevalent symptom in the dataset, this seems unlikely. In contrast to the category model, the GABDI-CV model appears to have low performance for most categories, but particularly for the cognitive and physiological categories, where total performance drops below 0.3. As the cognitive category has a significantly higher proportion of male sentences, it is interesting that overall, male performance of the GABDI-CV model is still better than female performance. However, with the cognitive and behavioral category being the categories with the smallest number of instances, the model might simply not have enough data to learn to recognize these categories successfully.

6.2. RQ2: Bias Mitigation

The second research question was: How can any present gender bias in terms of dataset imbalance, F1 score, predictive equality, or equal opportunity be mitigated? Different augmentation methods were compared to investigate how they impact gender bias. Results show that all three methods improve bias for different categories, giving no clear answer as to which one is most appropriate. One problem for the category and replicated model appears to be that while training data can be augmented, the size of the test data remains the same, and if one or two instances are still misclassified, the bias persists. Therefore, larger improvements are seen for the GABDI-CV model. However, bias often remains or reverses to favor the other gender. The averaged results of the GABDI-CV model show that oversampling appears to exhibit the highest improvement, although predictive equality remains biased. This might indicate that both augmentation with synonyms and with back-translation does not manage to capture the exact same semantic properties of the sentences they augment. For a task such as depression symptom detection, this might have a large impact on results, as there can be a lot of nuance. However, as comparing averaged results to results per category for the category model shows, even when averaged results indicate fairness, results for the individual categories might be unfair. Therefore, it is important to consider bias results on a category level.

The reversing of some bias suggests that the data augmentation does indeed impact bias, but it appears that the model is very sensitive to the amount of male or female sentences. This could indicate that it learns how depression symptoms are expressed more accurately for one gender than the other, depending on the amount of data for each gender. This is consistent with research finding that gender can be an informative feature for depression detection on social media [70] [71] [72]. However, while balancing the dataset for sensitive attributes also helped increase fairness for the CLPsych2015 dataset [12], the reversal of bias was not found for either the CLPsych2015 dataset [12] or the D-Vlog dataset [13]. Since almost all significance tests comparing fairness measures between the original and the augmented models show no significance, it is questionable whether the mitigation techniques are effective and these bias reversals replicable. Nevertheless, they might still be able to reduce bias. Despite not all bias being eliminated on the CLPsych2015 dataset [12] and the D-Vlog dataset [13] either, the repeated findings of bias and reduction through data augmentation methods emphasizes the need for depression detection datasets balanced for sensitive attributes.

Moreover, performance for the augmented methods was investigated. Especially as clinicians value good performance of a classifier even when it is unfair [87], it is important to ensure that performance does not decrease for the models trained on the augmented data. While overall performance for the models trained on the augmented data does often not improve, it appears to increase performance for some categories, especially for females. This is an interesting finding, as augmentation is also often used to increase performance [83]. However, slight decreases in performance can also be seen, and as overall performance does not increase, this suggests that while the dataset is small, it might contain sufficient information for the classifier to learn to detect depression symptoms. However, it could also suggest that more augmented data is needed to lead to an increase in performance. As the goal was merely a gender balance, for many categories the increase of data was small.

6.3. Ethical Considerations

There are several ethical considerations for this project. First, collecting data related to detecting depression symptoms poses privacy issues, as it is very sensitive information. Even though social media data is easily and publically accessible, this data needs to be handled with care. Specifically, information regarding a user's mental health and their gender is sensitive, and using it risks harming their privacy. To minimize this harm, all users were anonymized, and their usernames were not used in the training of the classifiers. In addition, user posts and comments shared in this thesis were only ones from the GABDI-Sen dataset and taken from the newly collected sentences not displaying symptoms of depression, and they were paraphrased to prevent user attribution [88]. Moreover, the project complies to all ethical and privacy regulations of Utrecht University.

Apart from it being sensitive information, inferring depression on a user can have undesirable use cases. For example, it can have a negative impact on someone if they are falsely predicted to be depressed. Identifying the expression of depressive symptoms instead of depression itself on a user could help bring more nuance to this issue, but the risks of inferring depression still need to be considered. Users could be considerably impacted by the identification of depression symptoms in their writing, and it could have an impact on their future actions. This raises the question of responsibility, especially when individuals are either wrongly identified as having depression symptoms or wrongly identified as not having depression symptoms. While in the former case, unnecessary stress might follow, in the later case, the individual might not be able to receive the support they need. It is therefore important to stress that while AI systems can be a helpful tool for detecting symptoms of depression, they cannot be held responsible for a depression diagnosis and therefore do not replace clinicians making these diagnoses.

