Utrecht University

Graduate Program in Artificial Intelligence
Utrecht University

"Use your words": Towards Gender Fairness for Multimodal Depression Detection

Utrecht University Master Thesis

by
Stan Jurgen Meyberg
6583709

First supervisor:
dr. Heysem Kaya,

Second examiner:
dr. Albert Gatt

Daily supervisor:
Gizem Soğancıoğlu MSc.

Utrecht, Wednesday 10^{th} July, 2024

ABSTRACT

Depression is a prevalent mental health disorder affecting both patients and society. The ability to identify at-risk individuals early, accurately, and without human intervention can be considered an important task as it enables timely and appropriate intervention and treatment. In recent years, numerous models have shown to be successful in detecting depression based on audiovisual cues. However, the growing use of machine learning (ML) systems for this task has raised concerns about potential biases within these systems.

This thesis explores gender fairness in multimodal depression detection using the D-Vlog dataset, which comprises vlogs derived from social media (YouTube). This study addresses the gender bias observed in previous models, particularly the performance disparity between genders. While previous studies have effectively used textual data to detect depression from social media, no research has yet applied this approach to the D-Vlog dataset. This study integrates the textual modality, experiments with various fusion strategies, and evaluates multiple bias mitigation techniques, aiming to improve both the fairness and performance of depression detection models developed using the D-Vlog dataset.

The methodology involves extracting the textual modality from the vlogs in the form of transcripts, followed by preprocessing steps to obtain both word and sentence embeddings and to prevent potential data leakage. A modality-based approach analyses the impact of the textual modality on performance and fairness, where uni- and multimodal models are trained using different modalities and fusion approaches. After applying various bias mitigation methods, the study assesses their effects on fairness and performance.

Experimental results reveal that incorporating the textual modality boosts the performance of both uni- and multimodal depression detection models, though a trade-off between performance and fairness is observed. Moreover, it was found that the choice of modality and specific feature embeddings may introduce additional gender bias into the model. In line with previous studies, the bias mitigation techniques did not consistently reduce the existing gender bias.

Despite the promising results, the study faces several limitations. The D-Vlog's collection and annotation process presents challenges such as self-disclosure bias, sampling bias, and label noise. Additionally, the model may be subject to conversational topic bias due to the collection process, despite the preprocessing steps taken to mitigate this effect.

This research provides a comprehensive assessment of the impact of incorporating the textual modality and various fusion approaches on the performance, bias, and fairness of depression detection models trained on the D-Vlog dataset. Furthermore, the research enhances the reproducibility of the experiments by open-sourcing the repository containing the re-implemented code for the D-Vlog model, addressing a gap left by previous studies that did not release their code.

Future research directions include the integration of existing video-language models or models specifically trained on multiple modalities, performing a cross-corpus validation using a clinically labelled dataset, and conducting a more in-depth analysis of textual features.

Keywords: Depression detection, multimodal AI, fairness, bias mitigation, D-Vlog, NLP

NEDERLANDSE ABSTRACT

Depressie is een veelvoorkomende psychische aandoending die zowel patiënten als de samenleving treft. Het vroegtijdig, nauwkeurig, en zonder menselijke tussenkomst identificeren van risicopersonen kan als een cruciale taak worden beschouwd, omdat dit een tijdige en passende interventie en behandeling mogelijk maakt. In de afgelopen jaren hebben meerdere modellen aangetoond succesvol te zijn in het detecteren van depressie op basis van audiovisuele signalen. Echter, de toenemende inzet van zulke machine learning (ML) systemen voor dit soort taken heeft tegelijkertijd zorgen doen toenemen over mogelijke vooroordelen of aannames (bias) die in het model zelf zitten.

Deze scriptie onderzoekt gendergelijkheid in multimodale depressie detectie modellen die gebruik maken van de D-Vlog dataset, wat een dataset omvat van vlogs afkomstig van sociale media (YouTube). Deze studie richt zich specifiek op de gender vooroordelen (bias) die in eerdere modellen zijn waargenomen, en dan met name de presentatieverschillen tussen mannen en vrouwen. Hoewel eerdere onderzoeken effectief gebruik hebben gemaakt van tekstuele gegevens om depressie uit sociale media te detecteren, heeft nog geen onderzoek deze aanpak toegepast op de D-Vlog dataset. Deze studie integreert de tekstuele modaliteit, experimenteert met verschillende fusie strategieën en evalueert meerdere bias mitigatie technieken, met als doel om zowel de eerlijkheid als de accuratesse van de depressie detectie modellen, ontwikkeld met behulp van de D-Vlog dataset, te verbeteren.

De methodologie omvat het extraheren van de tekstuele modaliteit uit de vlogs in de vorm van transcripties, gevolgd door bepaalde stappen om uiteindelijk een representatie voor zowel de individuele woorden alsmede de zinnen over te houden. Een modaliteitsgebaseerde aanpak analyseert de invloed van de tekstuele modaliteit op de accuratesse en eerlijkheid, waarbij uni- en multimodale modellen worden getraind met behulp van verschillende modaliteiten en verschillende fusie strategieën. Verder zijn verschillende bias mitigatie methodes toegepast waarna het effect van iedere methode beoordeelt is op effectiviteit.

Experimentele resultaten tonen aan dat het integreren van de tekstuele modaliteit de prestaties van zowel uni- als multimodale depressie detectie modellen verbetert, hoewel er een afweging tussen prestaties en eerlijkheid wordt waargenomen. Bovendien bleek dat de keuze van modaliteit en representatievorm van de gebruikte gegevens extra bias in het model kunnen introduceren. In lijn met eerdere studies verminderden de methodes om de vooroordelen te verminderen niet consistent deze bias.

Ondanks de veelbelovende resultaten kent het onderzoek een aantal beperkingen. Het verzamelen annotatieproces van de D-vlog brengt uitdagingen met zich mee zoals een vertekening doordat iedere persoon in de dataset zichzelf heeft gediagnosticeerd. Daarnaast kan het model onderhevig zijn aan een vertekening door een verschil in gespreksonderwerpen als gevolg van het verzamelproces, ondanks de voorbewerkingsstappen die zijn genomen om dit effect te verminderen.

Dit onderzoek biedt een uitgebreide beoordeling van de invloed van de integratie van de tekstuele modaliteit en verschillende fusie strategieën op de prestaties, bias en eerlijkheid van depressie detectie modellen getraind op de D-Vlog dataset. Bovendien verbetert het onderzoek de reproduceerbaarheid van de experimenten door het openbaar maken van de repository met de geïmplementeerde code voor het D-Vlog model.

Toekomstige onderzoeksrichtingen omvatten de integratie van bestaande video-taalmodellen of modellen die specifiek zijn getraind op meerdere modaliteiten, het uitvoeren van een cross-corpus validatie met behulp van een klinisch gelabelde dataset en het uitvoeren van een meer diepgaande analyse van tekstuele kenmerken.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Heysem Kaya, for his unwavering support, patience, and insightful critiques throughout my thesis process and for the sustenance he provided during the meetings.

I would also like to extend my gratitude to my daily supervisor Gizem Soğancıoğlu for her support and valuable advice during the thesis. I learned a great deal from both Professor Kaya and Gizem, not only during this thesis but also in the Human-centered Machine Learning course beforehand.

I am grateful to my second examiner, Professor Albert Gatt, for his patience, his feedback on the proposal, and the valuable advice he provided during the brief '10 minutes' he could spare, which really helped me in setting up the replication study.

Next, I would like to thank my fellow students from the thesis group (Mart Koek, Lara Gierschmann, and Sytse Backx) for the study sessions, their advice, and for enduring the many references I threw at them.

Furthermore, I would like to thank my friend Casper Smet for his interest in my project and his positive and constructive feedback while proofreading. Additionally, I would like to give a special mention to Thijs van den Berg, a friend with whom I worked on many different courses and projects throughout this master's program.

Lastly, I want to thank my family and friends for their unwavering support, encouragement, and for putting up with my shenanigans throughout this journey.

So long and thanks for all the cakes.

Stan Meyberg
Utrecht, 10-7-2024

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context

Depressive disorders are considered to be one of the most prevalent mental health disorders worldwide affecting both patients and society (Üstün et al., 2004). The World Health Organization (WHO) reported that approximately 280 million people in the world were depressed in 2023[1] and projected that by 2030, depression will be the most common mental disorder (Mathers & Loncar, 2006). In some situations, depression can even lead to suicide. A report by Hawton et al. (2013), points out that depression is the most common psychiatric disorder linked to suicide.

Taking into account the impact of the COVID-19 outbreak which increased the global prevalence of depression by 25%, the ability to identify at-risk individuals accurately and automatically would be of great value since this enables the application of preventive measures in the early stages (Bueno-Notivol et al., 2021; Dang et al., 2022). Therefore, the early and accurate detection of depression can be considered an important task since this allows for a timely and appropriate intervention and treatment. However, depression diagnosis remains a complicated task since this not only depends on the educational background, cognitive ability, and honesty of the subject but also on the motivation and experience of the clinicians tasked with identifying whether the subject is depressed (L. He et al., 2022). Furthermore, Craft and Landers (1998) state that traditional treatments for depression are time-consuming, costly, and sometimes even ineffective.

In recent years numerous models have been successfully introduced to the field of mental health care where the models were able to automatically and accurately detect mental health conditions based on audiovisual cues using a data-driven approach (L. He et al., 2022; Min et al., 2023; Y. Park et al., 2021; Rodrigues Makiuchi et al., 2019; Yoon et al., 2022). Not only models but also datasets have been released over the years which further contribute to the development of automatic depression detection (Gratch et al., 2014; Ringeval et al., 2019; Yoon et al., 2022). These developments hold great promise since they enable the development of systems capable of performing automatic screening with the purpose of early detection of mental health conditions which allows for proper and timely intervention.

1.2 Problem Definition

As mentioned above, the early and accurate detection of depression can be considered an important task since this allows for early and appropriate intervention and treatment. However, as Machine Learning (ML) systems are successfully applied in different contexts of the medical domain, among which is the detection of depression, concerns have been raised about the possible bias that could be contained inside of these system's data and algorithms (Shen et al., 2022; Timmons et al., 2023).

In their paper, Yoon et al. (2022) presented a multimodal depression dataset called D-Vlog, which consists of vlogs derived from social media and demonstrated that it is possible to detect whether a person is depressed based on acoustic and visual features using a Transformer-based multimodal deep learning model. However, the researchers reported a class imbalance in the dataset regarding gender, where the dataset had twice as many females as males. Subsequently, the researchers reported a big performance difference between genders where, counter-intuitively, the model performed better for the minority male class. This is surprising since learning algorithms typically over-classify the majority group when a class imbalance exists, leading to more frequent misclassification of the minority group (Johnson & Khoshgoftaar, 2019).

Furthermore, a study by Cheong, Kuzucu et al. (2023) specifically analysed the D-Vlog dataset for gender bias and confirmed the findings of Yoon et al. (2022). They also applied various bias

[1] https://www.who.int/news-room/fact-sheets/detail/depression

mitigation methods but found them inadequate in addressing the existing gender bias. They identified two main potential factors for this counter-intuitive phenomenon: either an incorrect data pre-processing approach or inherent gender differences in depression manifestation and diagnosis, suggesting that the choice of representation and modality could play a significant role.

At this time, none of the studies developing depression detection models using the D-Vlog dataset (Gimeno-Gómez et al., 2024; Sun et al., 2024; Yoon et al., 2022; Zhou et al., 2022, 2023) incorporate the textual modality even though earlier research successfully leveraged this modality for the task of depression detection. Therefore, we will introduce and add the textual modality as an additional component to the model.

Additionally, Cheong, Kuzucu et al. (2023) explored only three bias mitigation methods, while there are many more available that might effectively address the existing gender bias by targeting its root causes. Furthermore, Cheong, Kuzucu et al. (2023) only reported fairness measures for the bimodal model which means that no attempt was made to understand the impact of the unimodal models on the occurrence of gender bias in the model. Reporting fairness measures for the unimodal models could provide better insights into which modality contributes more to the bias in the system.

1.3 Aim

The first aim of this thesis is to add a textual modality to the existing model which will be done through the introduction of the transcripts from the D-Vlog dataset. Given that language use between depressed and non-depressed individuals of different genders may vary (de Jesús Titla-Tlatelpa et al., 2021; Jans-Beken, 2021), exploring the impact of this new modality on the diagnostic model regarding performance, bias, and fairness is beneficial. When adding the textual modality, both pre-trained word and sentence embeddings will be utilised. We will evaluate and compare these text-based models, as bias can also stem from the training data of these embeddings (Sogancioglu et al., 2023). Additionally, since the D-Vlog dataset already employs acoustic and visual embeddings, we will assess the performance of various fusion approaches that combine audio, video, and text modalities.

The second aim of this thesis is to address and potentially reduce the gender bias that is inside of the trained models. Potential sources of gender bias in the models will be investigated through both uni- and multimodal setups. This examination may reveal gender differences in depression manifestation and inform the selection of appropriate bias mitigation strategies.

Lastly, depending on the identified sources of bias, we will evaluate different sets of bias mitigation methods in the developed models to address and potentially reduce gender bias.
The main contributions of this thesis are as follows

- Comprehensive assessment of the impact of adding the textual modality on the overall performance, bias, and fairness of uni- and multimodal depression detection models.
- Exploration of various fusion approaches that combine audio, video, and text modalities, providing insights into which combinations yield the best results.
- Evaluation of various bias mitigation methods applied to the best performing unimodal textual model.
- Open-sourcing the repository[2] containing all my experiments as well as the code of the reconstructed and re-implemented D-Vlog model, thereby enhancing the credibility and reproducibility of this research.

1.4 Research Questions

Based on the problem definition and aim of the thesis, two research questions were formulated. Both research questions were subsequently split up into multiple subquestions. Both research

[2] https://github.com/StanMey/Master_thesis_DVlog

questions, and their respective subquestions, are shown below.

RQ 1: *Does the inclusion of the textual modality contribute to the performance of the multimodal depression detection model trained on the D-Vlog dataset in terms of the F1-score?*

> **RQ 1A**: *Which SOTA pre-trained word embeddings and sentence embeddings perform best in terms of F1-score for unimodal depression prediction on the D-Vlog dataset?*
>
> **RQ 1B**: *Which fusion approach to the audio, video and text modalities gives the best performance regarding the F1-score?*
>
> **RQ 1C**: *What is the impact of the fusion approaches of subquestion **1B** on the fairness of the model in terms of equal accuracy, equal opportunity, and predictive equality?*
>
> **RQ 1D**: *Is there a trade-off between fairness and performance for uni- and multimodal models with different fusion schemes?*

RQ 2: *To what extent can the application of bias mitigation techniques help reduce existing gender bias in the best model that has been obtained during subquestion **1B**?*

> **RQ 2A:** *What are the potential sources of gender bias (e.g., data imbalance, feature representation, modality) in the models trained on the D-Vlog dataset?*
>
> **RQ 2B:** *What bias mitigation approach is most effective in mitigating existing gender bias contained in the best model that has been obtained during subquestion **1B**?*

1.5 Readers Guide

Section 2.1 gives a general introduction to the field of fair machine learning. Building on this, section 2.2 outlines the fairness measures, providing a foundation for understanding how bias can be quantified. The performance measures used for evaluating a model are outlined in section 2.3. Section 2.4 specifies the three main bias mitigation strategies, while also providing a detailed account of the specific methods applied in this thesis. Finally, section 2.5 explains the internal workings of the models and the general approaches used for processing text whereas section 2.6 explains how bias can seep into both textual and multimodal models.

Chapter 3 reviews the existing research relevant to this thesis. Section 3.1 examines related studies on gender differences in depression detection. Section 3.2 then discusses previous approaches and models for multimodal depression detection, providing a detailed review of the D-Vlog paper and prior studies conducted specifically on the D-Vlog dataset. Lastly, section 3.3 addresses earlier research on fairness and bias in depression detection.

Chapter 4 describes the steps taken to answer each of the research questions. In section 4.2, the first research question is addressed, focusing on the impact of incorporating the textual modality into the existing model. Finally, in section 4.3, the potential bias mitigation strategies are explored and the approach to analysing gender bias is explained.

Chapter 5 covers both the statistics and the dataset split of the D-Vlog dataset in section 5.1 as well as all implementation details regarding the execution of the experiments in section 5.2.

In chapter 6 a selection of the results of all experiments conducted during this study are presented. This chapter starts by mentioning the results of the replication study in section 6.1. From section 6.2, the results of the uni- and multimodal experiments and the bias mitigation experiments

are shown.

Chapter 6 starts with answering the subquestions in subsection 7.1.1 after which both research questions are answered in subsection 7.1.2. Section section 7.2 contains the discussion whereas suggestions with regard to future research are discussed in section 7.3. Lastly, section 7.4 gives an overview of the ethical issues and concerns that may or have arisen during this study.

2 Background

This chapter starts by introducing the field of fair machine learning in section 2.1. Section 2.2 outlines the fairness measures utilised or mentioned throughout this thesis, providing a foundation for understanding how bias can be quantified. Section 2.3 describes the performance measures used for evaluating the effectiveness of a model. Section 2.4 discusses various bias mitigation approaches, specifically detailing methods that are applied in this thesis. Finally, section 2.5 explains the models and approaches used for processing text and gives examples of how bias can infiltrate these models.

2.1 Fair Machine Learning

ML algorithms have become ubiquitous in our modern society where they influence decisions made in various domains from hiring and finance to healthcare. ML algorithms can uncover patterns in historical evidence and learn from these examples to provide a more reliable basis for decision-making since these decisions are data-driven instead of being based on intuition or expertise (Barocas et al., 2023). However, learning from examples also carries serious risks since these examples may reflect historical prejudices against certain groups, cultural stereotypes, and existing demographic inequalities. Furthermore, since the algorithm will have to generalise to find these patterns they can become vulnerable to hidden or neglected biases in the data or algorithms rendering their decisions unfair and biased (Barocas et al., 2023; Mehrabi et al., 2021).

Mehrabi et al. (2021) identify two main types of bias: algorithmic bias and dataset bias. Dataset bias primarily arises from the characteristics and composition of the data (Mehrabi et al., 2021). This bias might result in biased outcomes when a ML algorithm is being trained. Mehrabi et al. (2021) give different subtypes of dataset bias such as measurement bias (bias from how features are chosen, utilised or measured), and representation bias (bias from how the population has been sampled).

Mitchell et al. (2021) decomposed dataset bias into statistical bias and societal bias. Statistical bias comes from measurement errors and non-representative sampling, whereas societal bias comes from the inherent bias that exists in the world due to past injustices and social structures.

Algorithmic bias is a result of biases introduced during the training process of the model. This type of bias is not directly present in the training data itself but is added by the model during the training process. For example, the algorithmic design choice made by the user (e.g. using certain architectures, optimisation functions, regularisations) can contribute to a biased outcome of the algorithm (Mehrabi et al., 2021).

Mehrabi et al. (2021) describe fairness as "the absence of any prejudice or favouritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" (p. 11). However, they also state that no universal definition exists giving the impact of the outlooks and preferences of different cultures as one of the reasons for the differences regarding the definition of fairness.

The field of fair ML concerns itself with uncovering and rectifying biases found in both data and algorithms in order to achieve fairness in ML systems (Mehrabi et al., 2021). Multiple fairness definitions and metrics have been put forth that can be used to assess whether ML systems are fair (Mehrabi et al., 2021). Some of these definitions and metrics will be discussed further in section 2.4. These fairness definitions and metrics can be broadly categorized into "individual" and "group" fairness based on the definition of the bias involved (Sogancioglu et al., 2023). Lastly, choosing an appropriate fairness measure for a given task is one of the most critical decisions when

designing a fair ML model or system since a less relevant measure can have damaging consequences when it is used in a real-world setting (Sogancioglu et al., 2024).

2.2 Fairness Measures

As described in section 2.1, a recent survey paper by Mehrabi et al. (2021) investigated and put forth different sources of biases that can affect AI applications as well as multiple fairness definitions and metrics. In the sections below, some of the commonly used "group" fairness measures are introduced and discussed. Furthermore, for the sake of brevity, the examples for the measures below will assume two demographic groups even though these measures do generalise to more complex cases. Lastly, individual fairness is beyond the scope of the thesis since we are mainly interested in whether both gender groups are treated equally and thus will not be covered.

2.2.1 Demographic Parity

Demographic parity, which is also known as statistical parity, requires a decision to be independent of the protected attribute. This constraint is formally defined as

$$Pr[\hat{Y} = 1|A = 0] = Pr[\hat{Y} = 1|A = 1], \tag{1}$$

where $Pr[\hat{Y} = 1|A = 0]$ corresponds to the (conditional) probability, $\hat{Y}$ stands for the predicted outcome and A stands for the protected attribute.

As can be seen from the equation, this measure is purely based on the predicted outcome while the true outcome Y is ignored. It is designed to ensure that an acceptance rate is the same regardless of the protected attribute (Barocas et al., 2023). Furthermore, the equations imply that the predicted probability has to be equal for both groups to ensure fairness. Partly for this reason, Hardt et al. (2016) considers this constraint to be flawed since statistical parity can be satisfied as long as the percentages of acceptance match. They argue that this behaviour can arise both naturally or with an unfair procedure. Moreover, they state that demographic parity would not allow the ideal predictor $\hat{Y} = Y$ under certain circumstances.

2.2.2 Disparate impact

Disparate impact is calculated by comparing the ratio of positive outcomes for both the majority and minority groups and is formally defined as

$$DI = \frac{Pr[\hat{Y} = 1|A = 0]}{Pr[\hat{Y} = 1|A = 1]}. \tag{2}$$

2.2.3 Equalised Odds

Equalised odds state that both demographic groups should have equal rates for all values of Y and is formally defined as

$$Pr[\hat{Y} = 1|A = 0, Y = y] = Pr[\hat{Y} = 1|A = 1, Y = y], y \in \{0, 1\}, \tag{3}$$

where $\hat{Y}$ stands for the predicted outcome, A stands for the protected attribute and Y stands for the true outcome.

This definition states that a predictor $\hat{Y}$ satisfies *equalised odds* with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y (Hardt et al., 2016). This means that for the outcome of $y = 1$ the constraint requires that the predictions have equal True Positive Rates (TPRs) for both values of A. Likewise, for $y = 0$ the constraint requires equalised False Positive Rates (FPRs) (Hardt et al., 2016). Equalised odds enforces the accuracy to be equally high in all demographics by punishing models that perform well only on the majority class and also allows the ideal predictor $\hat{Y} = Y$ (Hardt et al., 2016).

2.2.4 Equal Opportunity

Equal opportunity states that both demographic groups should have equal TPRs and is formally defined as

$$Pr[\hat{Y} = 1|A = 0, Y = 1] = Pr[\hat{Y} = 1|A = 1, Y = 1]. \tag{4}$$

This definition is considered a relaxation or weaker notion of equalized odds where the non-discrimination is only required to be within the "advantaged" outcome group ($y = 1$) (Hardt et al., 2016).

2.2.5 Predictive Equality

Predictive equality means that the accuracy of decisions, measured by the False Positive Rate (FPR), is consistent across both demographic groups (Corbett-Davies et al., 2017). This metric is formally defined as

$$Pr[\hat{Y} = 1|A = 0, Y = 0] = Pr[\hat{Y} = 1|A = 1, Y = 0]. \tag{5}$$

2.2.6 Equal Accuracy

Equal accuracy states that both demographic groups should have equal rates of accuracy and is defined as

$$Accuracy(\hat{Y}|A = 0) = Accuracy(\hat{Y}|A = 1), \tag{6}$$

where the accuracy refers to the accuracy of the model for each demographic group (Cheong, Spitale & Gunes, 2023).

2.3 Performance Measures

Performance measures play a crucial role in assessing the effectiveness of predictive models. To evaluate the performance of a developed model, both the predictions of the model and the ground truths are used. Using the predictions and ground truths, the following key metrics can be derived from the confusion matrix.

- True Positive (TP): denotes the number of correctly classified positive samples where the predicted label and the ground truth are an exact match.
- False Positive (FP): denotes the number of samples incorrectly classified as positive.
- False Negative (FN): denotes the number of samples incorrectly classified as negative.

From these basic metrics, it is also possible to calculate important rates such as the False Positive Rate (FPR) which denote the ratio of false positives to the actual negatives and is defined as

$$FPR = \frac{FP}{FP + TN}. \tag{7}$$

Another important rate is the True Positive Rate (TPR) which denote the ratio of true positives to the actual positives and is defined as

$$TPR = \frac{TP}{TP + FN}. \tag{8}$$

Moving beyond the key metrics and rates, the precision, recall and F1-score can offer extra insights into the model's performance.

2.3.1 Precision

Precision is a metric that indicates the proportion of instances correctly identified by the model. It reflects the model's capability to avoid misclassifying negative samples as positive (Hicks et al., 2022). The precision metric is defined as

$$Precision = \frac{TP}{TP + FP}. \tag{9}$$

2.3.2 Recall

The recall metric, also known as sensitivity, denotes the rate of correctly classified positive samples. It reflects the model's capability to find all the positive samples (Hicks et al., 2022). According to Hicks et al. (2022), this metric holds significant importance in medical studies as it is crucial to minimise the number of missed positive instances, leading to a high recall rate. The recall metrics is defined as

$$Recall = \frac{TP}{TP + FN}. \tag{10}$$

2.3.3 F1-score

The F1-score is the harmonic mean of the precision and recall metrics. This means that it penalises either metric whenever its values become extreme (Hicks et al., 2022). The F1 score is defined as

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{11}$$

2.3.4 Aggregating Performance Measures

The metrics described above can be aggregated using averaging techniques such as micro, macro, or weighted averaging. These averaging techniques prove useful for computing a more comprehensive review of the model since they can take into account the class and/or label distribution of the data on which it has been evaluated.

With micro-averaging the metrics are calculated globally by counting all true positives, false positives, and false negatives after which the final metrics are computed. Macro-averaging works exactly the other way around by calculating the metrics for each class after which the unweighted mean is calculated thereby accounting for class imbalance. However, this approach does not take label imbalance into account. Lastly, weighted averaging can be used to account for label imbalance. This approach entails calculating the metrics for each label after which the average is computed using the support.

2.4 Bias Mitigation

As described in section 2.1, dataset and algorithmic bias are the two main types of bias identified. To mitigate or to reduce the degree of the bias and to achieve ML fairness, researchers often have to apply specially designed bias mitigation methods. A reason for the need for specially designed bias mitigation methods comes from the fact that the naive approach of simply removing all discriminatory attributes, an approach known as fairness through unawareness, has been shown to not be sufficient to mitigate existing bias because of confounding variables (Pedreshi et al., 2008). Researchers can use three types of bias mitigation methods: pre-processing, in-processing, and post-processing (Cheong, Kuzucu et al., 2023; Hort et al., 2022). Each type of bias mitigation method and some of the specific bias mitigation methods used in this thesis are explained in more detail in the subsequent subsections.

2.4.1 Pre-processing

Pre-processing methods try to mitigate the bias at the data-level to prevent the bias from reaching the ML models. These methods typically involve collecting or resampling the training data to create a balanced or fair dataset (Cheong, Kuzucu et al., 2023). The data gets pre-processed early in the pipeline before training the model which means that the data is getting changed. However, this does not give direct control over the final outcome of the model.

Data Balancing

The data balancing strategy used by Yan et al. (2020) aims to create a balanced dataset by resampling the minority group. By randomly selecting samples from the minority group they obtained an augmented dataset which is balanced across all sensitive groups (Yan et al., 2020).

MixFeat

The MixFeat data augmentation method is different than the straightforward data balancing approach described above where instead of randomly oversampling samples, new synthetic samples are created by combining and mixing up existing samples with the same sensitive attribute to obtain balanced samples (Cheong, Spitale & Gunes, 2023). Given the dataset of size N where A, F, and V represent their respective modality cues, Cheong, Spitale and Gunes (2023) generate the synthetic samples as follows

$$\begin{aligned} A_k &= \lambda_A \cdot A_i + (a - \lambda_A) \cdot A_j \\ F_k &= \lambda_F \cdot F_i + (a - \lambda_F) \cdot F_j \\ V_k &= \lambda_V \cdot V_i + (a - \lambda_V) \cdot V_j \end{aligned} \tag{12}$$

where $i, j \in 1, ...N, i \neq j$ and $\lambda_A, \lambda_F, \lambda_V \sim$ Beta (0, 1).

The MixFeat approach builds upon the mixup approach by H. Zhang et al. (2018) in which they extend the training distribution by constructing linear interpolations of the feature vectors and their associated targets. However, as shown in Equation 12, the MixFeat method generates synthetic samples by only interpolating the feature vectors while leaving the associated target intact.

2.4.2 In-processing

In-processing methods are applied during the training process of the model and try to change or train the model in such a way that any bias that is in the model gets minimised through imposing some constraint (Hort et al., 2022). This approach requires both access to the model and the training data. An example of an in-processing approach is adversarial learning (B. H. Zhang et al., 2018). This type of learning trains two models at the same time: a predictor and a discriminator. While the predictor tries to accurately predict the target variable, the discriminator tries to predict the sensitive feature. The objective of this type of learning is to maximise the ability of the predictor to correctly predict the target variable while minimising the ability of the discriminator (B. H. Zhang et al., 2018). The researchers showed that by using this approach, the model learns to depend less on the sensitive attribute while almost satisfying the equal odds constraint (B. H. Zhang et al., 2018).

Loss Reweighing

Loss reweighing is a popular in-processing bias mitigation method where each object in the dataset gets assigned a different weight which the classifier takes into account during the calculation of the loss (Kamiran & Calders, 2012). The weight of each object in the dataset is based on how much that object has been deprived or favoured. It assigns a heavier weight to misclassified instances of

minority classes, thereby penalising these errors more heavily which should ultimately compensate for the bias in the dataset (Kamiran & Calders, 2012).

2.4.3 Post-processing

Post-processing methods are applied once the training of the classifier has been completed (Hort et al., 2022). These methods typically modify the predicted labels of the model, thereby directly controlling the outcome distribution, to mitigate the bias and are used when there is no access to the data and/or the model itself (Cheong, Kuzucu et al., 2023). Examples of these methods are Reject Option Classification (ROC) (Kamiran et al., 2012) and Equalised Odds post-processing (Hardt et al., 2016).

Equalised Odds Post-processing

The equalised odds method uses labelled training data, which also must include the values of the sensitive attribute, to construct a fair classifier which ensures that the equalised odds constraint described in subsection 2.2.3 gets satisfied (Hardt et al., 2016). This approach puts an additional model at the end of the existing pipeline which means that the original data and model do not get altered in any way.

2.5 Natural Language Processing

As one of the aims of this thesis is to incorporate the textual modality through the introduction of transcripts. This section will discuss the processing of natural language which, in turn, allows models to use this information to make downstream predictions.

In subsection 2.5.1, simple word embeddings will be discussed. Subsequently, in subsection 2.5.2, the attention mechanism will be explained. Following this, subsection 2.5.3 will go deeper into how the attention mechanism relates to the transformer architecture and will discuss pre-trained textual models and architectures.

2.5.1 Word Embeddings

Machine Learning (ML) and Deep Learning (DL) algorithms generally are not able to accept text directly. For this reason, the texts that serve as input to these algorithms must be transformed into a numerical vector representation so that the algorithm can process the information accordingly. These vector representations are called embeddings and can be learned.

The motivation for learning these embeddings is that by mapping them to a fixed-length encoding the meaning of words can be conserved which makes for an embedding that is representative and more informative. Word2vec (Mikolov, Chen et al., 2013) and GloVe (Pennington et al., 2014) are two examples of methods that can be used for learning high-quality word embeddings. These two types of word embeddings are static embeddings, which means that the model learns one static embedding for each word in the given vocabulary (Jurafsky & Martin, 2022).

These word embedding models are designed based on the idea that words that are used in a similar context have a similar meaning which entails that an embedding for a word can be learned by looking at their neighbouring words within a certain context. Through this approach, these types of models can learn word meaning from texts and are able to capture syntactic and semantic word similarities which can positively contribute to the performance of a model on a downstream task.

2.5.2 Attention

One of the problems encountered when using word embeddings is the fact that word embeddings are static. This means that each word has its unique embedding regardless of the context in which it is used. For this reason, one of the issues that arise when using word embeddings is the

problem of lexical ambiguity which Yepes (2017) describes as "the presence of two or more possible meanings within a single term or phrase" (p.137). One approach to mitigate this difficulty is by taking into account the context in which the word occurs while generating the embedding.

Sequence Modeling

A sentence can be thought of as a sequence of variable length that unfolds in time (Jurafsky & Martin, 2022). To model such a sequence, with the ultimate aim of determining the true meaning of a word in the sequence, the model should be able to handle both long-distance dependencies and variable-length sequences (Jurafsky & Martin, 2022). recurrent neural networks (RNNs), and their variants like gated recurrent units (GRUs) and long short-term memory (LSTM), can model the temporal nature of language through the usage of a hidden state and recurrent connections. The hidden state is a vector which represents information about all the preceding words handled thus far, making it possible to save some context and pass it on to the future time steps (Jurafsky & Martin, 2022).

Encoder-Decoder Networks

Encoder-decoder networks were originally introduced by Sutskever et al. (2014) and Cho et al. (2014). The architecture of such a model consists of both an encoder and decoder where the encoder learns to map a variable-length sequence to a fixed-length, contextualised representation after which the decoder learns to map the context vector back to a variable-length sequence (Jurafsky & Martin, 2022). Both the encoder and decoder of this model can either be a normal NN, an RNN (Cho et al., 2014), or an LSTM (Sutskever et al., 2014).

The Attention Mechanism

As described by Bahdanau et al. (2014), the issue with the encoder-decoder approach as mentioned above, is that the network needs to compress all the essential information of the input sequence into a fixed-length vector since this context vector is the only source of information for the decoder. Therefore, the final state is acting as a bottleneck making it difficult to handle longer sentences (Jurafsky & Martin, 2022). A solution to this problem, proposed by (Bahdanau et al., 2014) is a technique called "attention" and it works as an extension to the encoder-decoder network.

The attention mechanism allows the decoder to access the information from all the hidden states of the encoder instead of just the last one (Jurafsky & Martin, 2022). During the decoding process, the decoder decides which parts of the source sequence to pay attention to by giving a score to each of the hidden states of the encoder and taking a weighted sum of these vectors (Bahdanau et al., 2014). Ultimately, Bahdanau et al. (2014) shows that this approach has a positive impact on the ability of the model to yield good results on longer sentences.

2.5.3 Transformers

The transformer architecture was first proposed in the paper by Vaswani et al. (2017) where the recurrent layers that are commonly used in encoder-decoder architecture are replaced with multi-headed attention layers and the idea of positional encoders is introduced.

Overall Structure

The original transformer structure consists of an encoder-decoder structure where both the encoder and decoder can have multiple layers (Vaswani et al., 2017). The model architecture of the original transformer is shown in Figure 1. Each layer of the encoder has two sub-layers consisting of one self-attention layer and one fully connected feed-forward network layer. Each layer of the decoder has two self-attention layers and one fully connected feed-forward network layer where one of the self-attention layers has masking (Vaswani et al., 2017). Lastly, between each sub-layer, there is a

residual connection (K. He et al., 2016) which helps the model to learn and access information from lower layers followed by a normalisation operation which improves training performance (Jurafsky & Martin, 2022; Vaswani et al., 2017).

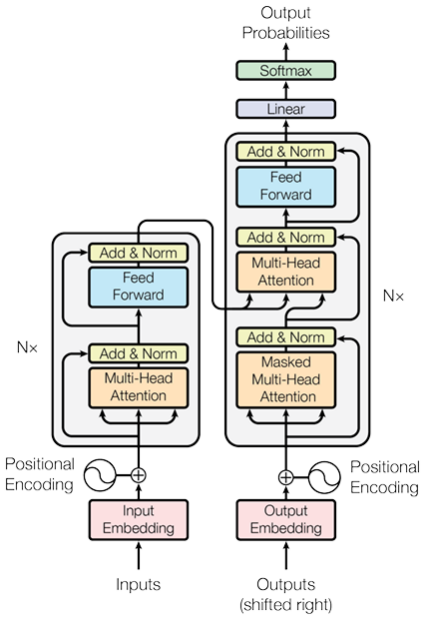

Figure 1: The architecture of the original Transformer model.
Retrieved from (Vaswani et al., 2017).

In order to perform the self-attention operation, the paper from Vaswani et al. (2017) introduces three weight matrices called the Query, Key, and Value (Q, K, and V respectively in Equation 13) which are learned during training. During the self-attention operation, these weights are used to project each input vector into a representation of its role as a key, query, or value (Jurafsky & Martin, 2022). This type of attention is called the Scaled Dot-Product Attention, where the output gets computed as a weighted sum of the values, with the weight assigned to each value determined by a compatibility function between the query and its corresponding key (Vaswani et al., 2017). Lastly, in practice, this self-attention function can be computed for an entire sequence of tokens using the following formula

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_h}})V \tag{13}$$

where Q, K, and V are the beforementioned Query, Key, and Value matrices and $\sqrt{d_h}$ is the square root of the input dimensionality of the queries and keys. Instead of performing a single self-attention operation, the researchers found it to be beneficial to perform the self-attention operation multiple times (Jurafsky & Martin, 2022; Vaswani et al., 2017). With multi-head self-attention, each sub-layer (or head) performs the self-attention function in parallel with its own set of parameters after which these heads are concatenated (Vaswani et al., 2017). This means that each head can learn different aspects of the relations among its inputs. (Jurafsky & Martin, 2022).

Positional Encoding

Since the transformer architecture does not use recurrence, information about either the relative or absolute positions of the tokens in the sequence has to be injected into the tokens beforehand (Vaswani et al., 2017). This can be done by adding positional encoding vectors to the input

embeddings at both the encoder and decoder stacks. In the original paper, Vaswani et al. (2017) uses sine and cosine functions as positional embeddings. The absence of recurrence and the addition of the positional encodings make the transformer architecture parallelizable (Jurafsky & Martin, 2022).

Cross-attention

Cross-attention is, except for the inputs, the same as the normal self-attention mechanism. The cross-attention mechanism was introduced in the paper by Vaswani et al. (2017) where the mechanism is able to combine the information from the output sequence of the encoder and the processed input sequence from the decoder. This information is combined by taking the Queries from the decoder sequence and the Keys and Values from the encoder after which the multi-head attention operation is applied Vaswani et al. (2017).

Importantly, cross-attention is not limited to sequences from a single modality. It can also be utilized to integrate and mix information from different modalities such as text, images, and audio thereby enhancing the model's ability to understand and generate better outputs. For example, Li et al. (2021) successfully applies cross-modal learning with cross-attention to use multimodal information from documents for the task of learning document representations. Additionally, Hasan et al. (2021) use cross-attention to exchange information between language, acoustic, vision, and humour-centric features in order to understand humour.

Pre-trained Language Models

The general approach when training language models is to pre-train a model on a large unlabelled dataset or corpus using different pre-training tasks in an unsupervised way followed by a smaller fine-tuning way where labelled data is used (Devlin et al., 2018). The pre-training phase is often expensive in terms of resources needed and time. However, this also means that the fine-tuning phase is relatively inexpensive and only needs a small annotated dataset to get good results (Delobelle et al., 2020).

As mentioned before, the original transformer structure consists of an encoder-decoder structure. However, in practice, there are also encoder-only or decoder-only architectures where the main difference is that the encoder-only architectures are used to learn embeddings whereas decoder-only architectures are mainly used to generate new texts.

One of the better-known encoder-only transformer-based models is the BERT model introduced by Devlin et al. (2018). BERT was pre-trained using two self-supervised tasks where the first task revolves around predicting some randomly masked tokens from the input tokens and the second task is the next sentence prediction task. Moreover, BERT uses a special classification token ([CLS]) which contains a representation of the sentence (Devlin et al., 2018). Another well-known encoder-only transformer-based model is called RoBERTa (Y. Liu et al., 2019) in which the pre-training approach of BERT was adjusted and optimised. Lastly, the aforementioned issue of lexical ambiguity can be overcome by using BERT and similar models since these models can produce contextual embeddings (Jurafsky & Martin, 2022).

One of the better known decoder-only transformer-based models is the GPT-family (Brown et al., 2020; Radford et al., 2018). These models are pre-trained and finetuned on specific tasks such as question-answering, summarization, and translation where the model ultimately is trained by predicting the next word in the sequence (Brown et al., 2020).

Sentence Embeddings

Where word embeddings transform individual words into representations that capture their meaning, sentence embeddings turn entire sentences or paragraphs into a fixed-length vector that captures the meaning of the entire sequence (Reimers & Gurevych, 2019). Two common methods of retrieving sentence embeddings are to either use the output of the first token of a BERT model (the [CLS] token) or to average the embeddings for an entire sequence (Jurafsky & Martin, 2022).

Additionally, Reimers and Gurevych (2019) proposed an approach to generating sentence embeddings where they used a pre-trained BERT model together with a siamese network structure to train the sentence embeddings on a dataset of sentence pairs.

2.6 Bias in Word Embeddings and Multimodal Models

One of the aims of this thesis is to investigate potential sources of bias in models. This section will discuss how models can learn these biases. Section 2.6.1 describes the occurrence of bias from the textual modality, while subsection 2.6.2 addresses the occurrence of bias in multimodal models.

2.6.1 Bias in Word Embeddings

Word embeddings and language models are not only able to learn word meaning from text, but they are also capable of learning implicit biases and stereotypes that are contained within the training data which is called semantic bias (Hovy & Prabhumoye, 2021; Jurafsky & Martin, 2022). These embeddings can also amplify the inherent bias from the training data which, for example, could lead to gendered terms becoming even more gendered in the embedding space (Jia et al., 2020). Moreover, the bias contained in the embeddings could seep down to downstream models and negatively impact the model's performance by favouring a certain group (Sogancioglu et al., 2023). As a result, both allocational harm and representational harm can occur when using biased embeddings (Blodgett et al., 2020). In their paper, Hovy and Prabhumoye (2021) provides an overview of how bias can be introduced during the various stages of a standard NLP system by giving examples from past research papers and sharing advice on how to counteract it.

According to Blodgett et al. (2020) allocational harms happen when a system unfairly allocates resources (e.g. loans) or opportunities (e.g. jobs) to different groups. A study from Bolukbasi et al. (2016) showed how word embeddings trained on news articles also exhibit gender stereotypes. For example, the researchers found that the embeddings suggest the analogy of 'father' is to 'doctor' as 'mother' is to 'nurse' when they checked for relational similarities. This behaviour could result in allocational harm since, for example, using biased word embeddings may contribute to resume screening tools which could favour certain demographics or genders while disadvantaging others based on the language that is being used. Moreover, Tan and Celis (2019) found that SOTA contextual models such as BERT also encode implicit racial and social biases learned from imbalanced corpora.

Blodgett et al. (2020) state that representational harms happen when a system portrays certain social groups negatively or unfavourably, belittles them or neglects to acknowledge their presence entirely. A study by Caliskan et al. (2017) showed, by looking at semantic similarity using GloVe word embeddings, that *European American names* are more likely to be closer to pleasant than to unpleasant as opposed to *African American names*. Therefore, in this case, representational harm could occur since using these embeddings could exacerbate bias against people with an African American name. Moreover, (Garg et al., 2018) specifically used the representational modelling power of embeddings and a textual corpus of 100 years of US census texts to show to what extent embeddings can capture societal shifts, historical trends and stereotypes.

2.6.2 Bias in Multimodal Models

The occurrence of bias is not limited to models trained on textual embeddings. Bias also exists in multimodal models trained using a multimodal approach where visual and language information are fused. For instance, Yin et al. (2021) showed evidence of geographic bias in two SOTA Vision-and-Language models by introducing a new commonsense reasoning evaluation benchmark. They found a significant difference in performance on images containing Western and non-Western cultural and geo-specific rituals suggesting that the commonsense learned by these models cannot generalise well across different regions. F. Liu et al. (2021) also encounter the problem of geographic bias after running benchmarks on a series of pre-trained visiolinguistic models and, as a response,

define a protocol for data collection resulting in the release of a multicultural and multilingual dataset. Ross et al. (2021) experimented on visually grounded word embeddings and found equal or larger social biases for grounded embeddings compared to ungrounded embeddings. Lastly, Cheong, Spitale and Gunes (2023) concluded from their experiments that using a multimodal approach for the task of depression detection seemed to balance out the bias of each modality resulting in more fairness. This finding is in line with the research done by Booth et al. (2021) where they used a multimodal approach for the task of hireability assessment from video interviews.

3 Related Work

This chapter begins with section 3.1, which shows some related work concerning gender differences in the detection of depression. Section 3.2 then discusses previous approaches and models for multimodal depression detection, which includes a detailed review of the D-Vlog paper and prior studies conducted on the D-Vlog dataset specifically. Lastly, section 3.3 addresses earlier research on fairness and bias in depression detection.

3.1 Gender Differences in Mental Health

Depression diagnosis remains a complicated and time-consuming task with the chance of misidentification and the fact that not all depressive individuals directly show depressive symptoms (L. He et al., 2022). In the past, there has been done some research regarding the differences in behavioural signals between depressed and non-depressed individuals. This research has subsequently led to developers choosing specific features when developing ML systems for the task of depression detection (Pampouchidou et al., 2016). First, some studies that found differences between depression and non-depression in terms of specific characteristics are described below. After that, some studies looking at specific differences in characteristics between men and women when it comes to depression are described.