Lastly, it needs to be considered that analyzing gender differences between men and women reinforces a binary view of gender and excludes non-binary individuals. As part of the LGBTQ+ community, non-binary individuals are often at a disadvantage [31] [32] [33]. Research into fairness especially aims to reduce inequalities, and should therefore be more focused on including non-binary individuals as well. Gender for this thesis was specifically annotated by finding phrases where users explicitly state their gender, which avoids assuming someone's gender. However, no nonbinary individuals were part of the datasets and the view of gender this thesis follows is mostly the folk view, treating gender as binary, assuming that individuals that identify as one gender show gender characteristics and behaviors that match the gender they identify with [89].

6.4. Limitations and Future Work

This study has multiple limitations. First, the small size of the dataset decreases the generalizability of the results. This is further confirmed by how few of the GABDI-CV results showed significant difference, suggesting that with different data, the gender differences could be different. Especially considering the multi-label classification task, a sufficient number of instances per label is necessary, not only to achieve good performance per label, but also to make meaningful conclusions, as a small test set is less likely to reflect the real world. The differences in performance and fairness between averaged results and results per label further emphasize this, showing that while averaged results can be relatively fair and have good performance, this does not mean performance for every symptom is high. Future work could aim to expand the BDI-Sen dataset or investigate gender differences on a similar dataset with more instances, such as the DepreSym dataset [8], if gender annotations are possible.

Another limitation to this study is that only data augmentation methods were tested. Data augmentation methods have been used for decreasing gender bias in past research on depression detection in social media [12] [13], making it a good technique to compare results, but other mitigation strategies might have achieved different outcomes. Future research could explore whether other pre-, in-, or post-processing techniques, such as reweighing, adversarial debiasing and equalized odds could further decrease the bias in the BDI-Sen dataset. This could provide more insight into the different mitigation techniques and their effectiveness for the mentalBERT classifier.

Moreover, this study does not consider the effect of other minority groups on depression symptom detection. The classifier might not only be biased in favor of males, but also in favor of whites. Next to a bias favoring whites, it has been shown that individuals of a higher socioeconomic status are more likely to be on social media, suggesting that the classifier might also learn a bias in favor of them, which is problematic as well [90]. The model might learn to cater mostly to individuals who are better off and already more likely to have access to help. Therefore, it is important to conduct more research into biases regarding other sensitive groups for depression detection in social media, as this could help ensure that these classifiers are more fair, not only regarding gender, but also other social groups.

Detecting depression symptoms is important to increase model interpretability and provide experts with models that are more easily understood. Moreover, it can ensure that models follow guidelines such as the DSM-V or diagnosis tools such as the BDI-II. With depression being a mental illness that is defined through the presence of its symptoms and depression treatment involving the reduction of these symptoms, finding symptoms in individuals can be very valuable [24]. However, datasets and models for symptom depression are limited, and more research is needed. Finding the presence of bias in datasets annotated for other depression symptoms, such as the PHQ-9, or collected on other social media, such as Twitter, could be important to gain a clearer understanding of how detection of depression symptoms can be improved.

6.5. Conclusion

This research aimed to investigate to what extent gender bias is present in the BDI-Sen dataset, how a classifier trained on it performs for different genders, and how any present gender bias can be mitigated. Some gender bias was found in the dataset and in the classifier. While most significantly different gender distributions showed higher distributions for females, most gender bias in the mentalBERT models was in favor of males, especially when measuring predictive equality. This suggests that males are more likely to be predicted as showing a symptom of depression when the symptom is not present, which might indicate that it is more difficult to distinguish between male expressions that indicate the presence of a symptom or the absence of a symptom, than for females. As research shows, women express symptoms of depression more often than men [26], but the way these symptoms are expressed might also be easier to detect. The datasets were augmented using synonyms, back-translation, and oversampling, and augmentation partially mitigated the bias. However, bias was still present after augmentation, suggesting the need for further research using other mitigation techniques apart from data augmentation. The small test set showed that the size of a dataset also plays a considerable role for gender bias analysis. With a small test set, a single misclassification can have a large impact on the fairness of a classifier. Therefore, especially for multi-label classification, it is important to have a sufficiently large dataset, as this can decrease bias, but also give a more realistic estimate of the gender bias. Machine learning models can be very beneficial tools in detecting symptoms of depression, and ensuring that they work fairly for different demographic groups could lead to an improvement of how unprivileged groups are treated and more accurate depression diagnoses from clinicians.

REFERENCES

1. Gratch, J., R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews.", *LREC*, pp. 3123–3128, Reykjavik, 2014.

2. Malhotra, A. and R. Jindal, "Deep learning techniques for suicide and depression detection from online social media: A scoping review", *Applied Soft Computing*, Vol. 130, p. 109713, 2022, https://www.sciencedirect.com/science/article/pii/S1568494622007621.

3. Low, D. M., L. Rumker, T. Talkar, J. Torous, G. Cecchi and S. S. Ghosh, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: Observational study", *J Med Internet Res*, Vol. 22, No. 10, p. e22635, Oct 2020, http://www.jmir.org/2020/10/e22635/.