3.1.1 Differences in Behavioural Signals Between Depressed and Non-depressed Individuals

Regarding audio, Min et al. (2023) extracted specific audio features such as Loudness, Spectral Flux, etc from their dataset based on prior works which studied and showed the relationship between the specific audio feature and depression. Moreover, they quantified the statistical differences in these features between the depression and non-depression groups. Lastly, they showed that their model used Loudness as one of the key indicators for detecting depression. Alghowinem et al. (2012) found that the use of MFCC, energy and intensity features led to better depression recognition rates.

Alghowinem et al. (2013) analysed head pose and movement for detecting depression using clinical interviews. They found that significant behavioural clues for diagnosing depression could be found in head movements, with slower head movements, less change of head direction and longer looking to the right or down as some of the specific behavioural clues to recognise a depressed person. Lastly, they concluded that depressed subjects expressed less positive emotions during the interviews (Alghowinem et al., 2013). The conclusion that depressed subjects expressed less positive emotions has also been supported by other studies (Min et al., 2023; Rottenberg et al., 2005). Moreover, the study by Alghowinem et al. (2012) also found that the accurate recognition of both depressed and non-depressed individuals was enhanced when positive emotions were expressed in spontaneous speech.

Regarding language use, the study done by Newell et al. (2018) showed both the increased use of self-references (such as 'I' and 'we') and the use of more negative emotion words to be fundamental markers of depression. Moreover, the study by Smirnova et al. (2018) identified distinct language patterns in individuals with mild depression compared to a healthy control group. The markers included longer written responses, more single-clause sentences, more language flow interruptions, and more responses being shifted into the past.

3.1.2 Gender Differences in Behaviours of Depressive Signals

With regards to gender-based depression detection, Stratou et al. (2013) showed that using a gender-dependent approach, in which they used gender-based classification for depression to

identify differences in performance when the two genders are separated, outperformed a gender agnostic approach (Stratou et al., 2013). The latter finding has also been concluded by the paper from Pampouchidou et al. (2016).

In their study, Stratou et al. (2013) also tried to identify an interpretable and intuitive set of predictive indicators for the task of both Depression and post-traumatic stress disorder (PTSD) detection. During their initial analysis, they reported that depressed men tend to display more frowning compared to non-depressed men and that women only display more frowning when they are non-depressed (Stratou et al., 2013). Related to this, the study from (Alghowinem et al., 2013) did not find gender differences in head movements for diagnosing depression.

In their study, Alghowinem et al. (2012) also investigated the influence of gender with regards to best audio features and found that Loudness was one of the best features for depression recognition in males while Log energy and shimmer features were better for females.

Lastly, regarding gender differences in language use, de Jesús Titla-Tlatelpa et al. (2021) analysed social media texts to identify linguistic markers of depression, leveraging word context to enhance detection. They found gender-specific classifiers to be more effective than age-specific classifiers suggesting that differences in how men and women express depression are more pronounced.

3.2 Multimodal Depression Detection

Since some research has been done in the past which have shown differences in behavioural signals between depressed and non-depressed individuals, it would make sense to use multiple modalities at the same time when trying to detect whether a person is depressed or not. Especially since multiple studies have found the benefits of combining multiple data types, or modalities regarding model performance as opposed to solely relying on one data type or modality (Alhanai et al., 2018; Cheong, Spitale & Gunes, 2023; Flores et al., 2022; Min et al., 2023). Furthermore, the literature reviews of Khoo et al. (2024) and Muzammel et al. (2021) both observed an increase in neural architectures and the usage of model-level fusion demonstrating the capability of handling high-dimensional data while modeling both internal and cross-modality relationships.

First, we describe how textual modality is applied in previous studies since this is the focus of one of the research questions. Thereafter, some studies using a multimodal approach will be referenced. Lastly, the D-Vlog paper will be explained.

3.2.1 Textual Modality

The usage of the textual modality has consistently held importance within the task of depression detection (Squires et al., 2023). Ever since De Choudhury et al. (2013) used Twitter data to measure and leverage behavioural cues (including language) to predict the risk of depression, a lot of other studies have been using the textual modality, next to other modalities, for the task of detecting depression as shown in Table 3.1. Regarding the choice of which textual features to use, Squires et al. (2023) showed in their survey that most earlier works relied on hand-crafted features while the more recent work increasingly rely on learned features such as word embeddings.

Some examples of hand-crafted features can be seen in the study by (Yang et al., 2016) where the researchers analysed the transcripts for certain personal characteristics which were later in the decision process fused with the audio and visual features. Alternatively, both Yang, Sahli et al. (2017) and Samareh et al. (2018) extract and compute basic statistics for words or sentences from the transcription files which include the number of words, the ratio of the number of laughs or sighs, number of filler words, etc. Using these kinds of features can enhance the intelligibility of depression detection models, which is a requirement for interpretable models.

Instead of designing specially hand-crafted features, with the increased prevalence of deep learning algorithms today, word embeddings are used by these algorithms to represent features numerically. Some examples of these embeddings that were specifically used in studies involving depression detection are both contextual and non-contextual word embeddings such as word2vec

(Marriwala & Chaudhary, 2023), BERT (Flores et al., 2022; J. Park & Moon, 2022; Rodrigues Makiuchi et al., 2019), and ELMo (Shen et al., 2022). Moreover, some studies also used embedding approaches that are capable of representing a sentence or even a whole document. Examples include Paragraph Vectors (Yang, Jiang et al., 2017; Yang, Sahli et al., 2017), sentence embeddings (Ray et al., 2019), and Doc2Vec (Alhanai et al., 2018).

3.2.2 Multimodal Approaches

The studies shown in Table 3.1 all perform the task of depression recognition using features extracted from different modalities such as text, video and audio. It can also be seen that almost all of these approaches use the audio modality and that the combination of the textual and audio modality is often used. Lastly, the table shows that most of the studies train and evaluate their models using the DAIC-WOZ dataset.

A crucial stage in multimodal recognition approaches involves the fusion of modalities which is the strategy for how the different modalities, or even the distinct features within the same modality, are combined. Four categories of fusion strategies are reported in the literature (Muzammel et al., 2021) and for this reason, the studies are separated by the type of fusion approach used.

Feature-level Fusion

Feature-level fusion, or early-fusion, concerns approaches which concatenate the extracted features into a single high-dimensional feature vector at an early stage (Muzammel et al., 2021). For example, in the study from Min et al. (2023) the researchers extracted audio and visual features from depression vlogs after which the XGBoost algorithm was trained using the concatenated features.

Decision-level Fusion

With decision-level fusion, or late-fusion, an integration step is performed after the decisions have been made on each modality by applying an algebraic combination rule over all the earlier predicted class labels (Muzammel et al., 2021). Samareh et al. (2018) implemented decision-level fusion by adopting an input-specific classifier for each of the visual, audio and text modalities after which the decisions were used to compute the modality-wise confidence score where the modality prediction with the highest confidence score was chosen as the final prediction. A different study by Ye et al. (2021) also adopted input-specific classifiers for the text and audio modalities where the decisions of both models were combined through a neural network to make the prediction.

Hybrid Fusion

The hybrid fusion approach performs both early- and late-fusion in the same model (Muzammel et al., 2021). An example of this approach is the study by Yang, Sahli et al. (2017) where, as the early-fusion step, the global-text features were computed and concatenated for one of the text-based models. The late-fusion consisted of fusing the predictions of the two separate audio-visual DL models with the two text-based models to ultimately use multimodal regression for the final depression prediction.

Model-level Fusion

Model-level fusion aims to learn joint representations that take into consideration the possible correlation and relationships between the representations of all the modalities involved (Khoo et al., 2024). These approaches often use different architectures for each modality after which the learned representations are concatenated followed by another neural architecture (Muzammel et al., 2021).

The studies from Alhanai et al. (2018), Marriwala and Chaudhary (2023) and J. Park and Moon (2022) all used an input-specific architecture for each of the audio and textual modality to learn a

Researcher	Dataset	Architecture	Feature Importance	Modality		
				Text	Video	Audio
Gimeno-Gómez et al., 2024	DAIC-WOZ, D-Vlog, E-DAIC	Transformer	✓		✓	✓
Min et al., 2023	self-collected	XGBoost	✓		✓	✓
Fang et al., 2023	DAIC-WOZ	BiLSTM, LSTM, attention fusion network		✓	✓	✓
Yadav and Sharma, 2023	DAIC-WOZ	BiGRU		✓		
Marriwala and Chaudhary, 2023	DAIC-WOZ	CNN, BiLSTM		✓		✓
Zhou et al., 2023	D-Vlog	CAM-BiLSTM			✓	✓
Zhou et al., 2022	D-Vlog, EATD-Corpus	TAMFN			✓	✓
Yoon et al., 2022	D-Vlog	Transformer, cross-attention			✓	✓
J. Park and Moon, 2022	DAIC-WOZ	BERT-CNN, CNN-BiLSTM		✓		✓
Flores et al., 2022	DAIC-WOZ	BiLSTM, self-attention	✓	✓	✓	✓
Shen et al., 2022	DAIC-WOZ, EATD-Corpus	GRU, LSTM		✓		✓
Ye et al., 2021	self-collected	Transformer, DNN		✓		✓
Toto et al., 2021	DAIC-WOZ	BiLSTM, self-attention	✓	✓		✓
Niu et al., 2020	AVEC2013, AVEC2014	2D & 3D CNN, LSTM, attention			✓	✓
Kaya et al., 2019	E-DAIC	GRU, KELM		✓	✓	✓
Ray et al., 2019	DAIC-WOZ	BiLSTM, attention	✓	✓	✓	✓
Rodrigues Makiuchi et al., 2019	E-DAIC	GCNN-LSTM, CNN-LSTM		✓	✓	✓
Alhanai et al., 2018	DAIC-WOZ	LSTM		✓		✓
Samareh et al., 2018	DAIC-WOZ (AVEC2017)	Random Forest	✓	✓	✓	✓
Yang, Jiang et al., 2017	DAIC-WOZ (AVEC2017)	DNN, Deep CNN		✓	✓	✓
Yang, Sahli et al., 2017	DAIC-WOZ	DCNN-DNN, Random Forest, Paragraph Vector	✓	✓	✓	✓
Yang et al., 2016	DAIC-WOZ	Decision Tree	✓	✓	✓	✓
Alghowinem et al., 2016	self-collected	SVM			✓	✓
Dibeklioğlu et al., 2015	self-collected	Logistic Regression	✓		✓	✓

Table 3.1: An overview of recent research papers with regards to multimodal depression detection.

representation after which these learned representations were concatenated and fed into a neural network to make the final prediction. Rodrigues Makiuchi et al. (2019) used a similar approach with the difference that the representations did not come from the end of modality-specific models but instead used the embeddings obtained from the first dense layer of each modality-specific model.

To further support the neural architecture at the end of the model by learning the relationships between the representations of all the modalities involved, some studies have used the attention mechanism to aid during the fusion process. Fang et al. (2023) proposes the attentional fusion network (AttFN) which fuses the features of the audio, video and text modalities by training a separate attention weight vector using the concatenated representations and the softmax function which evaluates the importance of each modality and multiplies this attention vector with the

vertically stacked representations of each modality. Shen et al. (2022) used a similar approach by training a separate attention weight vector with the difference that their attention scores were calculated for the concatenated representations. Zhou et al. (2022) use a different approach where three different feature embeddings (acoustic, visual, and early fusion) are extracted from the acoustic and visual input features. Hereafter, the time-aware attention multimodal fusion (TAMF) module fuses these features by learning the temporal importance of each combination of embeddings using a mixed attention vector which guides the embeddings during the fusion process.

The studies mentioned above use the attention method during the fusion approach. However, the study by Toto et al. (2021) uses the self-attention mechanism to learn task-specific representations from both the audio and text modalities where the attention layers are placed directly after the hidden layer of the BiLSTMs. Hereafter, the embeddings are fused using concatenation after which the classification occurs. The study by Flores et al. (2022) expands on this study by adding the video modality to the existing architecture which extracts facial features. The results of their experiments showed their model outperforming the original model. Yoon et al. (2022) also uses self-attention through the usage of an unimodal transformer model to learn significant cues within a specific modality as well as using cross-attention to learn about important relationships across the representations of the modalities. This architecture, dataset, and the performances of other studies using this dataset are described in more detail in subsection 3.2.3.

3.2.3 D-Vlog

The D-Vlog architecture, as proposed by Yoon et al. (2022), also makes use of model-level fusion where they use unimodal Transformer encoders for each modality to generate representations after which the cross-attention method is used to learn the relationships between both latent representations. Moreover, as shown in Table 3.1, many recent works involved using the DAIC-WOZ dataset which was captured in a clinical setting through interviews. Having access to the D-Vlog, which includes the behaviour of depressed individuals in their daily lives, enables us to use this dataset to train better models which may prove useful to accurately screen for depressed individuals on social media.

Since this thesis builds upon both the dataset and model designed and developed by Yoon et al. (2022), this section aims to give both a condensed summary of said paper as well as an overview of performances of past models trained on the D-Vlog dataset.

Dataset

The original D-Vlog dataset (Yoon et al., 2022) includes a collection of English YouTube videos posted between the 1st of January 2020 and the 31st of January 2021. The dataset consists of 555 depression and 406 non-depression vlogs where the depression vlogs were collected using 'depression daily vlog', 'depression journey', 'depression vlog', 'depression episode vlog', 'depression video diary', 'my depression diary', and 'my depression story' as the search keywords (Yoon et al., 2022). Likewise, the non-depression vlogs were collected using search keywords such as 'daily vlog', 'get ready with me (grwm) vlog', 'haul vlog', etc. Hereafter, each vlog was annotated by one of four college students who were recruited and trained on the annotation criteria. Each annotator had to identify whether the video had a 'vlog' format and whether the speaker had depression or not using certain criteria. The exact data statistics of the D-Vlog dataset for this thesis are discussed further in section 5.1.

Input Features

In the original model of the D-Vlog paper, a multimodal model is being used which uses both acoustic and visual features (Yoon et al., 2022).

The acoustic features were extracted using the open-source OpenSmile toolkit (Eyben et al., 2010). Using this toolkit, the researchers extracted 25 low-level acoustic descriptors (LLDs) from

the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2015) by segmenting each vlog into segments of one second and averaging the values of all LLDs per segment (Yoon et al., 2022).

For the visual features, the researchers extracted 68 facial landmarks (i.e., the x and y coordinates) from each frame in the vlog using the dlib (King, 2009) open-source software (Yoon et al., 2022). Facial landmark detection involves the identification of key facial features through shape prediction techniques on a given face. These landmarks were used directly, without any kind of transformation, as input features to the model.

To anonymize each speaker and avoid leakage of privacy, Yoon et al. (2022) only provides the non-intuitive extracted input features.

Architecture

To leverage multimodal inputs of video, Yoon et al. (2022) use two unimodal Transformer encoders that generate unimodal representations from the acoustic and visual input features after which the multimodal Transformer encoder takes both representations, combines them and gives back the final representation. This final representation is then used to predict whether the individual in the vlog is depressed. An overview of the whole architecture is shown in Figure 2.

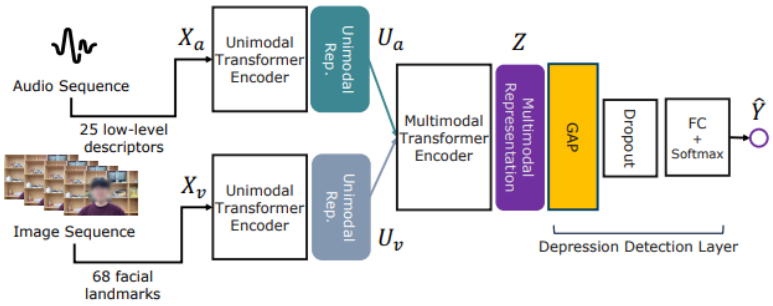

Figure 2: An overview of the architecture of the multimodal model used by Yoon et al. (2022). Retrieved from (Yoon et al., 2022).

To not be limited to either visual or acoustic modalities, Yoon et al. (2022) uses an unimodal Transformer encoder to generate representations for each modality. Each unimodal Transformer uses the original Transformer encoder as described by Vaswani et al. (2017) and can focus on significant cues within each modality.

To combine the representations generated by the unimodal Transformers, Yoon et al. (2022) employs a multimodal Transformer encoder which uses a cross-attention module to learn about the relationships between both modalities. The cross-attention approach was inspired by the paper from Hasan et al. (2021) in which humour was recognised from multimodal features using both modality-specific encoders for attending to each modality after which the cross-attention setup was used to understand the latent information between both representations. The cross-attention module still uses the Query, Key, and Value as described in subsection 2.5.3. However, instead of computing the self-attention scores on a single input sequence, with cross-attention the Query and Key/Value vectors are computed using a combination of input sequences from both unimodal representations as can be seen in Figure 3.

Lastly, The D-Vlog model employs a depression detection layer which uses the multimodal representation derived from the earlier steps as input to make the final depression prediction (Yoon et al., 2022).

Related Work on D-Vlog

Despite being introduced only recently in 2022, the D-Vlog dataset has already inspired the development of both uni- and multimodal models within the research community. These models along

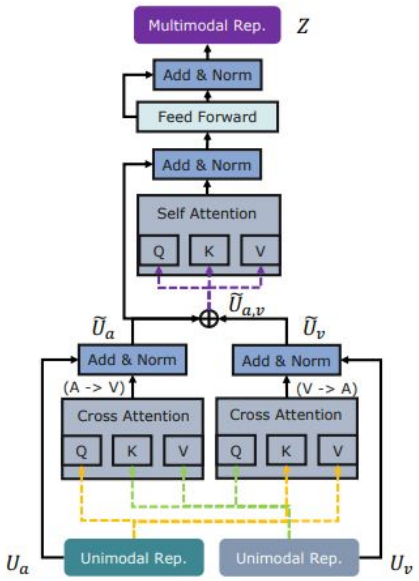

Figure 3: An illustration of the implementation of the cross-attention module in the multimodal transformer. Adapted from (Yoon et al., 2022).

with the reported performance measures are shown in Table 3.2. However, since previous studies reported their results with varying decimal precision, we standardised our reporting in the table based on the study (Gimeno-Gómez et al., 2024) with the fewest decimal places. Consequently, some of our results may be rounded up or down to maintain consistency.

The table shows that the recently published approach by Sun et al. (2024) significantly outperforms the earlier papers while only using the acoustic features. In their approach, they utilise a graph convolutional network (GCN) to obtain embeddings from the audio features while also explicitly introducing emotional features through a self-attention mechanism where ultimately both obtained embeddings are successfully combined by using another GCN (Sun et al., 2024). However, it must be taken into account that Sun et al. (2024) deviates from the other papers shown in Table 3.2 with regards to their evaluation approach. They evaluated their model using ten-fold cross-validation and use a 8:1:1 split ratio with regard to the training, validation, and testing subsets. Moreover, the researchers do not specify whether these folds share no YouTube channel with each other which could introduce some data leaking.

Researcher	Architecture	Modality	Original split	Performance measures (weighted)		
				Precision	Recall	F1-score
Sun et al., 2024	GRU, GCN	A	✗	0.92*	0.98*	0.95*
Gimeno-Gómez et al., 2024	Transformer	V, A	✗	0.74	0.84	0.78
Zhou et al., 2023	CAM-BiLSTM	V, A	✓	0.67	0.61	0.67
Zhou et al., 2022	TAMFN	V, A	✓	0.66	0.67	0.66
Yoon et al., 2022	Transformer, cross-attention	V, A	✓	0.65	0.66	0.64

Table 3.2: Depression detection results of related work on the D-Vlog dataset. To ensure consistency between all studies, all results are standardised to two decimal places, rounding some of the results in the process. Performance metrics that use a different experimental protocol are marked with an asterisk (*).

3.3 Fairness and Bias in Mental Health Analysis

Even though past research into the application of AI for the investigation and understanding of mental health, and especially detecting depression, has shown promising results, research in bias and fairness for mental health using ML has been limited (Cheong, Kuzucu et al., 2023; Cheong, Spitale & Gunes, 2023). However, these issues have received more attention in the last few years resulting in more papers. For example, recently a paper by Timmons et al. (2023) was published in which a call to action is presented to guide the development of *fair-aware* AI in psychological science.

Some of the papers found during the literature review analysed gender bias by only comparing the difference of the model using the performance metrics between both subgroups without using any kind of fairness measures (Min et al., 2023; Rejaibi et al., 2022; Yoon et al., 2022). Through their study, Rejaibi et al. (2022) found that gender does affect the performance of the model when they trained and validated the model for a specific gender. The quick overview of the papers found during the literature review regarding bias and fairness can be seen at Table 3.3.

3.3.1 Fairness and Bias Regarding the D-Vlog Dataset

Cheong, Kuzucu et al. (2023) not only applied different bias mitigation methods to the D-Vlog dataset, and the Depresjon and Psykose datasets, but also used different, carefully selected, fairness measures to analyse the degree to which the existing bias can be mitigated. They found that dataset bias was present in the D-Vlog dataset while the baseline model was deemed acceptably fair regarding algorithmic bias. Lastly, for the D-Vlog dataset, the researchers found that none of the applied bias mitigation methods (Data Augmentation, Loss Re-weighing, and ROC) were consistently effective at mitigating bias across the chosen fairness measures and all seem to worsen the bias present. Both the loss of information due to the data preprocessing method used and the difference in depression manifestation and diagnosis between men and women were mentioned by the researchers as possible factors for the behaviour described above (Cheong, Kuzucu et al., 2023). The paper by Bailey and Plumbley (2021) examined the DAIC-WOZ dataset and found that the dataset contained gender bias, which could negatively affect the accuracy of ml models trained on audio features. Moreover, they showed that features such as the mel-spectrogram can hold information that can serve as a proxy so that the gender bias in the dataset can be utilised.

By using the pre-processing data re-distribution approach of data subsampling, the researchers were able to remove the gender bias from the data. Lastly, their results suggest that using raw audio may be more robust to gender bias as opposed to using mel-spectogram features (Bailey & Plumbley, 2021). The latter is especially interesting since the D-Vlog uses the mel-spectogram as one of its features for the audio modality (Yoon et al., 2022).