4. Tigard, D. W., "Responsible AI and moral responsibility: A common appreciation", *AI and Ethics*, Vol. 1, No. 2, pp. 113–117, 2021, https://doi.org/10.1007/s43681-020-00009-0.

5. Trocin, C., P. Mikalef, Z. Papamitsiou and K. Conboy, "Responsible AI for digital health: A synthesis and a research agenda", *Information Systems Frontiers*, Vol. 25, No. 6, pp. 2139–2157, 2023, https://doi.org/10.1007/s10796-021-10146-4.

6. Zhang, Z., S. Chen, M. Wu and K. Zhu, "Symptom identification for interpretable detection of multiple mental disorders on social media", *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 9970–9985, 2022.

7. Pérez, A., J. Parapar, Á. Barreiro and S. Lopez-Larrosa, "BDI-Sen: A sentence

dataset for clinical symptoms of depression", *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2996–3006, 2023.

8. Pérez, A., M. Fernández-Pichel, J. Parapar and D. E. Losada, "DepreSym: A depression symptom annotated corpus and the role of LLMs as assessors of psychological markers", *arXiv preprint arXiv:2308.10758*, 2023.

9. Angwin, J., J. Larson, S. Mattu and L. Kirchner, "Machine bias", *ProPublica*, May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

10. Caliskan, A., J. J. Bryson and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases", *Science*, Vol. 356, No. 6334, pp. 183–186, 2017, https://www.science.org/doi/abs/10.1126/science.aal4230.

11. Blodgett, S. L., S. Barocas, H. Daumé III and H. Wallach, "Language (technology) is power: A critical survey of "bias" in NLP", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020.

12. Aguirre, C., K. Harrigian and M. Dredze, "Gender and racial fairness in depression research using social media", *16th Conference of the European Chapter of the Associationfor Computational Linguistics, EACL 2021*, pp. 2932–2949, Association for Computational Linguistics (ACL), 2021.

13. Cheong, J., S. Kuzucu, S. Kalkan and H. Gunes, "Towards gender fairness for mental health prediction", *International Joint Conferences on Artificial Intelligence Organization*, 2023, https://www.repository.cam.ac.uk/handle/1810/349873.

14. "Depressive disorder (depression)", *World Health Organization*, Mar 2023, https://www.who.int/news-room/fact-sheets/detail/depression.

15. Truschel, J., "Depression definition and DSM-5 diagnostic criteria", *Psy-*

com, Aug 2022, https://www.psycom.net/depression/major-depressive-disorder/dsm-5-depression-criteria.

16. American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (5th edition)*, American psychiatric publishing, 2013, https://doi.org/10.1176/appi.books.9780890425596.

17. Maurer, D. M., T. J. Raymond and B. N. Davis, "Depression: Screening and diagnosis", *American family physician*, Vol. 98, No. 8, pp. 508–515, 2018.

18. Evans-Lacko, S., S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, C. Benjet, R. Bruffaerts, W. Chiu, S. Florescu, G. de Girolamo, O. Gureje *et al.*, "Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the WHO World Mental Health (WMH) surveys", *Psychological medicine*, Vol. 48, No. 9, pp. 1560–1571, 2018, https://doi.org/10.1017/S0033291717003336.

19. "Depression and suicide risk in adults: Screening", *United States Preventive Services Taskforce*, Jun 2023, https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/screening-depression-suicide-risk-adults.

20. Kroenke, K., R. L. Spitzer and J. B. Williams, "The PHQ-9: Validity of a brief depression severity measure", *Journal of general internal medicine*, Vol. 16, No. 9, pp. 606–613, 2001, https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

21. Beck, A. T., "Manual for the Beck depression inventory-II", , 1996, https://cir.nii.ac.jp/crid/1370564063990947215.

22. Edelstein, B. A., L. W. Drozdick and C. M. Ciliberti, "Chapter 1 - Assessment of depression and bereavement in older adults", P. A. Lichtenberg (Editor), *Handbook of Assessment in Clinical Gerontology (Second Edition)*, pp. 3–43, Academic Press, San Diego, second edition edn., 2010, https://www.sciencedirect.com/

science/article/pii/B9780123749611100016.

23. Cohen, A., “The underlying structure of the Beck Depression Inventory II: A multidimensional scaling approach”, *Journal of Research in Personality*, Vol. 42, No. 3, pp. 779–786, 2008, https://www.sciencedirect.com/science/article/pii/S0092656607001122.

24. Borsboom, D., “A network theory of mental disorders”, *World psychiatry*, Vol. 16, No. 1, pp. 5–13, 2017, https://doi.org/10.1002/wps.20375.

25. Rosenfield, S. and D. Smith, “Gender and mental health: Do men and women have different amounts or types of problems”, *A handbook for the study of mental health: Social contexts, theories, and systems*, pp. 256–267, 2010.

26. Otten, D., A. N. Tibubos, G. Schomerus, E. Brähler, H. Binder, J. Kruse, K.-H. Ladwig, P. S. Wild, H. J. Grabe and M. E. Beutel, “Similarities and differences of mental health in women and men: A systematic review of findings in three large German cohorts”, *Frontiers in Public Health*, Vol. 9, 2021, https://www.frontiersin.org/articles/10.3389/fpubh.2021.553071.

27. Bracke, P., K. Delaruelle, R. Dereuddre and S. Van de Velde, “Depression in women and men, cumulative disadvantage and gender inequality in 29 European countries”, *Social Science and Medicine*, Vol. 267, p. 113354, 2020, https://www.sciencedirect.com/science/article/pii/S0277953620305736.

28. Barrigon, M. L. and F. Cegla-Schvartzman, *Sex, gender, and suicidal behavior*, pp. 89–115, Springer International Publishing, Cham, 2020, https://doi.org/10.1007/7854_2020_165.

29. Martin, L. A., H. W. Neighbors and D. M. Griffith, “The experience of symptoms of depression in men vs women: Analysis of the National Comorbidity Survey Replication”, *JAMA psychiatry*, Vol. 70, No. 10, pp. 1100–1106, 2013, https://doi.

org/10.1001/jamapsychiatry.2013.1985.

30. Bouman, W. P., N. Thorne and J. Arcelus, "Nonbinary gender identities", *Best Practice & Research Clinical Obstetrics & Gynaecology*, Vol. 88, p. 102338, 2023, https://doi.org/10.1016/j.bpobgyn.2023.102338.

31. James, S., J. Herman, S. Rankin, M. Keisling, L. Mottet and M. Anafi, "The report of the 2015 US transgender survey", *National Center for Transgender Equality*, 2016.

32. Reisner, S. L. and J. M. Hughto, "Comparing the health of non-binary and binary transgender adults in a statewide non-probability sample", *PLoS one*, Vol. 14, No. 8, p. e0221583, 2019, https://doi.org/10.1371/journal.pone.0221583.

33. Newcomb, M. E., R. Hill, K. Buehler, D. T. Ryan, S. W. Whitton and B. Mustanski, "High burden of mental health problems, substance use, violence, and related psychosocial factors in transgender, non-binary, and gender diverse youth and young adults", *Archives of sexual behavior*, Vol. 49, pp. 645–659, 2020, https://doi.org/10.1007/s10508-019-01533-9.

34. Cunningham, S., C. C. Hudson and K. Harkness, "Social media and depression symptoms: A meta-analysis", *Research on child and adolescent psychopathology*, Vol. 49, pp. 241–253, 2021, https://doi.org/10.1007/s10802-020-00715-7.

35. Duarte, F., "Reddit user age, gender, demographics (2024)", *Exploding Topics*, Apr 2024, https://explodingtopics.com/blog/reddit-users.

36. Proferes, N., N. Jones, S. Gilbert, C. Fiesler and M. Zimmer, "Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics", *Social Media+ Society*, Vol. 7, No. 2, 2021, https://doi.org/10.1177/20563051211019004.

37. Chowdhary, K. R., *Natural Language Processing*, pp. 603–649, Springer India, New Delhi, 2020, https://doi.org/10.1007/978-81-322-3972-7_19.

38. Boyd, R. L., A. Ashokkumar, S. Seraj and J. W. Pennebaker, “The development and psychometric properties of LIWC-22”, *Austin, TX: University of Texas at Austin*, Vol. 10, 2022.

39. Zhang, T., A. M. Schoene, S. Ji and S. Ananiadou, “Natural language processing applied to mental illness detection: A narrative review”, *NPJ digital medicine*, Vol. 5, No. 1, p. 46, 2022, https://doi.org/10.1038/s41746-022-00589-7.

40. Almeida, F. and G. Xexéo, “Word embeddings: A survey”, *arXiv preprint arXiv:1901.09069*, 2019.

41. Mikolov, T., K. Chen, G. Corrado and J. Dean, “Efficient estimation of word representations in vector space”, *arXiv preprint arXiv:1301.3781*, 2013.

42. Pennington, J., R. Socher and C. D. Manning, “Glove: Global vectors for word representation”, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

43. Garg, M., “Mental health analysis in social media posts: A survey”, *Archives of Computational Methods in Engineering*, Vol. 30, No. 3, pp. 1819–1842, 2023, https://doi.org/10.1007/s11831-022-09863-z.

44. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

45. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention is all you need”, *Advances in neural information processing systems*, Vol. 30, 2017.

46. Ji, S., T. Zhang, L. Ansari, J. Fu, P. Tiwari and E. Cambria, “MentalBERT:

Publicly available pretrained language models for mental healthcare", *Proceedings of LREC*, 2022.