Researcher	Dataset	Modality	Fairness measures	Fairness		Bias mitigation		
				Gender	Race	Pre	In	Post
Min et al., 2023	self-collected	V, A	F1	✓				
Cheong, Kuzucu et al., 2023	D-Vlog, Depresjon, Psykose	V, A	F1, SP, EOpp, EOdd, EAcc	✓		✓	✓	✓
Cheong, Spitale and Gunes, 2023	AFAR-BSFT	V, A	F1, DI, EAcc	✓	✓	✓*		
Yoon et al., 2022	D-Vlog	V, A	F1	✓				
Zanna et al., 2022	TILES	HE	F1, DI, EOdd		✓	✓	✓*	
Mosteiro et al., 2022	self-collected	HE	F1, SP, DI, EOpp, EOdd, EAcc	✓		✓	✓	
Rejaibi et al., 2022	DAIC-WOZ	A	F1	✓		✓		
Dang et al., 2022	LONGSCAN, FUUS, NHANES, UK Biobank	HE	F1, EOpp	✓		✓		✓*
Bailey and Plumbley, 2021	DAIC-WOZ	A	F1, SP, EOdd	✓		✓		
Y. Park et al., 2021	self-collected	HE	F1, DI, EOpp		✓	✓	✓	
Aguirre et al., 2021	CLPsych, MULTITASK	T	F1, EOpp, EOdd	✓	✓	✓		
Cong et al., 2018	RSDD	T	F1				✓*	

Table 3.3: An overview of the selected papers from the literature review regarding fairness and bias for the task of mental health analysis with the focus on depression detection. An asterisk (✓*) at a certain bias mitigation method means that the method is novel. Abbreviations. F1: F1-score. SP: Statistical Parity. DI: Disparate Impact. EOpp: Equal Opportunity. EOdd: Equal Odds. EAcc: Equal Accuracy. V: video. A: audio. T: text. HE: Health Data.

3.3.2 Bias Mitigation Approaches

A wide range of bias mitigation methods have been proposed and applied in recent years to impose fairness to ML models. Cheong, Spitale and Gunes (2023) proposed MixFeat, a simple preprocessing data augmentation strategy where new synthetic samples are created by mixing up features from existing samples with the same sensitive attribute. Their results showed that the proposed method was able to consistently improve the fairness of the model across both the uni- and multi-modal models. Additionally, the results indicated that using higher-level features (such as gaze, speech duration and sentiment of the speech) could be beneficial concerning performance and fairness when a smaller dataset is involved. Y. Park et al. (2021) conducted experiments on health data collected in a clinical setting regarding postpartum depression by analysing changes in fairness metrics when applying three different bias mitigation methods (reweighting, prejudice removal, and the removal of the sensitive label). They found that both the reweighting and prejudice removal mitigation methods were more effective in reducing the bias compared to the removal of the sensitive label mitigation method. Mosteiro et al. (2022) evaluated and mitigated gender bias in a model that predicts the future administrations of benzodiazepine based on clinical psychiatric data. The researchers applied both reweighing and prejudice remover methods to improve the fairness measures and found that reweighing was more effective than prejudice remover in reducing bias without compromising performance.

Zanna et al. (2022) used electrocardiogram (ECG) data collected in the wild and implemented a multi-task learning bias mitigation method to predict anxiety. Their results show that their method was able to reduce the bias in the base model at the expense of some accuracy. Aguirre et al. (2021) investigated the demographic fairness of depression classifiers trained on Twitter data based on self-reported diagnoses. Their analysis showed that the chosen datasets were not demographically representative and the resulting classifiers performed worse on people of colour in general. Lastly, the researchers proposed some recommendations to mitigate this bias (such as balancing or increasing the data size) and called for future new datasets to be constructed in a way

so that the population, especially underrepresented minorities, gets better represented (Aguirre et al., 2021). The study by Dang et al. (2022) also stresses the importance of prioritising adequate data collection as a means to improve fairness. In this study, the researchers used four public datasets covering different countries and populations to systematically investigate algorithmic bias in ML models. The researchers apply three pre-processing and two post-processing methods, one of the post-processing methods being a new proposed method. Their experimental results show that these bias mitigation techniques can improve fairness without sacrificing too much accuracy and that their proposed method tends to mitigate bias while preserving predictive performance. Lastly, as mentioned before, they stress the importance of the data collection process and they state that the choice for the combination of classifier type and bias mitigation algorithm depends on the use case at hand.

4 Methodology

This section starts with an overview of the methodology followed for the literature review. Hereafter, the sections dissect the proposed method and are divided into subsections per the research questions.

In section 4.2, the first research question is discussed where we specifically investigate the impact of adding the textual modality to the existing model. Lastly, in section 4.3, we delve deeper into the possible bias mitigation approaches that will be applied as well as explain how we approach the gender bias analysis.

4.1 Literature Review

The literary review for this thesis was done using a snowballing procedure (Wohlin, 2014) using the initial papers brought forth by both the first and daily supervisors as the start set. A keyword search was also used to retrieve relevant papers using Google Scholar.

For the literature review regarding past research on multimodal depression detection, the following keywords were used: "depression detection" AND "multimodal" AND "machine learning". Moreover, the review paper by (Squires et al., 2023) was used to gain insights into the domain of depression detection using AI. For this review, only papers were taken into consideration if they made use of either audio, visual, or textual features. Papers that used physiological biomarkers such as electroencephalography (EEG) or brain scans were excluded from this review.

A literature review has also been done on past works regarding bias and fairness for mental health. For this review, the following keywords were used: "depression detection" AND "bias mitigation" OR "fairness". Additionally, both the papers from Cheong, Kuzucu et al. (2023) and Cheong, Spitale and Gunes (2023) were used as initial papers for the snowballing procedure.

4.2 Comparison of Uni- and Multimodal Models

In this section, we aim to investigate the impact of incorporating the textual modality into a uni- and multimodal depression detection model, utilising the D-Vlog dataset. The research question that drives this analysis goes

> *"Does the inclusion of the textual modality contribute to the performance of the multimodal depression detection model trained on the D-Vlog dataset in terms of the F1-score?".*

To address this question, together with its subquestions that have been described in section 1.4, we will introduce the textual modality to both uni- and multimodal models after which the performance and fairness of both all the models will be evaluated. The methodology outlined in this section includes the data extraction and preprocessing steps, the feature extraction techniques for each modality, the model architectures, the selection process for the experiments, and the evaluation metrics. Additionally, Figure 10 gives the process workflow used to answer the research question above. Lastly, A visual overview of the complete feature extraction pipeline is presented in Figure 4. This figure illustrates all feature extraction models and demonstrates how each modality is utilised to extract the respective features.

4.2.1 Text-based Models for Depression Detection

Transcript Extraction

To add the textual modality, the transcripts of the vlogs have to be retrieved. As described in their paper Yoon et al. (2022), have made their dataset publicly available upon reasonable request.

However, only the de-identified anonymized data were provided consisting of the visual and audio modalities.

Upon request, Yoon et al. (2022) also provided the corresponding ID for each YouTube video from their dataset. Using these IDs and the youtube-transcript-api[1] library, the timestamped transcripts of 796 videos were retrieved. However, the main drawback of this approach is that YouTube allows people to turn off captions resulting in missing transcripts from videos that do exist. Moreover, some of these vlogs are not available anymore since YouTube allows people to take their videos offline.

Where the problem of vlogs being taken offline is difficult to overcome, the main drawback of the approach described above can be overcome by using a publicly available SOTA Automatic Speech Recognition (ASR) such as Whisper (Radford et al., 2023), to transcribe the speech using the audio signals. Furthermore, some of these ASR systems also support the prediction of word timestamps. This would mean that the original setup of Yoon et al. (2022), in which the audio and visual sequences are extracted and averaged per second, can be reproduced with the textual modality.

Whisper models are originally trained to predict approximate timestamps on speech segments rather than predicting word timestamps. For this reason, both WhisperX (Bain et al., 2023) and whisper-timestamped (Louradour, 2023) were considered since both models were specifically trained on predicting word-level timestamps. Ultimately, whisper-timestamped (Louradour, 2023) was chosen since it addresses some of the drawbacks of WhisperX, with more efficient usage of memory being one of the most important advantages for this study since the model had to run locally with limited hardware. The whisper-timestamped library uses Dynamic Time Warping (DTW) (Giorgino, 2009) under the hood to predict word timestamps using Whisper models.

Using the IDs of the vlogs and the pytube[2] library, 827 vlogs were retrieved and downloaded. Hereafter, the whisper-timestamped (Louradour, 2023) library was used to transcribe the audio in order to extract the word timestamps. A shortened and simplified example of a retrieved transcript is shown in section A.1. The statistics for the filtered dataset, meaning all vlogs that were downloaded, are discussed in more detail in section 5.1.

Textual Models

The review from Squires et al. (2023), showed that earlier work used both hand-crafted and high-level feature learning-based textual features for the task of depression detection. In this thesis, the latter will be used in the form of pre-trained word and sentence embeddings using a feature-based approach (Devlin et al., 2018). Both word and sentence embeddings were used to extract features from the textual data. The four embedding models that were used are described in more detail below.

- **word2vec*** (Mikolov, Sutskever et al., 2013) are non-contextual embeddings trained on part of the Google News dataset. The model contains 300-dimensional vectors for 3 million words and phrases.[3]
- **BioWordVec*** (Y. Zhang et al., 2019) are non-contextual embeddings that are trained using the fastText subword embedding model (Bojanowski et al., 2017) on the PubMed corpus. The model contains 200-dimensional vectors.[4]
- **all-mpnet-base-v2**[5] is a sentence embedding model which is fine-tuned on a concatenation of multiple datasets containing over 1 billion sentence pairs dataset with its intended use being a sentence and short paragraph encoder. The model maps sentences to a 768-dimensional vector space.
- **all-MiniLM-L12-v2**[6] is a sentence embedding model where a pre-trained distilled model

[1] https://pypi.org/project/youtube-transcript-api/
[2] https://pytube.io/
[3] https://code.google.com/archive/p/word2vec/
[4] https://github.com/ncbi-nlp/BioWordVec
[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[6] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

(Wang et al., 2020) is fine-tuned on a concatenated dataset containing over 1 billion sentence pairs. This model maps sentences to a 384-dimensional vector space.

Two of the pre-trained word embedding models mentioned above (marked by *) were also used in the paper by Sogancioglu et al. (2023) where they were specifically evaluated on fairness. Using these specific embedding models again will let us compare the results and conclusions of their paper with our results with regard to performance and fairness.

Textual Feature Preparation

In the original D-Vlog paper (Yoon et al., 2022), the extracted features from both modalities were segmented into seconds. Inspired by Yoon et al. (2022), we adopted the same seconds-based feature segmentation approach in our study for the word embedding models where the extracted textual features are segmented into 1-second intervals. First, the word embedding for each word in the text is retrieved from the model and linked to the start time of the word itself. Subsequently, the average value is computed for all words occurring within the same second to serve as the feature representation of that specific second. This seconds-based segmentation and averaging process allows us to leverage the temporal information from the features and to adopt the approach by Yoon et al. (2022).

However, one of the issues we encountered was that not all segments have words associated with them. Therefore, in our study, we investigated two experimental variants for handling these empty segments to address this completeness issue. Firstly, we implemented a zero-vector filling approach, where segments with no associated words were filled with zero-vectors. Secondly, we used an approach where empty segments were replaced by taking the average of the values of neighbouring segments that did contain words.

The approach with regard to the feature extraction for the sentence embedding models was more straightforward. The whisper-timestamped model (Louradour, 2023) not only predicts word timestamps but also segments the texts into sentences. The sentence embeddings are retrieved by using the pre-trained models which encode each sentence into a feature representation. Moreover, an experiment was run using spaCy's sentencizer[7] to evaluate the effectiveness of their sentence segmentation method compared to the original segmentation approach used during the transcript extraction process as described in subsection 4.2.1. The spaCy sentencizer was used to segment the whole text of a certain vlog into sentences after which the beforementioned sentence embeddings were used once more to retrieve the feature representations.

Lastly, certain data preprocessing steps were performed to prevent potential data leakage caused by explicit mentions of depression in textual data and the possible overfitting of the models on depression-related signals (Burdisso et al., 2024; Wolohan et al., 2018). We utilised a depression lexicon consisting of keywords which were derived from Cha et al. (2022) and Sogancioglu et al. (2023). Both studies developed a set of keywords indicative of depression-related sentiments and symptoms which mental health experts verified. Both sets were combined and were then used to specifically filter out segments containing explicit mentions of depression to avoid data leakage. In the seconds-based feature segmentation approach, we identified and only removed the keywords using the aforementioned synonym list before the averages of each segment were computed. In the sentence embedding approach, we screened and excluded entire sentences containing explicit mentions of depression using the synonym list. This preprocessing step is essential to mitigate any data leakage and ensure the reliability of the depression detection model when used in a real-world situation since individuals experiencing depression may not always explicitly state their condition. We trained the unimodal models both with and without this preprocessing step to gain insight into the impact of applying this preprocessing step. However, for the multimodal models, this preprocessing step will always be performed. A small sample of the depression lexicon with the

[7] https://spacy.io/api/sentencizer

keywords is given in section A.4. The full lexicon can be found in the paper by Cha et al. (2022) and in the repository[8] of the study of Sogancioglu et al. (2023).

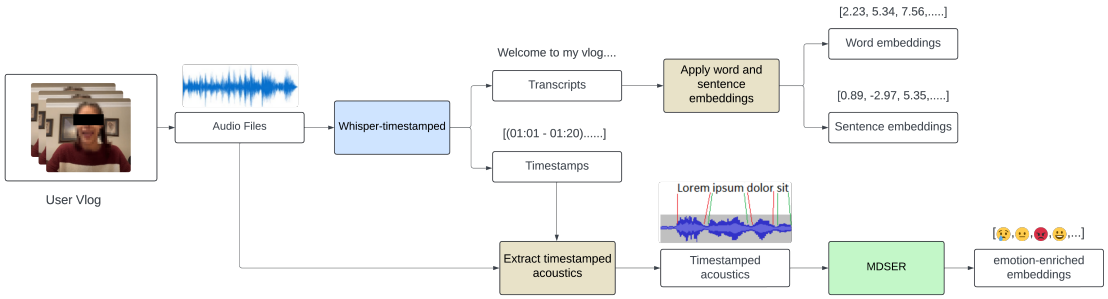

Figure 4: An overview of the feature extraction pipeline regarding the textual modality and the emotion-enriched embeddings.

4.2.2 Models using non-verbal cues

As described in subsection 3.2.3, Yoon et al. (2022) uses specific acoustic and visual features. However, section 3.1 shows studies using different feature representations which have also been shown to detect whether a person is depressed. For this reason, in addition to the feature representations detailed in Yoon et al. (2022), we also looked into some alternative feature representations which will be explained in more detail below.

Emotion-enriched Embeddings

In their paper, Wagner et al. (2023) explores the application of Transformer-based models in Speech Emotion recognition (SER) tasks, especially focusing on improving the recognition of valence (positive or negative emotional tone). They fine-tuned multiple models on the task of arousal, dominance, and valence (VAD) prediction using speech acoustics as input and released their best-performing Model for Dimensional Speech Emotion Recognition (MDSER)[9].

In our approach, we leverage both the hidden states and outputs from the pre-trained model as input features for the downstream task of depression detection. The hidden states and outputs from the pre-trained MDSER are returned on sentence level meaning that the model takes in acoustic sequences after which we retrieve the hidden states and the outputs for a particular sequence making these features sentence-based. As input sequence to the pre-trained MDSER, we use the timestamps of the sentences as predicted by the whisper-timestamped model described in subsection 4.2.1.

The input features are constructed by combining the hidden states and outputs from the pre-trained model in multiple ways. Firstly, the raw hidden states are extracted and used. Secondly, the hidden states are concatenated with the final outputs (the VAD predictions) to create a fused representation of both. Lastly, inspired by Sogancioglu et al. (2023), a normalisation technique was applied where the VAD predictions are Z-scored based on the training subset before they are concatenated with the hidden states.

4.2.3 Uni- and Multimodal Models

One of the aims of this thesis is to introduce the textual modality in order to study the effect of the textual features on the performance and the gender bias of both the uni- and multimodal models. In subsection 4.2.1 we already discussed that the textual modality gets incorporated through

[8] https://github.com/gizemsogancioglu/gender-bias-mental-health/blob/main/data/depression_synonyms.json
[9] https://github.com/audeering/w2v2-how-to

seconds-based word embeddings and sentence-based embeddings. This section will explain how these textual embeddings are fused with the other features in the uni- and multimodal models and explain how the modalities are chosen for the experiments.

Uni- and Bimodal Models

In their approach, Yoon et al. (2022) train both unimodal models and bimodal models. This means that for these models, the architectures that were proposed and used by Yoon et al. (2022) can and will be used for our experiments.

For the unimodal models, this means that the generated output of the unimodal transformer encoder will be used directly as the representation with which the depression detection layer will perform the classification operation.

In the case of the bimodal models, we again utilise the original approach detailed by Yoon et al. (2022), which uses cross-attention to generate a multimodal representation that is used by the depression detection layer for classification. A visual overview of the architecture is shown in Figure 2. In their paper, Yoon et al. (2022) mention that they compared their proposed fusion method against three commonly used multimodal fusion operations (Add, Concat, and Multiply). Since we are adding the textual modality, which is novel, we will again consider and implement the Concat (Kim et al., 2013) fusion approach as this fusion method has been referenced in previous work (Alhanai et al., 2018; Marriwala & Chaudhary, 2023; J. Park & Moon, 2022).

Synchronisation of Multimodal Features for Fusion

Sentence-based features typically span multiple seconds, whereas both the original audio and video embeddings use fixed seconds-based representations. These modalities need to be synchronised to combine these different feature modalities such that the cross-attention mechanism can be used.

To synchronise these modalities, we extracted the start and end timestamps for each sentence. These timestamps allowed us to calculate the averages of the seconds-based features for each sentence. Specifically, we aligned the timestamps of the sentences with the corresponding seconds-based features and computed the mean values of the features within those segments. This approach ensured that all modalities were represented on the same timescale and that the fusion between these modalities could be applied effectively. A visual example of this synchronisation process is shown at Figure 5.

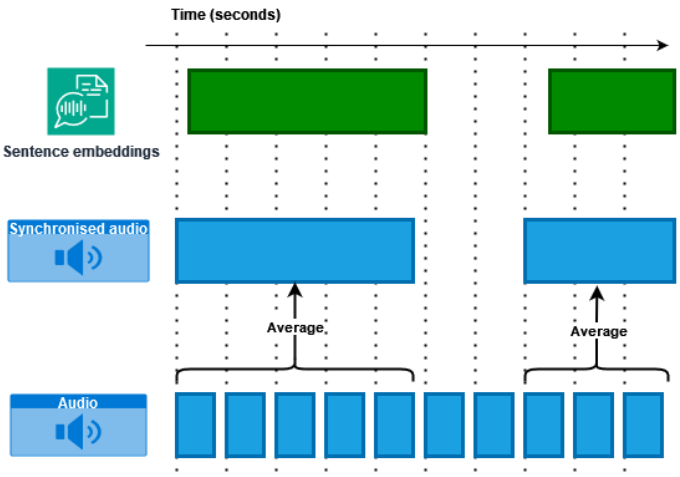

Figure 5: Illustrative example of the synchronisation process where the audio gets aligned with the sentence embeddings.

Trimodal Models

For the trimodal models, which will incorporate the textual, visual, and audio modalities, we will continue to experiment using the Concat fusion operation due to its ease of implementation and shown effectiveness in fusing the feature information of three modalities (Rodrigues Makiuchi et al., 2019).

However, the depression detection approach using cross-attention by Yoon et al. (2022) is specifically designed to be used for a model fusing two modalities. It therefore will not be directly feasible for a trimodal approach. Even though the feature fusion approaches put forth by Fang et al. (2023) and Shen et al. (2022), where they successfully train a separate modal attention layer which can effectively fuse the features from the input modalities, can be used for a trimodal model. We chose to stay close to the cross-attention module put forth by Hasan et al. (2021) and use their proposed module as inspiration. To fuse the latent information from the three input modalities, we put forth and experiment with two new module architectures which alter and re-use the original cross-attention module in such a way that it is able to handle three modalities. Both module architectures are discussed more elaborately below.

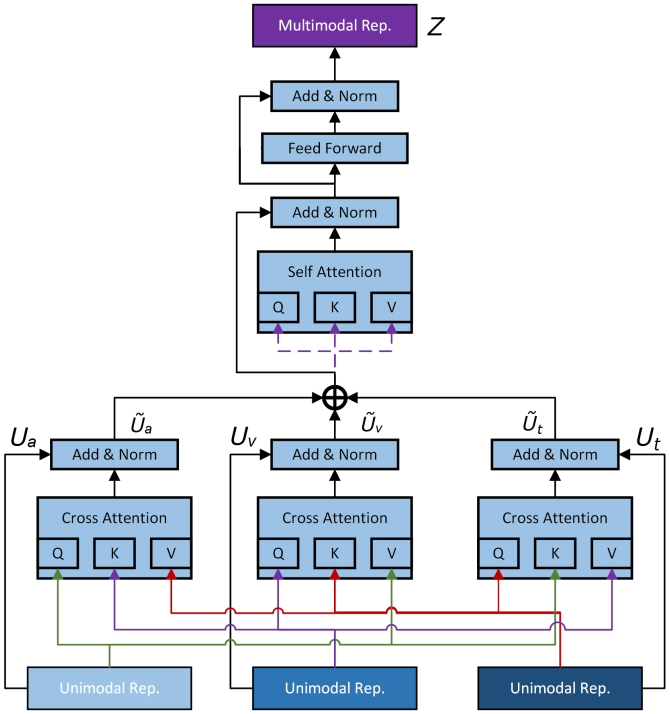

Figure 6: An illustration of the crisscross-attention module architecture.

Crisscross-attention Module The first architecture is called the crisscross-attention module and extends the original cross-attention module by adding an additional cross-attention block to the architecture. As shown in Figure 6, the multimodal transformer encoder takes as inputs the unimodal acoustic, visual, and textual representations (U_a, U_v, and U_t) and distributes them among the three cross-attention modules where each cross-attention module gets its Query, Key, and Value from a distinct modality. Moreover, following the approach by Yoon et al. (2022), the residual connection uses the representation from the modality that delivers the Query to the cross-attention module. After the residual connections and layer normalisation, we obtain three cross representations ($\widetilde{U}_a$, $\widetilde{U}_v$, and $\widetilde{U}_t$). To fuse the cross-modal information, we follow the approach

from Yoon et al. (2022), where the cross representations are concatenated and a straightforward Transformer encoder is used.

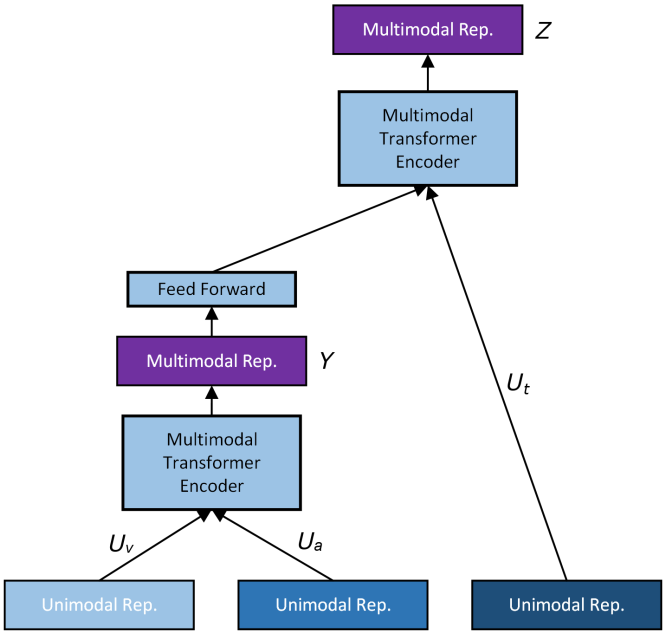

Figure 7: An illustration of the layered cross-attention module architecture.

Layered Cross-attention Module The second architecture is called the layered cross-attention module, where we fuse the three modalities by applying the original Multimodal Transformer Encoder proposed by Yoon et al. (2022), which includes the original cross-attention module, twice in succession. As shown in Figure 7, the multimodal transformer encoder first takes the unimodal acoustic and visual representations (U_a, U_v) as inputs and combines them using the original Multimodal Transformer Encoder to get the first multimodal representation Y. Hereafter, we use a single feed forward layer to map the dimensionality of Y back to the original size so it can be used during the next step. The last multimodal transformer encoder takes the multimodal representation Y and the textual representation (U_t) as inputs and fuses them as described above to generate the final multimodal representation Z.

Feature Selection and Model Evaluation Strategy

In order to properly assess the aforementioned research question, we used a systematic approach to select and evaluate features across the audio, visual, and text modalities for each of the uni- and multimodal models. This systematic approach consists of the following steps.

Data and Feature Extraction To prepare the data for model training, preprocessing and feature extraction was performed for each feature and their respective experimental variants for each modality. A quick overview of each modality's features are given below.

- **Audio Modality:** acoustic features as described in subsection 3.2.3, MDSER features.
- **Visual Modality:** visual features as described in subsection 3.2.3.
- **Text Modality:** seconds-based embeddings (word2vec, BioWordVec), sentence-based embeddings (mpnet, miniLM).

Unimodal Model Training and Evaluation For the unimodal models, each modality's feature representations and their experimental variants were used to train the models. Afterwards, the performances of these models were assessed based on the weighted F1-scores.

Selection of Optimal Features The feature configuration that achieved the highest weighted F1-score in the unimodal setup described above was selected for each modality. This step ensured that only the most effective features were considered for the multimodal models. However, we only consider the unimodal models that used features where the keywords were removed during this step. This ensures the reliability of the models trained hereafter, since it focuses on data from subjects that do not explicitly disclose depression, thus better reflecting real-world scenarios.

Multimodal Model Construction and Training The selected features from each modality, derived from the previous step, were then combined to form the input for the multimodal models. Through this integration, we aim for the models to leverage the complementary information the different modalities provide. Lastly, during this step, the various fusion techniques described in subsection 4.2.3 were explored.

4.2.4 Evaluation of the Classifiers

To evaluate and assess the impact of adding different modalities and fusion approaches, all uni- and multimodal classifiers will be evaluated using both performance and fairness metrics.

Evaluation of Classifier Performance

To measure the overall performance of the trained classifiers, several performance measures can be calculated. For this thesis, we will use the weighted average precision, recall and F1-score as performance measures since this enables us to compare the results with the original D-Vlog paper (Yoon et al., 2022). Moreover, during the model evaluation, weighted F1-scores are also specifically calculated for the different gender categories to provide a more detailed understanding of the effectiveness and fairness of the model when handling gender-specific data.

Evaluation of Classifier Fairness

To measure the fairness of the trained classifiers, we will use a combination of the approach from Cheong, Kuzucu et al. (2023) in which they use statistical parity, equal opportunity, equalised odds, and equal accuracy to evaluate their results and the study from Sogancioglu et al. (2024) in which equal opportunity, predictive equality, and equal accuracy are used to evaluate the fairness of their predictive models.

Both the statistical parity and equal odds measures will not be used. The statistical parity measure will be excluded since this measure is not relevant in the case of mental healthcare. Alternatively, the equalised odds measure has been excluded since we already include the predictive equality measure which, along with equal opportunity, makes up this measure. Ultimately, in this study, we will use equal opportunity (EqOpp), equal accuracy (EAcc), and predictive equality (PredEq) to evaluate the fairness and bias of the trained classifiers. The definitions of these measures are provided in section 2.2.

To quantify each measure of fairness, the formulas are transformed to ratios as described in the paper by Cheong, Kuzucu et al. (2023). For example, the formula for equal opportunity (EqOpp), as shown in Equation 4, gets altered to

$$EqOpp = \frac{Pr[\hat{Y} = 1|Y = 1, S = s_0]}{Pr[\hat{Y} = 1|Y = 1, S = s_1]}, \tag{14}$$

where S represents the demographic groups and s_0 and s_1 correspond to the unprivileged and privileged demographic groups respectively. Furthermore, since Yoon et al. (2022) uses a weighted F1-score as their performance measure, we formulate the equal accuracy (EAcc) accordingly.

$$EAcc = \frac{F1_{s_0}}{F1_{s_1}} \tag{15}$$

A score of 1 for a measure is regarded as the ideal score which entails that both measures are equal for both groups and the predictor can be deemed fair. Lastly, we adopt the approach of disparate impact proposed by Zanna et al. (2022), where they assume the acceptable lower and higher bound for the score to be 0.8 and 1.2 respectively.

Evaluation of Modality and Fusion Impact

To determine the impact of adding the textual modality and identify the best fusion approach, we compare the best-performing textual unimodal model with the best bi- and trimodal models regarding both their performance and fairness on the test set. The significance of these differences will be assessed by conducting two-tailed paired t-tests, using a significance level of $p < 0.05$ which is in line with the study from Sogancioglu et al. (2024).

4.3 Bias Mitigation and Fairness

In this section, we explore both the possible sources of the gender bias and explore strategies to address and reduce said bias in our depression detection models. This exploration is driven by the research question

> *"To what extent can the application of bias mitigation techniques help reduce existing gender bias in the best model that has been obtained during subquestion **1B**"?*

To address this question, together with the subsequent subquestions described in section 1.4, we will begin by evaluating the gender bias present in the best-performing model identified during the process of answering the first research question. Subsequently, we apply a series of bias mitigation techniques aimed at reducing this bias. The methodology outlined in this section will explain how the potential gender bias sources are analysed. Moreover, it will explain what kind of bias mitigation methods are to be applied and why. Additionally, Figure 11 gives the process workflow used to answer the research question above.

4.3.1 Analysing Potential Sources of Gender Bias

In their paper (Bailey & Plumbley, 2021) study gender bias by comparing the differences in performance after training the models with and without their bias mitigation approach. Cheong, Kuzucu et al. (2023) have a somewhat similar approach where they compare the differences in performance and fairness metrics in a gender-based approach. Lastly, Min et al. (2023) and Yoon et al. (2022) train and validate the model with a specific gender, where Yoon et al. (2022) additionally train the model with both genders and test with a specific gender. Based on the papers described before, our approach to analysing the potential sources of gender bias will be a combination of modality-based analysis and gender-based analysis.

For the modality-based analysis, multiple uni- and multimodal models will be trained using samples from both genders. These models will then be evaluated with samples specific to each gender. This approach will provide insights into whether certain features used by the model inherently favour one gender over another. By applying this approach to both uni- and multimodal models, we also gain insight into the impact of each modality on gender bias and whether certain fusion methods are more suitable for mitigating said bias. Lastly, both Cheong, Kuzucu et al. (2023) and Yoon et al. (2022) conduct a modality-based analysis exclusively on bimodal models. We will extend this analysis by also incorporating the performance and fairness metrics of the unimodal models and contribute to these studies by including the textual modality.

For the gender-based, we will train some of the models using a specific gender after which we evaluate the same model using the same gender. This will give us further insights into whether

certain features perform better for certain genders. This is especially informative since past studies surrounding gender-based classification for depression have reported improved performance when using a gender-dependent approach (Pampouchidou et al., 2016; Samareh et al., 2018; Stratou et al., 2013).

By analysing the potential sources of gender bias using this methodology, we aim to gain a better understanding of the factors contributing to the gender bias that emerged in the models while answering the initial research question.

4.3.2 Bias Mitigation Methods

In this thesis, we will not only analyse the potential sources of the bias but also try to mitigate this bias through the application of bias mitigation methods. As described in section 2.4, there are three types of bias mitigation methods (pre, in, and post). We will take pre-, in-, and post-processing methods into account while also taking into consideration the study by Sogancioglu et al. (2024), in which they found that clinicians did not find it to be acceptable to sacrifice the overall prediction performance to equalise the performance measures across gender groups which is behaviour that can be ascribed to post-processing methods.

In their paper, (Cheong, Kuzucu et al., 2023) already applied three different bias mitigation techniques (Data augmentation, Loss Re-weighting, and ROC) and found that none of these methods were consistently effective at mitigating the bias. However, since we use a model trained on the textual modality during the bias mitigation process, these models can again be considered.

Moreover, Cheong, Kuzucu et al. (2023) not only applied bias mitigation methods, but also identified two potential factors for the appearance of gender bias one of which was the way how the data pre-processing has been handled. Based on this assumption, we will also experiment using different durations as an approach towards bias mitigation. This experiment will be done on the cross-attention model trained using the original features as described in subsection 3.2.3, as well as the best-performing textual unimodal model.

Ultimately, MixFeat and oversampling were selected as the pre-processing methods. loss re-weighing was selected as the in-processing method, and equal odds was selected as the post-processing method as these were considered to be the most promising methods. These methods are all explained in section 2.4. While the application of the oversampling, loss reweighing and equal odds methods are used straightforwardly, for the MixFeat method five different variants are used which is explained in detail below.

MixFeat Bias Mitigation Variants

For the MixFeat (Cheong, Spitale & Gunes, 2023) bias mitigation method, we used five different variants, two of which rely solely on synthetic samples for performing the depression detection task. These variants are described below.

- **Group upsampling:** Here we upsample the minority sensitive group and use them in combination with the original instances.
- **Subgroup upsampling:** Here we upsample the minority (sensitive) group-minority label subgroup and use them in combination with the original instances.
- **Mixgender upsampling:** Here we upsample the minority group using a mixgendered approach which means that each new sample consists of a combination of a male and female sample and use them in combination with the original instances.
- **Synthetic sampling:** Here we sample from the 'depression' class, after which we only use these synthetic samples during training.
- **Synthetic mixgendered sampling:** Here we sample from the 'depression' class using the mixgendered approach, after which we only use these synthetic samples during training.

The synthetic experiments can be considered important as they evaluate the potential of using MixFeat to generate synthetic data for federated learning applications. If models trained on

synthetic data perform well, this approach could enhance the privacy of individuals with depression because these synthetic features are much more difficult to trace back to a specific individual. As a result, individual privacy can be safeguarded which could facilitate easier data sharing (Rieke et al., 2020).

Application of Bias Mitigation Methods

The bias mitigation methods described above were applied both at sentence-level and session-level, utilising different model architectures for each.

Sentence-level For the sentence-level bias mitigation approach, we used the straightforward D-Vlog unimodal model architecture described in subsection 4.2.3 to process all sentences from a subject and to perform the depression detection. This approach does mean that for the MixFeat methods, sentences from two samples are combined which may not hold any temporal connection.

Session-level For the session-level approach, we, inspired by Min et al. (2023), employed a support vector machine (SVM) where the keyword filtered sentence-based mpnet representations within a single session were averaged into a single representation. The pre-processing bias mitigation methods were then applied to these single representations. This approach provides both insights into the effectiveness of the textual modality on a simpler model as well as into how well the pre-processing methods worked on a simpler model. The session-level analysis was particularly important for understanding the impact of the variants of the MixFeat method on the sentence-level approach since MixFeat combines sentences from two samples without temporal connection using this approach, which could potentially affect the model performance differently.

5 Experimental Setup

This chapter begins with section 5.1, which explains the D-Vlog dataset and the new subset, including the rationale behind the explicit need for a new subset. Additionally, this section provides data statistics and an exploratory language analysis. In subsection 5.2.2, all implementation details regarding the execution of the experiments are described. Lastly, section 5.3 covers the bias mitigation methods employed in both the session- and sentence-level experiments

5.1 The D-Vlog Dataset

As described in subsection 4.2.1, we were only able to retrieve 827 of the transcripts for all the videos of the original D-Vlog dataset. Specifically, 96, 11, and 27 videos were missing from the training, validation, and test subsets respectively. This issue was similarly encountered by Gimeno-Gómez et al. (2024) where they managed to retrieve only 861 vlogs and subsequently used the original dataset split, excluding the unavailable videos. However, in our study, we aim to model the original imbalance present in the dataset rather than assume that the deleted videos do not affect the integrity of this imbalance. For this reason, a separate dataset (v2) was made consisting of all the vlogs for which the videos were available. The differences between the original D-Vlog dataset (v1) and the filtered dataset (v2) are discussed in more detail below.

5.1.1 Dataset Split

In their approach, Yoon et al. (2022) applied a standard train, validation, and test split of 7:1:2 but did not explicitly state whether any stratification was applied. However, Yoon et al. (2022) did specify for their collected dataset which sample belonged to which subset. This means that the experiments regarding the replication study could be done using the original dataset (v1) with the original folds.

For our experiments, a separate dataset split (v2) was made for the dataset consisting of the samples for which their videos could be downloaded. For this dataset, we adhered to the same dataset split ratio and ensured that no folds share a YouTube channel but did stratify the splits based on gender and the depression label. This approach ensured some alignment with the original paper while still allowing us to evaluate the model's performance under comparable conditions. The dataset splits described above, together with the size of each fold, are shown in Table 5.1.

Dataset	Gender	Train	Val	Test
v1 (Yoon et al., 2022)	Male	216	40	66
	Female	431	62	146
	Both	647	102	212
v2 (Ours)	Male	199	29	57
	Female	379	55	108
	Both	578	84	165

Table 5.1: The number of samples in the train, validation, and test folds in both datasets used in our approach.

5.1.2 Data Statistics

The descriptive statistics of both the original and the new dataset, grouped by label and gender, are presented in Table 5.2. The original dataset consists of 555 depression and 406 non-depression vlogs whereas the filtered dataset consists of 449 depression and 378 non-depression vlogs. Moreover,

the table shows that the status quo of the original D-Vlog dataset has been maintained in the new dataset since the new dataset also has twice as many females than males in the depression vlogs. This behaviour is also copied for the non-depression vlogs.

Dataset	Label	Gender	# samples	Avg. Duration
v1 (Yoon et al., 2022)	Depression	Male	182	583.74s
		Female	373	667.63s
	Non-depression	Male	140	438.77s
		Female	266	587.76s
v2 (ours)	Depression	Male	147	569.49s
		Female	302	650, 87s
	Non-depression	Male	138	433.17s
		Female	240	592.76s

Table 5.2: Descriptive statistics of the filtered D-Vlog dataset.

Lastly, Figure 8 illustrates the difference in vlog durations between the original and the filtered (v2) datasets regarding the depressed and non-depressed labels. The plot shows that the difference in distributions between the two datasets is barely discernible.

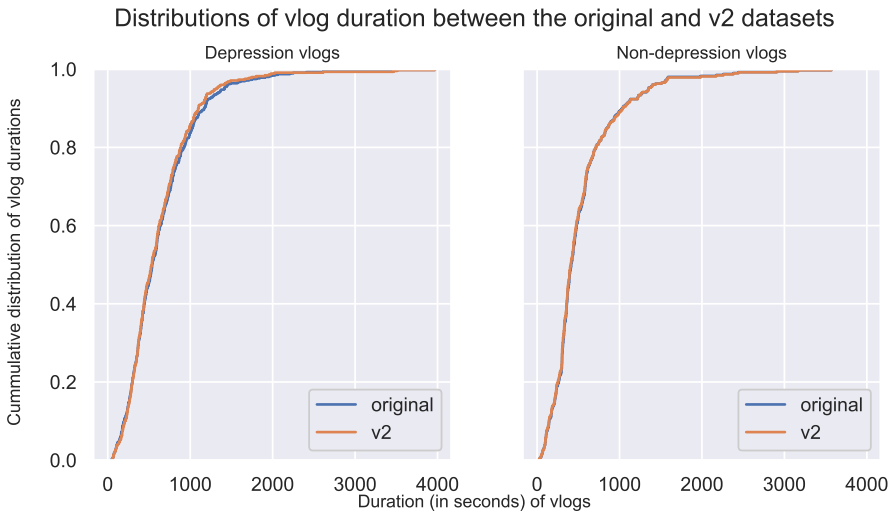

Figure 8: The difference in distributions of vlog duration between the original and v2 D-Vlog dataset.

5.1.3 Exploratory Language Analysis

One of the aims of this thesis is to introduce the textual modality using the transcripts of the vlogs. As a first step, an exploratory analysis of the extracted transcripts was conducted. This preliminary analysis was motivated by the question of whether distinctive linguistic features are prevalent among different label-gender combinations. By computing the most frequently occurring words and n-grams within each subgroup, we can identify potential differences in their utterances.

During preprocessing, all newlines, trailing whitespaces, punctuation, and numbers were removed. The transcripts were converted to lowercase, stopwords were eliminated, and the words were lemmatized.

The initial exploration revealed no interesting or noticeable differences between subgroups when words and n-grams were computed straightforwardly. However, it did reveal one video where after approximately 11 minutes the person switches to a different language resulting in the ASR model

to predict 'jobless' for almost each speech utterance after this moment. As a result, the extracted text for this video was excluded from the analysis.

Consequently, we refined the analysis by filtering out words and n-grams used by the other subgroups before computing the counts for each subgroup. This ensured that the unique utterances from each subgroup became visible. The resulting analysis, presented in Table 5.3, shows the differences between the depressed and non-depressed subgroups for unigrams and bigrams. The table indicates that the depressive subgroup tends to use more words related to symptoms or consequences of depression. This exploratory method was also applied to the label-gender combinations, with the results presented in section A.3.

Depression			**Normal**		
n-gram	word(s)	count	n-gram	word(s)	count
1	psychiatrist	82	1	clemson	118
	antidepressant	77		stereotype	78
	selfesteem	42		tammy	58
	worthless	33		component	57
	counseling	33		phd	57
	ptsd	33		concealer	44
	counseler	28		spanish	43
	biscuit	33		ethnicity	40
2	im depressed	55	2	student dialogue	162
	depression know	42		year third	159
	know depression	36		new student	140
	depressive episode	32		active voice	77
	dealing depression	30		true dialogue	72
	suffering depression	28		think dialogue	47

Table 5.3: The distinctive linguistic features prevalent in both the depressed and the non-depressed subgroup.

5.2 Implementation Details

To study the impact of adding the textual modality to the existing D-vlog architecture, we utilised the *Depression Detection* architecture proposed by Yoon et al. (2022). As described in subsection 3.2.3, this architecture consists of two unimodal Transformer encoders after which a multimodal Transformer encoder is used to incorporate the learned representations. While the proposed architecture proposed by Yoon et al. (2022) is compelling, no public architecture implementation was available at the time of this thesis. For this reason, the architecture proposed by Yoon et al. (2022) was re-implemented based on the descriptions and illustrations provided by Yoon et al. (2022) and Cheong, Kuzucu et al. (2023).

To improve the reproducibility of our implementation, we used the confection library[1], which offers a configuration system capable of parsing config files. This means that each experiment is saved as a separate config file, improving the transparency and reproducibility of each experiment.

5.2.1 Replication Study

Contrary to the implementation by Yoon et al. (2022), we used PyTorch (Paszke et al., 2019) to implement the model primarily due to our past experiences with its architecture and tooling. As mentioned, the architecture was re-implemented where sensible choices were made for certain components for which no specific information was provided in the reference literature. One such choice was the addition of an additional embedding layer between the convolution layers and the

[1] https://github.com/explosion/confection

positional encoder to map both unimodal representations (defined as d_u by Yoon et al. (2022)) to the same dimensionality as proposed by Zerveas et al. (2021).

All models were trained using a desktop containing an Intel Core I9-9900K processor and GTX1660 OC 6G graphics card. Moreover, for the replication study, the models were trained following the experimental settings as described by Yoon et al. (2022). This means we used the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.002. Furthermore, the batch size, epochs, sequence length (t), dropout and the unimodal representation dimensionality (d_u) are set to 32, 50, 596, 0.1, and 256. Lastly, For each of the experiments, the models are trained and evaluated using different seeds (0, 1, 42, 1123, 3407) following the studies by Cheong, Kuzucu et al. (2023) and Picard (2021). This means that the reported results are the averaged results. For the replication study, we specifically experimented with the number of heads in the cross-attention module (8 and 16) as these are not mentioned in the paper. Moreover, both the Concat method and cross-attention module are used as fusion methods as Yoon et al. (2022) also communicate the performances for these methods in their paper.

5.2.2 Uni- and Multimodal Experiments

The setups of the models for the uni- and multimodal experiments largely follow the same setup as described in subsection 5.2.1. This means that all the models use the same hyperparameters whenever applicable and all the models are trained and evaluated with the same 5 seeds mentioned before.

During the replication study shown in Table 6.1, we discovered that using 16 heads in the cross-attention module yielded better performance on average compared to using 8 heads on the v2 dataset. Therefore, for the subsequent experiments involving multimodal models, each cross-attention module was configured with 16 heads. However, the unimodal transformer encoder is implemented with 8 heads, aligning with the original implementation by Vaswani et al. (2017).

As discussed earlier in subsection 4.2.1, we employed both seconds-based and sentence-based approaches for incorporating textual features. For the experiments which involved seconds-based features, we followed the approach by Yoon et al. (2022). We used the mean duration of the vlogs, which is 596 seconds, as the sequence length. In line with this, we used the mean number of sentences per vlog, which is 104, for the sentence-based experiments.

Lastly, when sentence-based features or the MDSER features were used as one of the modalities in the experiments, synchronisation was performed on the other features that were involved. Consequently, no experiments were performed using the combination of seconds-based word embeddings and MDSER features. This is because synchronising the word embeddings effectively turns them into sentence representations, and the unimodal experiments in section 6.2 demonstrate that these are inferior in terms of downstream performance with respect to sentence embeddings.

5.3 Bias Mitigation Setup

For both the session-level and sentence-level experiments, the specific features used for running the experiments were the best performing sentence-based keyword-filtered features from the unimodal experiments shown in Table 6.2 which were the mpnet features.

5.3.1 Session-level Experiments

For the session-level experiments, the SVM model from scikit-learn[2] was employed where all sentences within a session were averaged into a single representation as described in subsection 4.3.2. For the model optimisation, hyperparameters were tuned using a grid search approach. The hyperparameters considered were C (regularisation parameter), Gamma (kernel coefficient), and Kernel (specifies the kernel type) using the parameter grid from Table 5.4. For all hyperparameters not specified in this table, the default values were used, including a polynomial kernel function with a

[2] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

degree of 3. The best estimator, with its corresponding hyperparameters, was selected based on the performance of the validation subset.