47. Adarsh, V., P. Arun Kumar, V. Lavanya and G. Gangadharan, "Fair and explainable depression detection in social media", *Information Processing Management*, Vol. 60, No. 1, p. 103168, 2023, https://www.sciencedirect.com/science/article/pii/S0306457322002692.

48. Burdisso, S. G., M. Errecalde and M. Montes-y Gómez, "A text classification framework for simple and effective early depression detection over social media streams", *Expert Systems with Applications*, Vol. 133, pp. 182–197, 2019, https://doi.org/10.1016/j.eswa.2019.05.023.

49. Gallegos Salazar, L. M., O. Loyola-González and M. A. Medina-Pérez, "An explainable approach based on emotion and sentiment features for detecting people with mental disorders on social networks", *Applied Sciences*, Vol. 11, No. 22, p. 10932, 2021, https://doi.org/10.3390/app112210932.

50. Kuang, J., J. Xie and Z. Yan, "What symptoms and how long? An interpretable AI approach for depression detection in social media", *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023*, p. 4, 2023, https://aisel.aisnet.org/icis2023/dab_sc/dab_sc/4.

51. Zogan, H., I. Razzak, X. Wang, S. Jameel and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media", *World Wide Web*, Vol. 25, No. 1, pp. 281–304, 2022, https://doi.org/10.1007/s11280-021-00992-2.

52. Chancellor, S. and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review", *NPJ digital medicine*, Vol. 3, No. 1, p. 43, 2020, https://doi.org/10.1038/s41746-020-0233-7.

53. Coppersmith, G., M. Dredze, C. Harman, K. Hollingshead and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter", *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pp. 31–39, 2015.

54. Losada, D. E., F. Crestani and J. Parapar, "Overview of erisk 2019 early risk prediction on the internet", *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pp. 340–357, Springer, 2019.

55. Cohan, A., B. Desmet, A. Yates, L. Soldaini, S. MacAvaney and N. Goharian, "SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions", *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 1485—-1497, Association for Computational Linguistics, 2018, https://www.aclweb.org/anthology/C18-1126.

56. Jamil, Z., *Monitoring tweets for depression to detect at-risk users*, Ph.D. Thesis, Université d'Ottawa/University of Ottawa, 2017.

57. Yoon, J., C. Kang, S. Kim and J. Han, "D-vlog: Multimodal vlog dataset for depression detection", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 12226–12234, 2022.

58. Sap, M., G. Park, J. Eichstaedt, M. Kern, D. Stillwell, M. Kosinski, L. Ungar and H. A. Schwartz, "Developing age and gender predictive lexica over social media", *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1146–1151, 2014.

59. Losada, D. and F. Crestani, "A test collection for research on depression and language use", *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016*,

pp. 28–39, Evora, Portugal, September 2016.

60. Parapar, J., P. Martín-Rodilla, D. E. Losada and F. Crestani, "Overview of erisk 2023: Early risk prediction on the internet", *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 294–315, Springer, 2023.

61. Parikh, R. B., S. Teeple and A. S. Navathe, "Addressing bias in artificial intelligence in health care", *Jama*, Vol. 322, No. 24, pp. 2377–2378, 2019, https://doi.org/10.1001/jama.2019.18058.

62. Yeh, M.-H., B. Metevier, A. Hoag and P. Thomas, "Analyzing the relationship between difference and ratio-based fairness metrics", *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 518–528, 2024.

63. Zanna, K., K. Sridhar, H. Yu and A. Sano, "Bias reducing multitask learning on mental health prediction", *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, IEEE, 2022.

64. Barocas, S., M. Hardt and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*, The MIT Press, 2023.

65. Zhao, J., T. Wang, M. Yatskar, V. Ordonez and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2, 2018.

66. Kamiran, F. and T. Calders, "Data preprocessing techniques for classification without discrimination", *Knowledge and information systems*, Vol. 33, No. 1, pp. 1–33, 2012, https://doi.org/10.1007/s10115-011-0463-8.

67. Zhang, B. H., B. Lemoine and M. Mitchell, "Mitigating unwanted biases with adversarial learning", *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

68. Hardt, M., E. Price, E. Price and N. Srebro, "Equality of opportunity in supervised learning", D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett (Editors), *Advances in Neural Information Processing Systems*, Vol. 29, Curran Associates, Inc., 2016, https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

69. Benton, A., M. Mitchell and D. Hovy, "Multi-task learning for mental health using social media text", *arXiv preprint arXiv:1712.03538*, 2017.

70. Preoţiuc-Pietro, D., J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz and L. Ungar, "The role of personality, age, and gender in tweeting about mental illness", *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pp. 21–30, 2015.

71. Amir, S., M. Dredze and J. W. Ayers, "Mental health surveillance over social media with digital cohorts", *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 114–120, 2019.