Hyperparameter	**Values**
C	0.1, 1, **10**, 100
Gamma	**1**, 0.1, 0.01, 0.001
Kernel	linear, rbf, **poly**, sigmoid

Table 5.4: The parameter grid for the session-level experiments. The 'best' parameters found are highlighted in bold.

With these 'best' parameters identified, all bias mitigation approaches were applied to the training dataset. All models were trained using their respective augmented datasets on the five different seeds used for running all the experiments after which the the averages of the performance and fairness were computed and collected.

5.3.2 Sentence-level Experiments

For the sentence-level experiments, the unimodal model described in subsection 4.2.3 was used for all the experiments. There were three different bias mitigation variations for the sentence-level experiments.

Increasing Sequence Length For the original v1 bimodal model, we experimented with sequence lengths (n) of 596, 750, and 900. For the sentence-based unimodal model, which typically has a maximum sequence length of 104, we additionally trained the model with sequence lengths (n) of 140 and 180.

Gender-specific Training For this variation, the unimodal textual model was trained using samples from only male or female subjects. Afterwards, the model was evaluated using the same gender on which it was trained.

Bias Mitigation Methods The final variation involved applying the bias mitigation methods described in subsection 4.3.2 to the training data . The implementation of the oversampling and reweighing methods was straightforward.

For the Equalized Odds method, we used the EqOddsPostprocessing[3] method from the aif360 library (Bellamy et al., 2018). In this approach, the post-processing model is trained using the training set. Once trained, the model is applied to the test set during the evaluation process.

For the MixFeat method, we used an alpha value of 1.0 when drawing samples from the beta distribution. This choice was based on the default value used in the mixup implementation[4], as the MixFeat paper (Cheong, Spitale & Gunes, 2023) did not specify an alpha value.

[3] https://aif360.readthedocs.io/en/stable/modules/generated/aif360.algorithms.postprocessing.EqOddsPostprocessing.html

[4] https://github.com/facebookresearch/mixup-cifar10

6 Results

In this section, we present a selection of the results from the extensive number of experiments conducted during this study. Given the sizeable number of experiments performed, we have chosen to present a selection of results to highlight specific experiments that most effectively illustrate the core findings and support answering the research questions. A complete overview of the training performance of all uni- and multimodal models, including performance metrics on the validation and test subsets, is provided in the appendix in section A.5.

Throughout this section, we will implicitly refer back to the research (sub)questions outlined in section 1.4. This approach helps maintain a clear connection between our experimental results and the primary objective of this study.

6.1 Replication Study

As mentioned in subsection 5.2.1, no public architecture implementation of the D-Vlog architecture was available at the time of this thesis. For this reason, we developed our own implementation from scratch following the descriptions and illustrations from Yoon et al. (2022) and Cheong, Kuzucu et al. (2023). To ensure the validity and accuracy of our implementation, we conducted a replication study of which the results are shown in Table 6.1.

Based on the initial results of the averages on the v1 dataset, we have chosen to include the standard deviation for the overall F1-scores to provide a clearer understanding of the performance consistency of the models as Table 6.1 shows the re-implemented v1 bimodal models using cross-attention having lower performance than reported in the original paper. However, it was observed that for certain seeds, the model's performance on the v1 dataset matched the F1-scores reported in the original paper. For instance, the weighted F1-score for the bimodal model using 8 heads in the cross-attention (cross8) with seed 0 is approximately 0.65, demonstrating that our implementation can achieve comparable results under specific conditions.

Moreover, for the v1 dataset, Table 6.1 shows that the unimodal models still exhibit a performance difference between genders, with a better performance towards males, although this disparity is less pronounced than the disparity reported in the original paper. This gender performance gap persists across all bimodal variants, albeit with a smaller difference. This is further evidenced by the EqOpp metric falling within the acceptable range for two of the three bimodal models.

For the v2 dataset, the results show that the performance differences between the genders become even smaller and, in some cases, reverse, with females now having better performance. This trend continues with the bimodal models. However, all bimodal models now fall outside the acceptable range regarding the PredEq metric which indicates that the female class has a higher FPR rate than the male class. Lastly, the F1-score for the bimodal model using 16 heads in the cross-attention (cross16) outperforms the model using 8 heads which means that all subsequent experiments involving the cross-attention module are configured with 16 heads, as mentioned in subsection 5.2.2.

Dataset	Modality	Fusion	F1-score Overall	std	F1-score Male	Female	Fairness EAcc	EqOpp	PredEq
v1 (Yoon et al., 2022)	audio	-	0.586	-	-	-	-	-	-
	visual	-	0.564	-	-	-	-	-	-
	both	concat	0.611	-	-	-	-	-	-
		cross	0.635	-	0.754	0.545	-	-	-
v1	audio	-	0.602	0.058	0.674	0.568	1.19	**1.32**	1.06
	visual	-	0.608	0.040	0.684	0.572	1.19	**1.25**	0.82
	both	concat	0.613	0.037	0.650	0.594	1.09	**1.24**	**1.22**
		cross8	0.598	0.074	0.657	0.572	1.15	1.15	0.80
		cross16	0.534	0.077	0.572	0.516	1.11	1.14	1.12
v2	audio	-	0.581	0.164	0.567	0.584	1.00	0.99	0.81
	visual	-	0.608	0.033	0.599	0.612	0.98	0.95	0.90
	both	concat	0.613	0.023	0.627	0.604	1.05	0.96	**0.77**
		cross8	0.536	0.130	0.527	0.536	1.00	0.86	**0.69**
		cross16	0.544	0.107	0.545	0.543	1.02	0.95	**0.75**

Table 6.1: A comparison of the performance results on the v1 and v2 datasets from the replication study with the original results as communicated by Yoon et al. (2022). Values where the standard deviation exceeds 10% of the overall F1-score are underlined. Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Abbreviations: equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq).

6.2 Unimodal Results

To investigate the impact of incorporating the textual modalities and to specifically gain insights into which word or sentence embeddings perform best for unimodal depression detection and are suitable for use in the multimodal approach, we trained multiple models using both word embeddings and sentence embeddings. A selection of these results is shown in Table 6.2 where both approaches are combined.

For the seconds-based approach, the results in Table 6.2 indicate that the zero-vector filling approach is more effective for the word2vec features that still included depression-related keywords, whereas the average-filling approach proves more effective once these keywords were removed. Additionally, the performance of the BioWordVec-based features was much lower than that of the word2vec-based features. This outcome is expected, as the BioWordVec embeddings are trained on a medical corpus (Y. Zhang et al., 2019), whereas the subjects typically do not use medical terms in their vlogs.

For the sentence-based approach, the results indicate that the mpnet sentence embeddings outperform the minilm sentence embeddings and that the mpnet sentence embeddings perform as well as the word2vec word embeddings on the downstream task. Moreover, the segmentation method from spaCy shows a slight advantage over the original approach.

Overall, Table 6.2 shows that, in almost every instance, filtering out depression-related keywords increases the gender performance gap, resulting in better performance for females. This suggests that these models, at least to some extent, focus on these explicit mentions of depression for the male group. Moreover, the removal of explicit mentions of depression also causes the PredEq metric to shift into the unacceptable range regarding fairness.

As part of the experiments, we also explored the use of emotion-enriched embeddings from acoustic signals. Using outputs from MDSER, three different setups were trained and evaluated. Table 6.3 shows that all three setups perform on par with each other, both on performance and fairness, but that the setup using a concatenation of the hidden layers and the VAD predictions as features performed slightly better than the other setups. More importantly, these experiments also demonstrated that these emotion-enriched features are more informative than the acoustic signals used

			F1-score			Fairness		
Feature	**type**	**with keyw**	**Both**	**Male**	**Female**	**EAcc**	**EqOpp**	**PredEq**
w2v	avg	yes	0.912	0.930	0.902	1.03	0.97	**0.43**
w2v	zero	yes	0.945	0.947	0.944	1.00	0.98	0.93
w2v*	avg	no	0.929	0.909	0.939	0.97	0.93	1.05
w2v	zero	no	0.913	0.863	0.939	0.92	0.92	**1.90**
w2v	sent	no	0.878	0.831	0.903	0.92	0.91	**1.26**
biow	avg	yes	0.866	0.842	0.878	0.96	0.91	0.94
biow	avg	no	0.859	0.807	0.886	0.91	0.89	**1.24**
mpnet	sent	yes	0.950	0.936	0.957	0.98	0.95	1.01
mpnet*	sent	no	0.928	0.888	0.950	0.93	0.90	**1.33**
mpnet	spacy	no	0.931	0.891	0.952	0.94	0.97	**1.87**
minilm	sent	no	0.892	0.848	0.915	0.93	0.94	**1.42**

Table 6.2: The performance and fairness results of the textual unimodal model experiments. For the w2v and biow features 'type' refers to the approach used when handling empty segments. The best-performing unimodal textual model that does not use the keywords is marked with an asterisk (*). Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Abbreviations: equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq).

by Yoon et al. (2022). Lastly, regarding gender-specific performance, one of the models trained on the emotion-enriched embeddings had a higher performance for the male group than for the female group. This result, together with the results from the unimodal models in the replication study and the textual modality experiments, highlights that the choice in feature embedding can have varying impacts on the performance of a certain group.

	F1-score			Fairness		
Features	**Both**	**Male**	**Female**	**EAcc**	**EqOpp**	**PredEq**
audio	0.581	0.567	0.584	1.00	0.99	0.81
hidden	0.634	0.611	0.645	0.95	1.00	1.13
hidden + VAD*	0.636	0.624	0.642	0.98	1.04	1.16
hidden + z-score VAD	0.624	0.627	0.621	1.01	1.11	1.11

Table 6.3: Results from the experiments using features extracted using MDSER. The unimodal audio-based model trained on the v2 dataset from Table 6.1 has also been included in this table. Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Abbreviations: equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq).

6.3 Multimodal Results

To investigate which fusion approach provides the best performance and to gain insight into the impact of these fusion approaches on the fairness of the models, multiple experiments were conducted. Models were trained using the best-performing unimodal features discussed in section 6.2 and the original features, described in subsection 3.2.3, with various fusion approaches. First, the results of the bimodal models will be reviewed after which the trimodal results will be discussed.

6.3.1 Bimodal Experiments

For the bimodal experiments, the models were run for both the seconds-based and sentence-based approaches using the Concat and cross-attention fusion methods. In Table 6.4, the results for all bimodal feature-fusion combinations is shown.

For the seconds-based approach, the table shows that the cross-attention fusion method yielded better performance when visual features were used as the second modality than the Concat fusion method. Conversely, the Concat fusion method performed better when audio features were used as the second modality. Moreover, when visual features were the second modality, the Concat fusion method made the model fairer across all fairness metrics, albeit with some sacrifice in performance.

For the sentence-based approach, models using the Concat fusion method performed better compared to those using cross-attention fusion. Moreover, the sentence-based approach caused the PredEq metric to shift into the unacceptable range for all models, resulting in better performance for females.

			F1-score			**Fairness**		
M1	**M2**	**Fusion**	**Both**	**Male**	**Female**	**EAcc**	**EqOpp**	**PredEq**
w2v	visual	cross	0.922	0.888	0.940	0.95	0.91	**1.27**
		concat	0.917	0.912	0.920	0.99	0.96	0.90
	audio	cross	0.907	0.881	0.921	0.98	0.92	**1.25**
		concat	0.922	0.895	0.937	0.95	0.92	1.14
mpnet	visual	cross	0.918	0.873	0.942	0.93	0.93	**1.56**
		concat	0.932	0.895	0.952	0.94	0.91	**1.24**
	audio	cross	0.931	0.891	0.952	0.94	0.94	**1.62**
		concat*	0.937	0.888	0.963	0.92	0.91	**1.77**
	hiddenVAD	cross	0.922	0.877	0.946	0.93	0.90	**1.34**
		concat	0.933	0.891	0.955	0.93	0.91	**1.39**

Table 6.4: Results from the bimodal model experiments using the top performing features from the unimodal experiments. The w2v stands for the time-based approach whereas mpnet stands for the sentence-based approach. The best-performing bimodal model is marked with an asterisk (*) after the fusion method. Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Abbreviations: equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq), M1: the features of the first modality, M2: the features of the second modality.

6.3.2 Trimodal Experiments

For the trimodal experiments, the three fusion methods described in Figure 4.2.3 were applied to both seconds-based and sentence-based approaches. In Table 6.5, a selection of models is shown where the two best-performing models for each feature-fusion combination are displayed. Overall, the Concat fusion approach was found to be the most effective for both approaches. However, for the seconds-based approach, the proposed crisscross attention fusion achieved the same overall performance as the Concat fusion. Lastly, the results show that the fusion with hiddenVAD as one of the additional modalities made the models far more unfair regarding the PredEq metric compared to the other feature-fusion approaches.

			F1-score			Fairness		
M1	**M2 + M3**	**Fusion**	**Both**	**Male**	**Female**	**EAcc**	**EqOpp**	**PredEq**
w2v	visual, audio	criss*	0.926	0.905	0.937	0.97	0.95	**1.27**
		concat*	0.926	0.895	0.943	0.95	0.92	**1.22**
mpnet	visual, hiddenVAD	criss	0.911	0.855	0.941	0.91	0.87	**1.58**
		concat	0.925	0.891	0.942	0.95	0.94	**1.43**
	audio, visual	layered	0.919	0.887	0.936	0.95	0.95	**1.24**
		concat*	0.926	0.888	0.946	0.94	0.91	**1.28**

Table 6.5: Results from the trimodal model experiments using the top performing features from the unimodal experiments and the proposed fusion methods. The best-performing trimodal model is marked with an asterisk (*) after the fusion method. Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Abbreviations: equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq), M1: the features of the first modality, M2 + M3: the features of the second and third modalities.

6.3.3 Comparative Analysis of Best Uni-, Bi-, and Trimodal Models

Given the sizeable number of experiments performed, we evaluated the best-performing unimodal, bimodal, and trimodal models to determine whether there are any significant differences between the multimodal approaches and the unimodal textual model. Since the sentence-based approach yielded the best-performing models overall, we focused on the best-performing sentence-based model from each experiment.

In Table 6.6, the comparison is shown between these three models. However, no significant difference was found between the performances and fairness metrics of the unimodal and the other models. Since the multimodal models did not prove to be significantly better on any of the metrics, the unimodal model using mpnet sentence features was chosen to use in the bias mitigation experiments. Lastly, one notable result from this table is that the outcomes for the male gender appear to be very stable.

	F1-score			Fairness		
Model	**Both**	**Male**	**Female**	**EAcc**	**EqOpp**	**PredEq**
unimodal	0.928	0.888	0.950	0.93	0.90	1.33
bimodal	0.937	0.888	0.963	0.92	0.91	1.77
trimodal	0.926	0.888	0.946	0.94	0.91	1.28

Table 6.6: Comparative results of the best uni-, bi-, and trimodal models. Abbreviations: equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq).

6.4 Bias Mitigation Results

As described in subsection 4.3.2, we applied a series of bias mitigation methods to evaluate whether the existing gender bias can be reduced. The results of these experiments are shown in this section.

6.4.1 Sequence Length Experiments

As described in subsection 5.3.2, we experimented with increasing the sequence length of the input the model takes in. The results of these experiments are shown in Table 6.7. The results indicate that for the original bimodal cross-attention model trained using the original dataset (bimodal v1),

the EqOpp metric worsens as performance increases while the PredEq metric improves. Conversely, for the textual unimodal model, although there is a slight increase in performance, the fairness metrics remain largely unchanged.

			Fairness		
Model	**n**	**F1-score**	**EAcc**	**EqOpp**	**PredEq**
bimodal v1	596	0.534	1.11	1.14	1.12
	750	0.577	1.10	1.16	1.12
	900	0.599	1.15	**1.25**	1.05
unimodal mpnet	104	0.928	0.93	0.90	**1.33**
	140	0.931	0.94	0.91	**1.34**
	180	0.937	0.95	0.94	**1.36**

Table 6.7: Results from the experiments where the sequence length (n) of the input for the models gets increased. Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Abbreviations: equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq).

6.4.2 Gender-specific Analysis

Inspired by Yoon et al. (2022), a gender-specific analysis was conducted using some of the newly introduced features in this research to investigate how gender affects the detection of depression for these features. The results in Table 6.8 indicate that for the models using the emotion-enriched features (hiddenVAD), the male class benefits more when both classes are represented in the training dataset, while the female class only shows a slight improvement when gender-specific training gets applied. Regarding the textual features, gender-specific training slightly improves the results for the sentence-based features (mpnet), whereas this trend is inverted for the seconds-based word embedding features (w2v).

		F1-score		
Train	**Test**	**hiddenVAD**	**w2v**	**mpnet**
Male	Male	0.595	0.902	0.895
Female	Female	0.645	0.935	0.954
Both	Male	0.624	0.909	0.888
	Female	0.642	0.939	0.950

Table 6.8: Results of gender-specific training for the emotion-enriched (hiddenVAD), sentence-based (mpnet), and seconds-based (w2v) features.

6.4.3 Bias Mitigation Methods

For the experiments regarding the application of various bias mitigation methods, the methods were applied both at the session-level and sentence-level, utilising different models and feature approaches for each.

For the session-level experiments, all pre-processing and post-processing bias mitigation methods were applied to the averaged representation of each session. The results of these experiments are shown in Table 6.9. The results show that the textual modality remains effective even when a simpler model is used. However, the results also show the performance of each model decreasing with the application of the bias mitigation methods, with the subgroup and synthetic mixgendered methods showing drastic performance drops. Notably, the table shows that for these two specific

bias mitigation methods, the PredEq metric flips, indicating a higher FPR for the male class. The EAcc metric remains consistent, suggesting that the existing performance gap persists.

For the Equal Odds post-processing method, Table 6.9 shows an improvement in fairness regarding the PredEq metric while also displaying a performance that is only slightly lower than that of the original model without bias mitigation which aligns with previous research (Sogancioglu et al., 2024).

			Fairness		
Bias mitigation	**Method**	**F1-score**	**EAcc**	**EqOpp**	**PredEq**
None	–	0.921	0.96	0.88	**0.72**
Pre: oversampling	-	0.917	0.95	0.90	0.94
Pre: mixfeat	Group	0.916	0.94	0.86	**0.72**
	Mixg	0.914	0.94	0.82	**0.50**
	Subgroup	0.886	0.95	0.98	**1.27**
	Synth	0.920	0.96	0.92	0.94
	Synth mixg	0.823	0.96	1.00	1.13
Post: Equal Odds	-	0.915	0.94	0.88	0.97

Table 6.9: Results from the session-level experiments using the pre-processing bias mitigation methods. Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Abbreviations: Mixg: mixgendered. Synth: synthetic. equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq).

For the sentence-level experiments, the results in Table 6.10 show that, even though sentences from two samples may be combined without maintaining any temporal connection, the performance does not decrease that much and even improves for some methods. Specifically, the subgroup variant maintains performance levels whereas the reweighing method results in higher performance.

Overall, the EAcc metrics remain largely unchanged compared to the original method, as do the EqOpp metrics. However, the PredEq metrics increase for each mitigation method, indicating that the FPR of the male group increases more compared to the FPR for the female group. Regarding the synthetic variants, both experiments demonstrate the potential of using the MixFeat approach to generate synthetic data, as both models do not necessarily underperform in performance with respect to the other models. Lastly, Table 6.10 shows that the Equal Odds post-processing method was not able to fully mitigate the bias regarding the PredEq metric.

			Fairness		
Bias mitigation	**Method**	**F1-score**	**EAcc**	**EqOpp**	**PredEq**
None	-	0.928	0.93	0.90	**1.33**
Pre:oversampling	-	0.925	0.92	0.91	**1.81**
Pre:mixfeat	Group	0.927	0.92	0.90	**1.55**
	Mixg	0.927	0.92	0.92	**1.84**
	Subgroup	0.928	0.93	0.91	**1.45**
	Synth	0.911	0.91	0.91	**1.77**
	Synth mixg	0.923	0.93	0.95	**1.84**
In:reweighing	-	0.934	0.93	0.90	**1.59**
Post:Equal Odds	-	0.926	0.93	0.90	**1.49**

Table 6.10: Results from the sentence-level experiments using the bias mitigation methods. Fairness values outside the acceptable range of 0.8-1.2 are highlighted in bold. Scores that are significantly better than the model trained on the original data ($p < 0.05$), are highlighted in green; otherwise in red. Abbreviations: Mixg: mixgendered, Synth: synthetic, equal accuracy (EAcc), equal opportunity (EqOpp), predictive equality (PredEq).

7 Discussion and Conclusion

7.1 Conclusions

In this section, each of the sub-questions gets answered based on the findings from the experiments. The section will conclude in subsection 7.1.2 where the main research questions of this thesis will be addressed.

7.1.1 Sub-questions

For RQ 1A, we extracted the transcripts using a publicly available ASR Whisper model trained to predict approximate timestamps on speech segments. Subsequently, both pre-trained word and sentence embeddings were utilised, and a myriad of experiments were conducted based on seconds-based and sentence-based approaches. Additionally, a preprocessing step was implemented to prevent potential data leakage by removing explicit mentions of depression synonyms. All unimodal models were trained using the architecture proposed by Yoon et al. (2022). The results indicated that mpnet sentence embeddings, specifically those using spaCy's sentencizer, performed best in terms of F1-score for the unimodal model, having only a slight advantage over the model utilising word2vec word embeddings.

To determine the best fusion approach for combining audio, video, and text modalities for RQ 1B, multiple multimodal models using various fusion methods were trained. Experiments were conducted using the Concat, cross-attention module, and two novel fusion approaches (crisscross attention and layered cross-attention). Additionally, a synchronisation method was used to align seconds-based features with sentence-based features. Initially, unimodal models were trained using each modality's feature representation and their experimental variants. The features achieving the highest weighted F1-score for each modality were then combined to form the input for the multimodal models. While the unimodal models already performed very well in terms of F1-score. The results indicated that the bimodal model using mpnet sentence embeddings combined with aligned audio features, and employing the Concat fusion method, performed best in terms of F1-score. This increase in performance for the multimodal approach over an unimodal approach is consistent with prior work (Cheong, Spitale & Gunes, 2023; Flores et al., 2022; Yoon et al., 2022).

To answer RQ 1C, the performance and fairness metrics were used from the experiments conducted to answer the previous subquestions. With regard to the equal accuracy and equal opportunity metrics, the impact of the fusion approaches can be considered limited. However, for the predictive equality metric, certain fusion methods do affect the fairness of the model and may even worsen it. For instance, the cross-attention fusion approach often causes predictive equality to fall outside the acceptable range, whereas the predictions of the unimodal models using the same separate features were considered fair for this metric.