72. de Jesús Titla-Tlatelpa, J., R. M. Ortega-Mendoza, M. Montes-y Gómez and L. Villaseñor-Pineda, "A profile-based sentiment-aware approach for depression detection in social media", *EPJ data science*, Vol. 10, No. 1, p. 54, 2021, https://doi.org/10.1140/epjds/s13688-021-00309-3.

73. Kamiran, F., A. Karim and X. Zhang, "Decision theory for discrimination-aware classification", *2012 IEEE 12th international conference on data mining*, pp. 924–929, IEEE, 2012.

74. Burger, J. D., J. Henderson, G. Kim and G. Zarrella, "Discriminating gender on Twitter", *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 1301–1309, 2011.

75. Knowles, R., J. Carroll and M. Dredze, "Demographer: Extremely simple name

demographics", *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 108–113, 2016.

76. Vashisth, P. and K. Meehan, "Gender classification using twitter text data", *2020 31st Irish Signals and Systems Conference (ISSC)*, pp. 1–6, IEEE, 2020.

77. Yang, Y.-C., M. A. Al-Garadi, J. S. Love, J. Perrone and A. Sarker, "Automatic gender detection in Twitter profiles for health-related cohort studies", *JAMIA open*, Vol. 4, No. 2, p. ooab042, 2021, https://doi.org/10.1093/jamiaopen/ooab042.

78. Liu, Y., L. Singh and Z. Mneimneh, "A comparative analysis of classic and deep learning models for inferring gender and age of Twitter users", *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications-DeLTA,*, 2021.

79. Fabian, B., A. Baumann and M. Keil, "Privacy on Reddit? Towards large-scale user classification", *European Conference on Information Systems (ECIS)*, 2015.

80. Thelwall, M. and E. Stuart, "She's Reddit: A source of statistically significant gendered interest information?", *Information processing & management*, Vol. 56, No. 4, pp. 1543–1558, 2019, https://doi.org/10.1016/j.ipm.2018.10.007.

81. Tigunova, A., P. Mirza, A. Yates and G. Weikum, "RedDust: a large reusable dataset of Reddit user traits", *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6118–6126, 2020.

82. Fosch-Villaronga, E., A. Poulsen, R. Søraa and B. Custers, "A little bird told me your gender: Gender inferences in social media", *Information Processing Management*, Vol. 58, No. 3, p. 102541, 2021, https://www.sciencedirect.com/science/article/pii/S0306457321000480.

83. Bayer, M., M.-A. Kaufhold and C. Reuter, "A survey on data augmentation for text classification", *ACM Computing Surveys*, Vol. 55, No. 7, pp. 1–39, 2022,

https://doi.org/10.1145/3544558.

84. Rančić, S., S. Radovanović and B. Delibašić, "Investigating oversampling techniques for fair machine learning models", *Decision Support Systems XI: Decision Support Systems, Analytics and Technologies in Response to Global Crisis Management: 7th International Conference on Decision Support System Technology, ICDSST 2021, Loughborough, UK, May 26–28, 2021, Proceedings*, pp. 110–123, Springer, 2021.

85. Wei, J. and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6383–6389, Association for Computational Linguistics, Hong Kong, China, Nov. 2019, https://www.aclweb.org/anthology/D19-1670.

86. Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation", *arXiv preprint arXiv:2207.04672*, 2022.

87. Sogancioglu, G., P. Mosteiro, A. A. Salah, F. Scheepers and H. Kaya, "Fairness in AI-based mental health: Clinician perspectives and bias mitigation", *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society 2024*.

88. Benton, A., G. Coppersmith and M. Dredze, "Ethical research protocols for social media health research", *Proceedings of the first ACL workshop on ethics in natural language processing*, pp. 94–102, 2017.

89. Larson, B. N., "Gender as a variable in natural-language processing: Ethical considerations", *2017 Proceedings of the First Workshop on Ethics in Natural Language Processing*, pp. 30–40, 2017, https://scholarship.law.tamu.edu/facscholar/832.