For RQ 1D, the trade-off between fairness and performance for uni- and multimodal models was assessed. The results indicated a clear trade-off for the unimodal models. Models trained using the acoustic and visual features were considered fair but lacked with regard to performance, whereas most models using textual features were more accurate but introduced some unfairness, with the model using seconds-based word2vec embeddings being the exception on this point. This trade-off persisted in the multimodal models, where some models were able to slightly increase the performance but also increased unfairness while using features that were considered fair in the unimodal context.

For RQ 2A, the trained models were screened for potential sources of gender bias. During this step, both modality-based and gender-based analyses were performed. Our findings indicate that it is not necessarily the unbalanced dataset that introduces gender bias into the model but rather the choice in modality, the specific feature embeddings and subsequent preprocessing steps that

were used. For instance, before the filtering preprocessing steps, most predictions using textual features were fair based on the chosen metrics. After the filtering, these predictions turned unfair where the female group had a much higher accuracy compared to the male group. This finding shows how embeddings may impact the model downstream and subsequently introduce bias by favouring a certain class or group which is consistent with prior work (Sogancioglu et al., 2023). Moreover, the results showed that when these features were used in a multimodal setup, they increased unfairness even though the other features used in an unimodal context were considered fair. Lastly, the results indicated that certain fusion approaches had a larger impact on the increase of bias than others.

Lastly, for RQ 2B, different bias mitigation variations were experimented upon. Originally, as stated by the research question, the best model was to be selected from the answer to RQ 1B. However, ultimately, the best unimodal textual model was chosen for the sake of experimental brevity and because there were no significant differences in scores found between the best unimodal and multimodal models. Increasing the sequence length, as proposed by Cheong, Kuzucu et al. (2023), does not fix the gender bias from the models since we only observed a slight increase in the performance of both the bimodal v1 model and the textual unimodal model while the fairness metrics remained largely unchanged. With regards to the application of the bias mitigation methods, our results are consistent with the conclusions made by Cheong, Kuzucu et al. (2023). where we do not observe a consistent improvement in performance or fairness, with the PredEq metric even consistently worsening for all mitigation methods.

7.1.2 Main Research Questions

As stated in section 1.4, this thesis aims to both introduce the textual modality as an additional feature space as well as gaining insight into the bias situation of the models. For this purpose, two research questions were devised which will be answered below.

> Does the inclusion of the textual modality contribute to the performance of the multimodal depression detection model trained on the D-Vlog dataset in terms of the F1-score?

The inclusion of the textual modality contributes positively to the performance of both the uni- and multimodal depression detection models trained on the D-Vlog dataset. The results of the experiments demonstrated that unimodal models using the textual modality outperformed all other unimodal models that used either visual, acoustic, or emotion-enriched features. Furthermore, the experiments revealed that incorporating the textual modality in fusion with other modalities could potentially further enhance the performance of the depression detection models in terms of performance.

> To what extent can the application of bias mitigation techniques help reduce existing gender bias in the best model that has been obtained while answering the previous research question?

Our experimental results showed that the choice of modality and specific feature embeddings may introduce additional gender bias into the model. Subsequently, the results showed that the application of the chosen bias mitigation techniques and their respective experimental variants did not result in a consistent reduction of the existing gender bias in the best unimodal textual model. While these findings indicate that the current bias mitigation techniques were not effective for the unimodal textual model trained on the D-Vlog dataset, this does not imply that these techniques cannot be useful. They may still prove beneficial when applied to different model architectures, fusion approaches or feature representations.

7.2 Discussion

In this section, we will discuss several challenges and limitations encountered during the research.

7.2.1 Absence of Available Code

One of the significant limitations of this study is the absence of available source code from the original paper (Yoon et al., 2022) as well as from a subsequent study (Cheong, Kuzucu et al., 2023) where a replication study was performed. The lack of a public architecture implementation posed certain challenges in ensuring our code remained faithful to the initial research. As a result of this limitation, the original model was reconstructed and re-implemented based on the descriptions, illustrations and parameters provided in the literature. This process of reconstructing the model could introduce some discrepancies in our implementation and results, as there is a possibility that the variations we (un)consciously introduce in our implementation may affect our final results.

This limitation underlines the importance of open-source code in research for reproducibility and validation purposes. As a response to this and to contribute to the transparency and reproducibility of research in this domain, we make the code open-source by making the repository publicly accessible[1].

7.2.2 Dataset Constraints in Comparative Analysis

As previously mentioned in section 5.1, we were unable to retrieve all the transcripts for the videos in the original D-Vlog dataset. Consequently, we created a separate dataset consisting of all vlogs that were available while taking the original class imbalance into account. Despite our results demonstrating that we outperform models from the other studies trained on the original v1 dataset using the original experimental protocol, a direct comparison is not feasible since we used a different data split. Therefore, while our findings are promising, they should be interpreted with caution considering the differences in dataset composition.

7.2.3 Training with Uniform Hyperparameters

At this time, all uni-, bi-, and multimodal models were trained using the same hyperparameters wherever applicable. The only variation in hyperparameters was the number of heads in the cross-attention, which was specifically experimented upon for replication purposes. Consequently, this approach could imply that some models may have been trained with suboptimal parameters. Suboptimal training can impact the performance and fairness of the models, which means that the results for some models will not fully capture their true potential regarding performance and fairness (Nguyen et al., 2023).

7.2.4 Data in the Wild vs Clinical Data

The D-Vlog dataset consists of self-identified depressed individuals. A benefit of collecting such a dataset is that it easily scales since the data is more readily available and less intensive to procure. However, it also has the potential to introduce problematic *self-disclosure* bias and label noise making the collected samples less representative (Aguirre et al., 2021). The subjects in the vlogs may selectively share only certain symptoms or experiences, omitting aspects of their condition. Moreover, the reliance on self-identification could lead to inaccuracies and inconsistencies in how depression is labelled and interpreted. Both these issues can affect the generalisability of the model and potentially the performance if the model is used in the real world.

7.2.5 Variety of the Collected Data

The D-Vlog dataset was collected using specific search keywords to collect both depression and non-depression vlogs in which both self-identified depressed and non-depressed individuals talk about topics that relate to them. This approach can lead to spurious correlations that models might exploit to achieve higher precision (Hovy & Prabhumoye, 2021). For example, non-depressed vlogs might contain a wider variety of topics and tones, while depressed vlogs may focus more narrowly

[1] https://github.com/StanMey/Master_thesis_DVlog

on specific themes related to depression. This can cause the models to overfit on depression-related signals (Wolohan et al., 2018).

As can be seen in Table 5.3, the depressive subgroup already tends to use more words related to symptoms or consequences of depression. Consequently, the model might learn to associate certain topics, keywords, or emotional tones with depression, and develop some conversational topic bias rather than identifying underlying depressive symptoms (Wolohan et al., 2018).

As described in subsection 4.2.1, we did perform additional data preprocessing steps to prevent potential data leakage caused by explicit mentions of depression symptoms in the textual data. However, we do still have specific depression vlogs in which people speak openly about being depressed. In contrast, in real life they may not be so open about this which means that we might still end up with a model having some conversational topic bias (Wolohan et al., 2018).

7.3 Future Research

This section discusses several possibilities for future research.

7.3.1 Additional Features and Architectures

This thesis investigated the impact of adding the textual modality and various fusion approaches, including two novel experimental fusion methods, on the overall performance and fairness of uni- and multimodal depression detection models. We utilised the original acoustic and video embeddings and introduced emotion-enriched embeddings to enhance our models.

Given that there are gender differences in depression manifestation and diagnosis (Cheong, Kuzucu et al., 2023), future research could explore the addition of features that model depression in more meaningful ways. For instance, Sun et al. (2024) and Cheong, Spitale and Gunes (2023) successfully extracted features using sentiment analysis and applied them to the depression detection task, indicating the potential value of such features in this context.

Additionally, integrating existing pre-trained, grounded video-language models or models trained using Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) could be highly beneficial. These multimodal models often leverage vast datasets to understand the interplay between multiple modalities, which might allow them to capture subtle cues and correlations that regular unimodal models might miss ultimately resulting in a more comprehensive understanding and better performance.

Lastly, with regards to alternative architectures, it would be beneficial to use a GCN model on the D-Vlog dataset while leveraging the textual modality. Previous research by Sun et al. (2024) showed their model outperforming other approaches while only using the acoustic features from the D-Vlog. Using such a model can also benefit interpretability and feature importance analysis, which will be discussed in section 7.3. Burdisso et al. (2024) used the GCN model for a qualitative analysis on the DAIC-WOZ dataset, demonstrating the model's interpretability property by displaying the segments the model focused on when predicting whether a person was depressed.

7.3.2 Cross-corpus Analysis

In the original D-Vlog paper by Yoon et al. (2022), the researchers performed a cross-corpus validation using the clinically labelled DAIC-WOZ dataset alongside their proposed model and dataset. This analysis aimed to assess how the D-Vlog dataset could contribute to depression research and focused on the acoustic and visual features of both datasets.

Given the addition of the textual modality in this study, a valuable future research step would be to replicate this cross-corpus validation using the textual modality as an additional modality. This would involve evaluating whether the findings reported by Yoon et al. (2022) still hold when incorporating the textual modality, providing a more comprehensive understanding of how the D-Vlog dataset and model could contribute to depression research.

7.3.3 Hyperparameter Search

As described in subsection 7.2.3, at this time, all uni-, bi-, and multimodal models were trained using the same hyperparameters wherever applicable. This approach could have led to some models being trained with suboptimal parameters, potentially impacting their fairness and performance. For this reason, applying hyperparameter tuning to some of the models could benefit both the performance and fairness of these models since past research has demonstrated the effectiveness of parameter tuning with regards to enhancing the fairness of models without sacrificing accuracy (Nguyen et al., 2023; Tizpaz-Niari et al., 2022).

7.3.4 Feature Importance Analysis

In this thesis, we primarily examined the impact of adding various combinations of features using specific fusion methods on the overall performance and fairness of uni- and multimodal depression detection models. However, a more in-depth analysis of (textual) feature importance would provide valuable insights into the contribution of each feature and modality.

Moreover, by specifically doing this analysis using the textual modality, we can conduct a topic analysis to uncover any potential biases in conversational topics within the dataset. This, in turn, can inform us about the presence and nature of any conversational topic biases.

One potential approach, inspired by Min et al. (2023), is to utilise explainable models in conjunction with interpretable features. While this approach may not be suitable for word and sentence embeddings directly, an alternative could involve training a depression detection model using separate sentences as input. This method would allow us to rank sentences for each sample, helping to identify which sentences are more critical for the model's predictions. Moreover, analysing these rankings could reveal important linguistic patterns and cues associated with depression.

Furthermore, to analyse the importance of combined modalities, we could implement a depression detection model using the fusion approaches proposed by Fang et al. (2023) or Shen et al. (2022). These approaches involve training a separate attention weight vector, which should reflect the internal relationships and importance of each modality. By examining these attention weights, we can gain a better understanding of how different modalities contribute to the model's decision-making process.

Lastly, another potential approach regarding explaining the importance of the individual modalities to the eventual prediction is to follow the approach by Gimeno-Gómez et al. (2024). The authors used the Integrated Gradients (Sundararajan et al., 2017) attribution method which can give attribution scores per modality over a select window.

7.4 Ethical Considerations

This section discusses some of the key ethical issues and concerns that may or have arisen during this project. This section will address issues such as the protection of participant privacy, ensuring data security, any issues with the dataset that is being used, and the potential implications of the usage of the model.

Ethics and Privacy Quick Scan Prior to this thesis, the Ethics and Privacy Quick Scan (Utrecht University, n.d.) from the Research Institute of Information and Computing Sciences of the Utrecht University was conducted where it classified this research as low-risk with no fuller ethics review or privacy assessment required.

Privacy and Anonymity of Subjects The first consideration centres around the privacy of the persons in the vlogs. The authors of the original D-Vlog dataset (Yoon et al., 2022) have provided de-identified anonymized data consisting of the extracted acoustic and visual features. However, since one of the aims of this thesis is to add the textual modality to the model the

provided dataset proved to be inadequate. To retrieve the captions of the vlogs, the authors supplied identifies for each vlog, which were used to extract the textual content.

To retrieve these captions, it was necessary to download the corresponding vlogs, a process not condoned by YouTube's terms of service and to which participants did not explicitly consent. To address these privacy concerns, strict measures were implemented: the videos were downloaded solely for caption extraction purposes, and the extraction process was conducted locally to prevent any kind of data transmission to third parties. Subsequently, the videos were promptly deleted from the researcher's machine to minimise any potential privacy risks.

However, despite these precautions, the risk remains that the usage of captions could facilitate the identification process of the individuals in the vlogs. For the DAIC-WOZ interviews, where the researchers transcribed interviews from the audio stream, the captions were de-identified by removing identifying information (Gratch et al., 2014) to mitigate this risk. For our dataset, the captions were transformed into word and sentence embeddings which enables us to preserve the individual's privacy and anonymity while still enabling a researcher to do meaningful analysis.

By employing the safeguards described above, we try to balance the necessity of retrieving the textual data which was needed for performing the research with the need to protect the privacy and anonymity of the vlog participants.

Usage of Binary Labels The next consideration regards the usage of binary labels for gender. Since the D-Vlog dataset has been labelled using the sex categories of male and female, we will continue to use these labels as binary categories. Nevertheless, we acknowledge that many individuals do not fit into these binary categories and that they often experience higher mental health disorders at a higher rate compared to their heterosexual counterparts (McDonald, 2018).

Interpretation and Application of a Depression Detection Model The accuracy of the implemented depression detection scores in this study warrants cautious interpretation, particularly within the context of early-stage depression detection. As stated by Min et al. (2023), the trade-off between false positives and negatives comes with significant implications since false positives may lead to unnecessary interventions while false negatives may result in missed opportunities for early intervention.

Furthermore, it is important to clarify the purpose of the depression detection model with regards to the application. The primary goal of this model would be to assist the practitioner with the early-stage detection of depression instead of replacing this clinical judgement. This distinction is crucial, especially in light of the regulatory framework of the AI act (European Parliament, 2023; Meaker, 2023), which emphasises addressing and mitigating possible risks specifically created through the use of AI systems.

Environmental Impact This ethical concern relates to the environmental impact of developing and training ML models. These processes are often resource-intensive, require substantial amounts of electricity and hardware, and contribute to a large carbon footprint (Dhar, 2020; Lacoste et al., 2019). To lessen this impact, pre-trained models are utilised whenever possible to reduce resource consumption. Furthermore, we chose to train all models using GPU instead of CPU as this was the more efficient hardware for this task and the model.

Acronyms

AI Artificial Intelligence. 23, 26
ASR Automatic Speech Recognition. 27, 38, 50
AttFN attentional fusion network. 19

CLIP Contrastive Language-Image Pre-Training. 53
CPU central processing unit. 55

DL Deep Learning. 10, 18
DTW Dynamic Time Warping. 27

EAcc equal accuracy. 33, 34, 43–49
ECG electrocardiogram. 24
EEG electroencephalography. 26
eGeMAPS extended Geneva Minimalistic Acoustic Parameter Set. 21
EqOpp equal opportunity. 33, 42–49

FN False Negative. 7
FP False Positive. 7
FPR False Positive Rate. 6, 7, 42, 48

GCN graph convolutional network. 22, 53
GPU graphics processing unit. 55
GRU gated recurrent unit. 11
grwm get ready with me. 20

LLD low-level acoustic descriptor. 20, 21
LSTM long short-term memory. 11

MDSER Model for Dimensional Speech Emotion Recognition. ix, 29, 32, 40, 43, 44
MFCC Mel-frequency ceptral coefficients. 16
ML Machine Learning. 1, 5, 6, 8–10, 16, 23–25, 55

NLP Natural Language Processing. 14
NN Neural Network. 11

PredEq predictive equality. 33, 42–49, 51
PTSD post-traumatic stress disorder. 17

RNN recurrent neural network. 11
ROC Reject Option Classification. 10, 23, 35
RQ 1A Which SOTA pre-trained word embeddings and sentence embeddings perform best in terms of F1-score for unimodal depression prediction on the D-Vlog dataset?. 3, 50
RQ 1B Which fusion approach to the audio, video and text modalities gives the best performance regarding the F1-score?. 3, 50, 51
RQ 1C What is the impact of the fusion approaches of subquestion **1B** on the fairness of the model in terms of equal accuracy, equal opportunity, and predictive equality?. 3, 50
RQ 1D Is there a trade-off between fairness and performance for uni- and multimodal models with different fusion schemes?. 3, 50
RQ 2A What are the potential sources of gender bias (e.g., data imbalance, feature representation, modality) in the models trained on the D-Vlog dataset?. 50
RQ 2B What bias mitigation approach is most effective in mitigating existing gender bias contained in the best model that has been obtained during subquestion **1B**?. 51

SER Speech Emotion recognition. 29
SOTA state-of-the-art. 3, 14, 27, 56
SVM support vector machine. 36, 40

TAMF time-aware attention multimodal fusion. 20
TP True Positive. 7
TPR True Positive Rate. 6, 7

VAD arousal, dominance, and valence. 29, 43

WHO World Health Organization. 1

Bibliography

Aguirre, C., Harrigian, K., & Dredze, M. (2021, April). Gender and racial fairness in depression research using social media. In P. Merlo, J. Tiedemann & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 2932–2949). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.256

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., et al. (2012). From joyous to clinically depressed: Mood detection using spontaneous speech. *FLAIRS Conference*, *19*.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Hyett, M., Parker, G., & Breakspear, M. (2016). Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, *9*(4), 478–490.

Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G., & Breakspear, M. (2013). Head pose and movement analysis as an indicator of depression. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 283–288.

Alhanai, T., Ghassemi, M. M., & Glass, J. R. (2018). Detecting depression with audio/text sequence modeling of interviews. *Interspeech*, 1716–1720.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bailey, A., & Plumbley, M. D. (2021). Gender bias in depression detection using audio features. *2021 29th European Signal Processing Conference (EUSIPCO)*, 596–600.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018, October). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. https://arxiv.org/abs/1810.01943

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information (L. Lee, M. Johnson & K. Toutanova, Eds.). *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

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

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021). Bias and fairness in multimodal machine learning: A case study of automated video interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 268–277.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Bueno-Notivol, J., Gracia-García, P., Olaya, B., Lasheras, I., López-Antón, R., & Santabárbara, J. (2021). Prevalence of depression during the covid-19 outbreak: A meta-analysis of community-based studies. *International journal of clinical and health psychology*, *21*(1), 100196.

Burdisso, S., Reyes-Ramírez, E., Villatoro-Tello, E., Sánchez-Vega, F., López-Monroy, P., & Motlicek, P. (2024). Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. *arXiv preprint arXiv:2404.14463*.

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

Cha, J., Kim, S., & Park, E. (2022). A lexicon-based approach to examine depression detection in social media: The case of twitter and university community. *Humanities and Social Sciences Communications*, *9*(1), 1–10.

Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023, August). Towards gender fairness for mental health prediction [AI for Good]. In E. Elkind (Ed.), *Proceedings of the thirty-second international joint conference on artificial intelligence, IJCAI-23* (pp. 5932–5940). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2023/658

Cheong, J., Spitale, M., & Gunes, H. (2023). "it's not fair!" – fairness for a small dataset of multi-modal dyadic mental well-being coaching. https://doi.org/10.17863/CAM.97210

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., & Tao, C. (2018). Xa-bilstm: A deep learning approach for depression detection in imbalanced data. *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 1624–1627.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Craft, L. L., & Landers, D. M. (1998). The effect of exercise on clinical depression and depression resulting from mental illness: A meta-analysis. *Journal of Sport and Exercise Psychology*, *20*(4), 339–357.

Dang, V. N., Cascarano, A., Mulder, R. H., Cecil, C., Zuluaga, M. A., Hernández-González, J., & Lekadir, K. (2022). Fairness and bias correction in machine learning for depression prediction: Results from four different study populations. *arXiv preprint arXiv:2211.05321*.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the international AAAI conference on web and social media*, *7*(1), 128–137.

de Jesús Titla-Tlatelpa, J., Ortega-Mendoza, R. M., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2021). A profile-based sentiment-aware approach for depression detection in social media. *EPJ data science*, *10*(1), 54.

Delobelle, P., Winters, T., & Berendt, B. (2020, November). RobBERT: A Dutch RoBERTa-based Language Model. In T. Cohn, Y. He & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 3255–3265). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.292

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

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, *2*(8), 423–425.

Dibeklioğlu, H., Hammal, Z., Yang, Y., & Cohn, J. F. (2015). Multimodal detection of depression in clinical interviews. *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 307–310.

European Parliament. (2023, December 9). *Artificial intelligence act: Deal on comprehensive rules for trustworthy ai* [Press release]. Retrieved January 29, 2024, from https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190–202.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.

Fang, M., Peng, S., Liang, Y., Hung, C.-C., & Liu, S. (2023). A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, *82*, 104561.

Flores, R., Tlachac, M., Toto, E., & Rundensteiner, E. (2022). Audiface: Multimodal deep learning for depression screening. *Proceedings of the 7th Machine Learning for Healthcare Conference*, *182*, 609–630. https://proceedings.mlr.press/v182/flores22a.html

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gimeno-Gómez, D., Bucur, A.-M., Cosma, A., Martínez-Hinarejos, C.-D., & Rosso, P. (2024). Reading between the frames: Multi-modal depression detection in videos from non-verbal cues. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald & I. Ounis (Eds.), *Advances in information retrieval* (pp. 191–209). Springer Nature Switzerland.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, *31*(7). https://doi.org/10.18637/jss.v031.i07

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., & Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3123–3128. http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

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

Hasan, M. K., Lee, S., Rahman, W., Zadeh, A., Mihalcea, R., Morency, L.-P., & Hoque, E. (2021). Humor knowledge enriched transformer for understanding multimodal humor. *Proceedings of the AAAI conference on artificial intelligence*, *35*(14), 12972–12980.

Hawton, K., i Comabella, C. C., Haw, C., & Saunders, K. (2013). Risk factors for suicide in individuals with depression: A systematic review. *Journal of affective disorders*, *147*(1-3), 17–28.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., et al. (2022). Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, *80*, 56–86.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, *12*(1), 5979.

Hort, M., Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). Bias mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and linguistics compass*, *15*(8), e12432.

Jans-Beken, S. J. T. (2021). *You sound depressed: The effect of depression and gender on self-referencing and the use of negative emotion words*. Tilburg University.

Jia, S., Meng, T., Zhao, J., & Chang, K.-W. (2020). Mitigating gender bias amplification in distribution by posterior regularization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2936–2942. https://doi.org/10.18653/v1/2020.acl-main.264

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54.