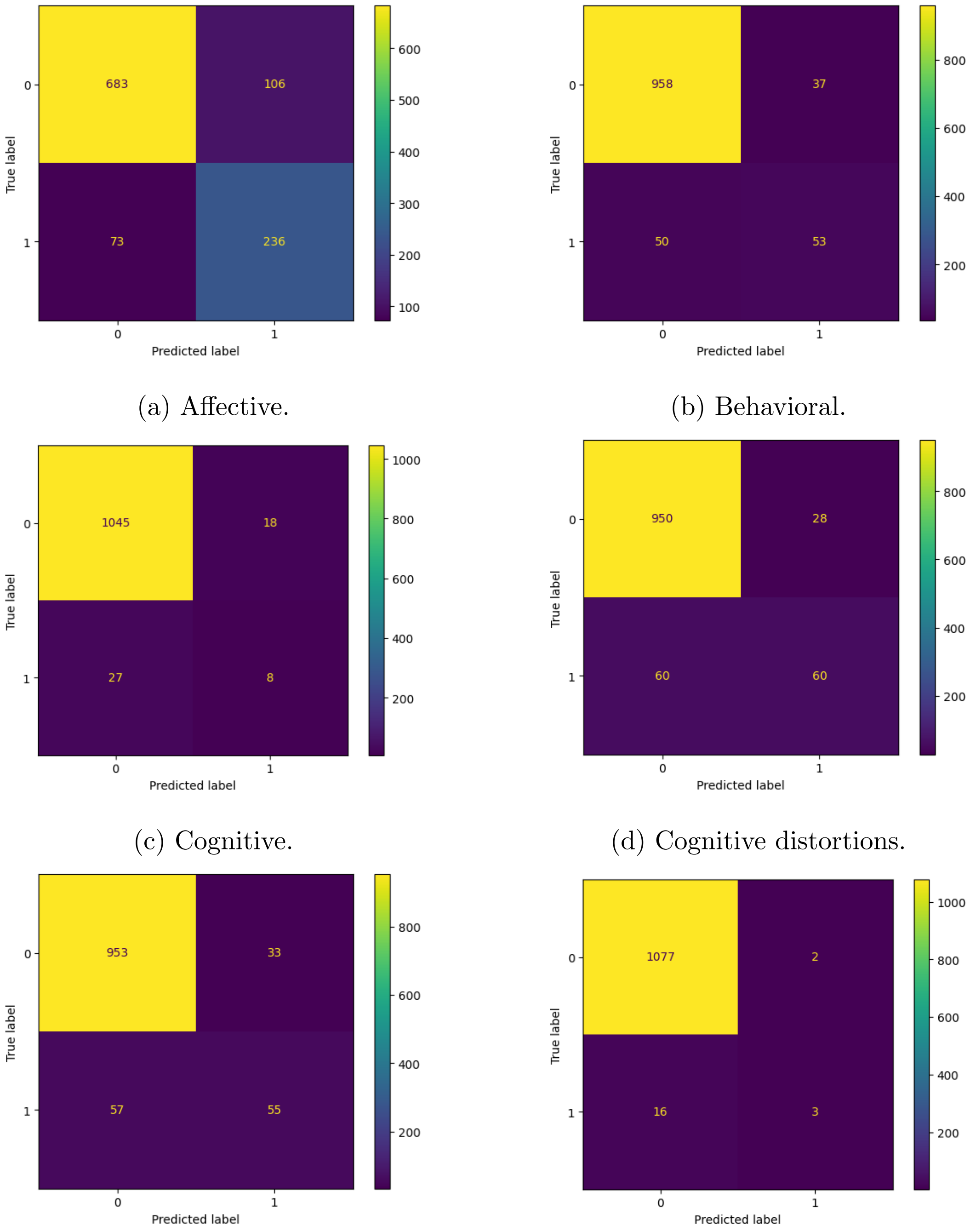
90. Hargittai, E., “Potential biases in big data: Omitted voices on social media”, *Social Science Computer Review*, Vol. 38, No. 1, pp. 10–24, 2020, https://doi.org/10.1177/0894439318788322.

APPENDIX A: Keywords for Depression Symptoms Used during the Creation of the GABDI-Sen Dataset

DSM-5 Keywords: depressed, depression, loss of interest, loss of pleasure, sad, empty, hopeless, irritable, diminished interest, decreased interest, diminished pleasure, decreased pleasure, weight loss, weight gain, decrease in appetite, increase in appetite, insomnia, hypersomnia, fatigue, fatigued

Additional BDI-II Keywords: tired, loss of energy, listless, listlessness, exhausted, exhaustion, breakdown, worthlessness, worthless, useless, inadequate, inadequacy, uselessness, guilt, regret, shame, sin, remorse, regretful, shameful, remorseful, indecisiveness, indecision, indecisive, hesitation, hesitate, uncertainty, uncertain, doubt, doubtful, death, dying, suicide, die, kill, failure, failing, punishment, penalty, punished, punish, suffering, suffer, torture, self-dislike, self-hate, self-hatred, cry, agitation, agitated, unrest, restless, turmoil, social withdrawal, loneliness, lonely, alone, difficulty concentrating, loss of interest in sex

APPENDIX B: Confusion Matrices GABDI-CV Model



(a) Affective.

(b) Behavioral.

(c) Cognitive.

(d) Cognitive distortions.

(e) Motivational.

(f) Physiological.

Figure B.1: Confusion matrices of the GABDI-CV model per category.

APPENDIX C: Number of Sentences after Augmentation

C.1. Category Model

Category	Sentences count
Affective	594
Behavioral	357
Cognitive	135
Cognitive distortions	357
Motivational	251

Table C.1: Number of sentences per category in the augmented training set for the category model.

C.2. GABDI-CV Model

Category	Sentences count
Affective	407
Behavioral	150
Cognitive	64
Cognitive distortions	178
Motivational	175
Physiological	23

Table C.2: Number of sentences per category in the augmented training set for the GABDI-CV model.

APPENDIX D: F1 Scores of the Category and CV Model

D.1. Category Model

Category	F1 Score	Original	Synonyms	Back-Translation	Oversampling
Affective	Male	**0.880**	0.863	0.846	0.863
	Female	0.831	0.825	**0.839**	**0.839**
	Total	**0.852**	0.842	0.842	0.850
Behavioral	Male	**0.870**	0.783	0.783	0.783
	Female	0.870	0.870	0.870	0.870
	Total	**0.870**	0.826	0.826	0.826
Cognitive*	Male	**1**	0.923	0.923	0.923
	Female	0.667	0.667	0.500	0.667
	Total	**0.889**	0.842	0.762	0.842
Cognitive dist	Male	**1**	0.941	0.941	0.941
	Female	**0.750**	0.615	0.667	**0.750**
	Total	**0.909**	0.800	0.800	0.848
Motivational*	Male	0.933	**1**	0.875	**1**
	Female	0.842	0.800	0.842	0.842
	Total	0.882	0.882	0.857	**0.909**

Table D.1: F1 scores for the category model.

D.2. GABDI-CV Model

Category			**Original**	**Synonyms**	**Back Transl.**	**Oversampling**
Affective	**F1 Score**	Male	**0.720**	0.623	0.625	0.656
		Female	0.506	**0.511**	0.490	0.460
		Total	**0.732**	0.675	0.728	0.717
Behavioral	**F1 Score**	Male	0.254	0.288	**0.470**	0.393
		Female	**0.608**	0.428	0.433	0.399
		Total	0.528	0.469	**0.589**	0.536
Cognitive*	**F1 Score**	Male	0.139	0.047	**0.231**	0.097
		Female	0.114	**0.124**	0.118	0.086
		Total	0.254	0.119	**0.302**	0.163
Cognitive dist.	**F1 Score**	Male	0.583	0.633	**0.728**	0.663
		Female	**0.441**	0.264	0.404	0.370
		Total	0.542	0.488	**0.646**	0.525
Motivational*	**F1 Score**	Male	0.614	0.523	0.408	**0.661**
		Female	0.451	0.363	**0.598**	0.393
		Total	0.574	0.459	0.531	**0.575**
Physiological	**F1 Score**	Male	0.157	**0.200**	0.149	0.199
		Female	0.059	0	**0.203**	0.093
		Total	0.216	0.044	0.183	**0.229**

Table D.2: F1 scores for the GABDI-CV model.

APPENDIX E: T-test Results GABDI-CV Model

		Original	**Synonyms**	**Back Trans**	**Oversampling**
Total	F1 Score	0.548	0.895	0.463	0.215
Affective	TP Rate	0.078	**0.013**	**0.042**	0.173
	FP Rate	0.281	0.343	0.446	0.907
Behavioral	TP Rate	0.791	0.669	0.797	0.998
	FP Rate	0.291	0.607	0.221	0.285
Cognitive	TP Rate	1	0.109	0.141	0.418
	FP Rate	0.921	0.422	0.435	0.236
Cog. Distortions	TP Rate	0.074	**0.045**	0.627	0.285
	FP Rate	0.458	0.487	0.487	0.761
Motivational	TP Rate	0.074	**0.006**	**0.032**	0.358
	FP Rate	0.161	0.385	0.463	0.964
Physiological	TP Rate	0.332	0.842	0.748	0.909
	FP Rate	0.170	0.944	0.312	0.239

Table E.1: P-values for each gender distribution per model. Numbers in bold pass the significance threshold $p < 0.05$.

		Synonyms	**Back Trans**	**Oversampling**
Affective	Equal opp	0.740	0.674	0.740
	FPR Ratio	0.198	0.247	0.198
Behavioral	Equal Opp	0.447	0.548	0.447
	FPR Ratio	0.104	**0.034**	0.104
Cognitive	Equal Opp	0.463	0.748	0.463
	FPR Ratio	0.293	0.286	0.293
Cog Distortions	Equal Opp	0.871	0.759	0.871
	FPR Ratio	0.089	0.926	0.089
Motivational	Equal Opp	0.979	0.212	0.979
	FPR Ratio	0.514	0.958	0.514
Physiological	Equal Opp	0.655	0.855	0.655
	FPR Ratio	0.093	0.059	0.093

Table E.2: P-values of fairness distributions between original and augmented models. Numbers in bold pass the significance threshold p <0.05 .