Jurafsky, D., & Martin, J. H. (2022). *Speech and language processing* (3th ed.) Stanford. Retrieved June 23, 2022, from https://web.stanford.edu/~jurafsky/slp3/

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, *33*(1), 1–33.

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

Kaya, H., Fedotov, D., Dresvyanskiy, D., Doyran, M., Mamontov, D., Markitantov, M., Akdag Salah, A. A., Kavcar, E., Karpov, A., & Salah, A. A. (2019). Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 27–35. https://doi.org/10.1145/3347320.3357691

Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors*, *24*(2). https://doi.org/10.3390/s24020348

Kim, Y., Lee, H., & Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. *2013 IEEE international conference on acoustics, speech and signal processing*, 3687–3691.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, *10*, 1755–1758.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of 3rd International Conference on Learning Representations (ICLR)*.

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Li, P., Gu, J., Kuen, J., Morariu, V. I., Zhao, H., Jain, R., Manjunatha, V., & Liu, H. (2021). Selfdoc: Self-supervised document representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5652–5660.

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., & Elliott, D. (2021, November). Visually grounded reasoning across languages and cultures. In M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10467–10485). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.818

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Louradour, J. (2023). Whisper-timestamped.

Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. *Measurement: Sensors*, *25*, 100587.

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, *3*(11), e442.

McDonald, K. (2018). Social support and mental health in lgbtq adolescents: A review of the literature. *Issues in mental health nursing*, *39*(1), 16–29.

Meaker, M. (2023). The EU just passed sweeping new rules to regulate AI. Retrieved January 29, 2024, from https://www.wired.com/story/eu-ai-act/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, *54*(6). https://doi.org/10.1145/3457607

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Min, K., Yoon, J., Kang, M., Lee, D., Park, E., & Han, J. (2023). Detecting depression on video logs using audiovisual features. *Humanities and Social Sciences Communications*, *10*(1), 1–8.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163.

Mosteiro, P., Kuiper, J., Masthoff, J., Scheepers, F., & Spruit, M. (2022). Bias discovery in machine learning models for mental health. *Information*, *13*(5), 237.

Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine*, *211*, 106433.

Newell, E. E., McCoy, S. K., Newman, M. L., Wellman, J. D., & Gardner, S. K. (2018). You sound so down: Capturing depressed affect through depressed language. *Journal of Language and Social Psychology*, *37*(4), 451–474. https://doi.org/10.1177/0261927X17731123

Nguyen, G., Biswas, S., & Rajan, H. (2023). Fix fairness, don't ruin accuracy: Performance aware fairness repair using automl. *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 502–514. https://doi.org/10.1145/3611643.3616257

Niu, M., Tao, J., Liu, B., Huang, J., & Lian, Z. (2020). Multimodal spatiotemporal representation for automatic depression level detection. *IEEE transactions on affective computing*, *14*(1), 294–307.

Pampouchidou, A., Simantiraki, O., Fazlollahi, A., Pediaditis, M., Manousos, D., Roniotis, A., Giannakakis, G., Meriaudeau, F., Simos, P., Marias, K., Yang, F., & Tsiknakis, M. (2016). Depression assessment by fusing high and low level features from audio, video, and text. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 27–34. https://doi.org/10.1145/2988257.2988266

Park, J., & Moon, N. (2022). Design and implementation of attention depression detection model based on multimodal analysis. *Sustainability*, *14*(6), 3569.

Park, Y., Hu, J., Singh, M., Sylla, I., Dankwa-Mullan, I., Koski, E., & Das, A. K. (2021). Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA network open*, *4*(4), e213909–e213909.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 560–568.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Picard, D. (2021). Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning*, 28492–28518.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

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

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng & X. Wan (Eds.), *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. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, *71*, 103107.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 1–7.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., & Pantic, M. (2019). Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 3–12. https://doi.org/10.1145/3347320.3357688

Rodrigues Makiuchi, M., Warnita, T., Uto, K., & Shinoda, K. (2019). Multimodal fusion of bert-cnn and gated cnn representations for depression detection. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 55–63.

Ross, C., Katz, B., & Barbu, A. (2021, June). Measuring social biases in grounded vision and language embeddings. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou (Eds.), *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 998–1008). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.78

Rottenberg, J., Gross, J. J., & Gotlib, I. H. (2005). Emotion context insensitivity in major depressive disorder. *Journal of abnormal psychology*, *114*(4), 627.

Samareh, A., Jin, Y., Wang, Z., Chang, X., & Huang, S. (2018). Predicting depression severity by multi-modal feature engineering and fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Shen, Y., Yang, H., & Lin, L. (2022). Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6247–6251.

Smirnova, D., Cumming, P., Sloeva, E., Kuvshinova, N., Romanov, D., & Nosachev, G. (2018). Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in psychiatry*, *9*, 105.

Sogancioglu, G., Kaya, H., & Salah, A. A. (2023). The effects of gender bias in word embeddings on patient phenotyping in the mental health domain. *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. https://doi.org/10.1109/ACII59096.2023.10388203

Sogancioglu, G., Kaya, H., Salah, A. A., Mosteiro, P., & Scheepers, F. (2024). *Exploring perspectives of clinicians on gender fairness in mental healthcare: Insights from interviews and the impact of bias mitigation methods* [Unpublished manuscript. Department of Information and Computing Sciences, Utrecht University.].

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics*, *10*(1), 1–19.

Stratou, G., Scherer, S., Gratch, J., & Morency, L.-P. (2013). Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 147–152.

Sun, C., Jiang, M., Gao, L., Xin, Y., & Dong, Y. (2024). A novel study for depression detecting using audio signals based on graph neural network. *Biomedical Signal Processing and Control*, *88*, 105675. https://doi.org/https://doi.org/10.1016/j.bspc.2023.105675

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International conference on machine learning*, 3319–3328.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Tan, Y. C., & Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, *32*.

Timmons, A. C., Duong, J. B., Fiallo, N. S., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, *18*(5), 1062–1096. https://doi.org/10.1177/17456916221134490

Tizpaz-Niari, S., Kumar, A., Tan, G., & Trivedi, A. (2022). Fairness-aware configuration of machine learning libraries. *Proceedings of the 44th International Conference on Software Engineering*, 909–920. https://doi.org/10.1145/3510003.3510202

Toto, E., Tlachac, M., & Rundensteiner, E. A. (2021). Audibert: A deep transfer learning multimodal classification framework for depression screening. *Proceedings of the 30th ACM international conference on information & knowledge management*, 4145–4154.

Üstün, T. B., Ayuso-Mateos, J. L., Chatterji, S., Mathers, C., & Murray, C. J. L. (2004). Global burden of depressive disorders in the year 2000. *The British Journal of Psychiatry*, *184*(5), 386–392. https://doi.org/10.1192/bjp.184.5.386

Utrecht University. (n.d.). *Ethics and privacy*. Retrieved June 26, 2024, from https://www.uu.nl/en/research/institute-of-information-and-computing-sciences/ethics-and-privacy

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

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–13.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, 1–10.

Wolohan, J., Hiraga, M., Mukherjee, A., Sayyed, Z. A., & Millard, M. (2018, August). Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In M. Sinha & T. Dasgupta (Eds.), *Proceedings of the first international workshop on language cognition and computational models* (pp. 11–21). Association for Computational Linguistics. https://aclanthology.org/W18-4102

Yadav, U., & Sharma, A. K. (2023). A novel automated depression detection technique using text transcript. *International Journal of Imaging Systems and Technology*, *33*(1), 108–122.

Yan, S., Huang, D., & Soleymani, M. (2020). Mitigating biases in multimodal personality assessment. https://doi.org/10.1145/3382507.3418889

Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., & Sahli, H. (2016). Decision tree based depression classification from audio video and language information. *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 89–96.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H. (2017). Multimodal measurement of depression using deep learning models. *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 53–59.

Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C., & Jiang, D. (2017). Hybrid depression classification and estimation from audio video and text information. *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, 45–51.

Ye, J., Yu, Y., Wang, Q., Li, W., Liang, H., Zheng, Y., & Fu, G. (2021). Multi-modal depression detection based on emotional audio and evaluation text. *Journal of Affective Disorders*, *295*, 904–913.

Yepes, A. J. (2017). Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. *Journal of biomedical informatics*, *73*, 137–147.

Yin, D., Li, L. H., Hu, Z., Peng, N., & Chang, K.-W. (2021, November). Broaden the vision: Geo-diverse visual commonsense reasoning. In M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 2115–2129). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.162

Yoon, J., Kang, C., Kim, S., & Han, J. (2022). D-vlog: Multimodal vlog dataset for depression detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12226–12234.

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

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021). A transformer-based framework for multivariate time series representation learning. *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2114–2124.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, *6*(1), 52.

Zhou, L., Liu, Z., Shangguan, Z., Yuan, X., Li, Y., & Hu, B. (2022). Tamfn: Time-aware attention multimodal fusion network for depression detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 669–679.

Zhou, L., Liu, Z., Yuan, X., Shangguan, Z., Li, Y., & Hu, B. (2023). Caiinet: Neural network based on contextual attention and information interaction mechanism for depression detection. *Digital Signal Processing*, *137*, 103986. https://doi.org/https://doi.org/10.1016/j.dsp.2023.103986

A Appendix

A.1 Extracted Transcript Example


```
{
  "text": " So I wanted to come on here and sit down with you guys...",
  "segments": [
    {
      "start": 0.6,
      "end": 6.02,
      "text": " So I wanted to come on here and...",
      "tokens": [
        50364, 407, 286, 1415, 281, 808, 322, 510, 293, 1394, 760,...
      ],
      "temperature": 0.0,
      "no_speech_prob": 0.03191100433468819,
      "confidence": 0.868,
      "words": [
        {
          "text": "So",
          "start": 0.6,
          "end": 0.76,
          "confidence": 0.725
        },
        {
          "text": "I",
          "start": 0.76,
          "end": 0.84,
          "confidence": 0.853
        },
        {
          "text": "wanted",
          "start": 0.84,
          "end": 1.04,
          "confidence": 0.938
        },
        ...
    ]}]
}
```


Figure 9: A shortened and simplified example of a retrieved transcript using whisper-timestamped.

A.2 Process Workflows for Answering Research Questions

A.2.1 Process Workflow First Research Question

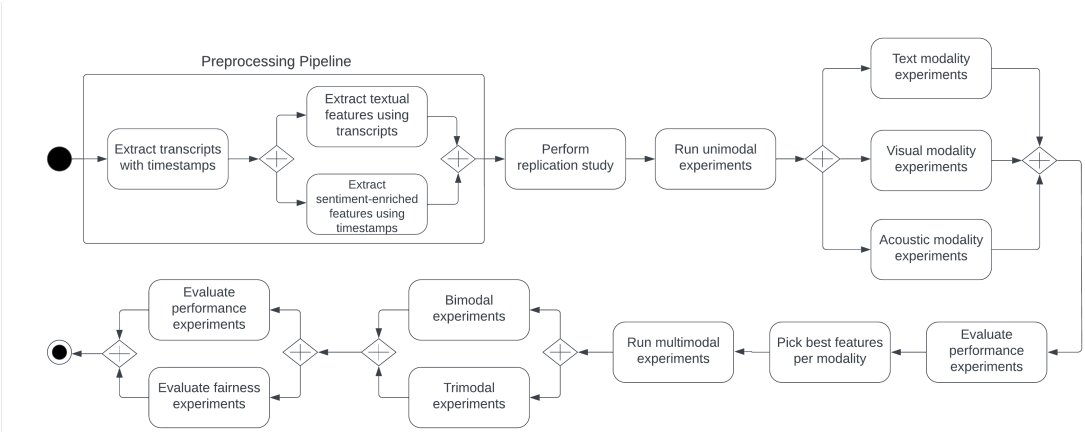

Figure 10: The research process workflow used for answering the first research question. The Preprocessing Pipeline refers to the feature extraction pipeline shown in Figure 4.

A.2.2 Process Workflow Second Research Question

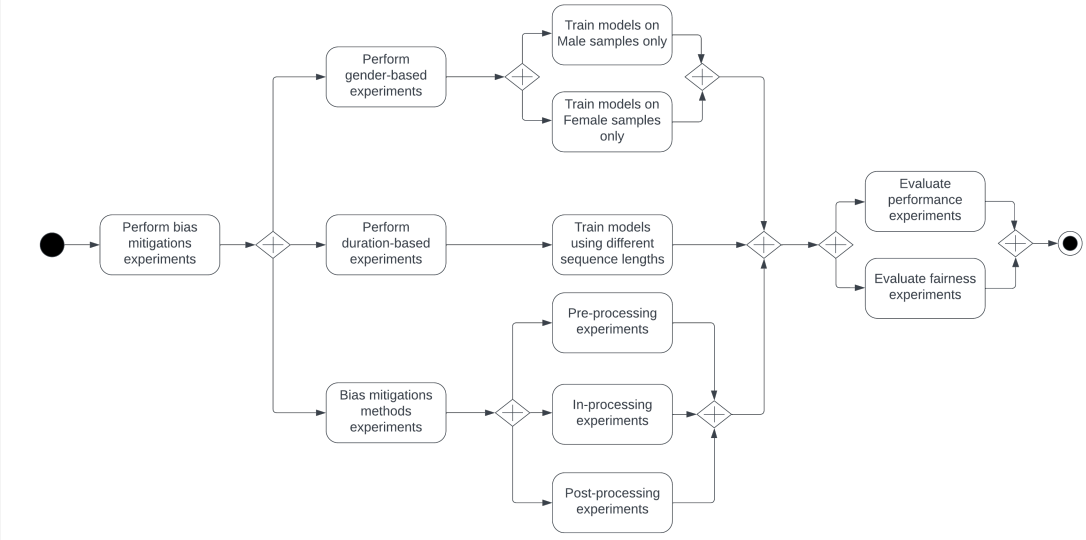

Figure 11: The research process workflow used for answering the second research question.

A.3 The Label-gender Combination Linguistic Features Table

	Depression				**Normal**			
	female		male		female		male	
n-gram	word(s)	count	word(s)	count	word(s)	count	word(s)	count
1	biscuit	23	neardeath	18	thesis	39	tammy	58
	psychiatric	16	gagging	11	sephora	34	assassin	21
	legitimately	15	cranium	9	weaving	33	shadowing	18
	decis	15	cratum	9	binder	29	creed	17
	ian	13	kava	9	cosmetology	22	timber	16
	waving	11	hazing	8	scent	21	paralysis	15
	hotline	11	ordinary	8	rite	21	surgeon	15
	smacker	11	gon	7	bracelet	20	hearthstone	14
	postpartum	10	wim	7	planner	20	spawn	13
	azda	10	hof	7	sloth	20	uganda	12
2	penn teller	16	neardeath experience	18	double crochet	35	treasure hunt	34
	im angry	13	ice bath	14	peer pressure	26	string pattern	26
	depressed people	12	glory sun	14	chain one	23	jerk bait	25
	keep thinking	10	glory glory	12	cosmetology school	21	assassin creed	17
	chronic pain	10	santa highest	9	rite aid	21	mr gate	16
	leave room	10	low testosterone	9	passive voice	18	adult adhd	14
	going cry	10	moving la	8	going chain	16	sleep paralysis	13
	lip smacker	10	vera juice	8	pink one	16	governing body	12
	decis decis	10	beat as	7	nursing school	15	student number	11
	pain crisis	10	rock bottom	7	love color	15	gender role	10

Table A.1: The distinctive lemmatised linguistic features prevalent for all of the label-gender combinations.

A.4 Sample of Depression Synonyms Lexicon

anxiety, insomnia, suicidal, dysfunction, fatigue, imbalance, nervousness, sad, lonely, unhappy, no focus, disturbed, isolation, lack of interest, low interaction, sleep problems, loss of meaning in life, stressed, uneasiness, instability, moody, emotional, low self-esteem, have no emotional support, depressed, depression, suicide, broken, killme, worthless, selfharm, pain
affect unhappy, affects lack, anhedonia, anhedonias, apathetic, apathetic behavior, apathetic behaviour, apathy, cannot see a future, decreased interest, decreased mood, demoralisation, demoralization, depressed, depressed mood, depressed state, depressing, depression, depressions, depression moods, depression psychic, depression symptom, depressive state, depressive symptom, depressive symptoms, despair, diminished pleasure, emotional depression, emotional indifference, emotionally apathetic, emotionally cold, emotionally detached, emotionally distant, emotionally subdued, feeling blue, feeling despair, feeling depressed, feeling down, feeling empty, feeling helpless, feeling hopeless, feeling isolated, feeling lost, feeling low, feeling of despair, feeling of hopelessness, lost feeling, low mood, melancholia, melancholic, melancholy, mental depression, miserable, mood depressed, mood depression, mood depressions, morose mood, morosity, negative about the future, no hope for the future, nothing matters, powerlessness, sad, sad mood, sadness, stuporous, symptoms of depression, torpid, unhappiness, unhappy, worthless, worthlessness, anxiety depressed, anxious depressed, anxious depressed mood, anxious sad, anxiousdepressed, appeared depressed, appeared sad, appears depressed, appears sad, appetite anhedonia, appetite hopelessness, appropriately sad, battling depression, became sad, became very depressed, become depressed, become increasingly depressed, becomes depressed, becoming depressed, becoming more depressed, being sad, bluesdepression, bpii current episode depressed, catatonia depressed mood, characterlogical depression however, chronically depressed, comorbid depression, concentration anhedonia, consistently depressed mood, constricted sad, current depressive episode, current epidose depressed, current episode depressed, dailydepression, dark thoughts, decreased motivation, deperssion, depession, deppression, deppressive, depreesion, deprerssion, depresed...

A.5 Complete Training Performance Overview of All Uni- and Multimodal Models

A.5.1 Replication Study Training Results

			Val performance				Test performance			
Dataset	**Modality**	**Fusion**	**Precision**	**Recall**	**F1 (avg)**	**F1 (std)**	**Precision**	**Recall**	**F1 (avg)**	**F1 (std)**
v1	audio	-	0.620	0.606	0.591	0.074	0.626	0.610	0.602	0.058
	visual	-	0.630	0.614	0.613	0.017	0.626	0.608	0.608	0.040
	both	concat	0.634	0.625	0.620	0.015	0.623	0.619	0.613	0.037
		cross8	0.623	0.614	0.586	**0.082**	0.627	0.622	0.598	**0.074**
		cross16	0.558	0.606	0.553	**0.108**	0.542	0.587	0.534	**0.077**
v2	audio	-	0.573	0.614	0.576	**0.166**	0.578	0.618	0.581	**0.164**
	visual	-	0.670	0.662	0.660	0.020	0.629	0.613	0.608	0.033
	both	concat	0.708	0.698	0.697	0.021	0.624	0.615	0.613	0.023
		cross8	0.731	0.669	0.655	**0.082**	0.622	0.576	0.536	**0.130**
		cross16	0.741	0.664	0.648	0.052	0.642	0.582	0.544	**0.107**

Table A.2: Overview of the performance of the models trained for the replication study on the validation and test split. Values where the standard deviation exceeds 10% of the averaged F1-score are highlighted in **bold**.

A.5.2 Unimodal Training Results

	Val performance			Test performance		
Method	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
hidden	**0.743**	**0.738**	**0.734**	0.636	0.636	0.634
hidden + VAD	0.727	0.726	0.724	**0.639**	**0.638**	**0.636**
hidden + z-score VAD	0.739	0.731	0.725	0.629	0.628	0.624

Table A.3: Overview of the performance of the models trained on the validation and test split using emotion-enriched embeddings. The highest values per split and metric are highlighted in **bold**.

			Val performance			Test performance		
Feature	**type**	**with keyw**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
w2v	avg	yes	0.944	0.943	0.943	0.912	0.912	0.912
		no	0.921	0.919	0.919	0.929	0.928	0.929
	zero	yes	0.937	0.936	0.936	0.946	0.945	0.945
		no	0.922	0.921	0.921	0.913	0.913	0.913
	sent	yes	0.926	0.924	0.923	0.947	0.944	0.944
		no	0.864	0.864	0.864	0.882	0.879	0.878
biow	avg	yes	0.951	0.950	0.950	0.869	0.867	0.866
		no	0.939	0.938	0.938	0.862	0.859	0.859
	zero	yes	0.941	0.940	0.940	0.878	0.876	0.876
		no	0.942	0.940	0.940	0.865	0.863	0.862
	sent	yes	0.915	0.914	0.914	0.895	0.886	0.885
		no	0.829	0.826	0.826	0.834	0.822	0.818
mpnet	sent	yes	**0.979**	**0.979**	**0.978**	0.952	0.950	0.950
		no	0.929	0.929	0.928	0.929	0.928	0.928
	spacy	yes	0.963	0.962	0.962	**0.959**	**0.958**	**0.957**
		no	0.921	0.919	0.919	0.934	0.931	0.931
minilm	sent	yes	0.970	0.969	0.969	0.935	0.933	0.933
		no	0.941	0.938	0.938	0.904	0.893	0.892

Table A.4: Overview of the performances of the unimodal models trained on the textual modality using the validation and test split. The highest values per split and metric for the models without keyword filtering are highlighted in **bold**. The highest values for the models trained with keyword filtering are underscored.

A.5.3 Multimodal Training Results

			Val performance			Test performance		
M1	**M2**	**Fusion**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
w2v	visual	cross	0.920	0.919	0.919	0.923	0.922	0.922
		concat	0.925	0.924	0.923	0.919	0.918	0.917
	audio	cross	0.912	0.907	0.906	0.912	0.908	0.907
		concat	0.909	0.907	0.907	0.923	0.922	0.922
mpnet	visual	cross	0.936	0.933	0.933	0.923	0.919	0.918
		concat	0.941	0.940	0.940	0.933	0.932	0.932
	audio	cross	**0.946**	**0.943**	**0.943**	0.934	0.931	0.931
		concat	0.934	0.933	0.933	**0.938**	**0.937**	**0.937**
	hiddenVAD	cross	0.937	0.936	0.936	0.924	0.922	0.922
		concat	0.935	0.933	0.933	0.934	0.933	0.933

Table A.5: Overview of the performance of the bimodal models trained on a combination of modalities using the validation and test split. The highest values per split and metric for the models are highlighted in **bold**. Abbreviations: M1: the features of the first modality. M2: the features of the second modality.

			Val performance			Test performance		
M1	**M2 + M3**	**Fusion**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
w2v	visual. audio	criss	0.912	0.910	0.909	**0.928**	**0.926**	**0.926**
		layered	0.915	0.912	0.911	0.920	0.919	0.918
		concat	0.920	0.919	0.919	0.926	**0.926**	**0.926**
mpnet	visual. audio	criss	0.927	0.921	0.921	0.916	0.907	0.905
		layered	0.927	0.924	0.923	0.926	0.920	0.919
		concat	0.934	0.933	0.933	0.927	**0.926**	**0.926**
	visual. hiddenVAD	criss	0.932	0.929	0.928	0.914	0.912	0.911
		layered	0.925	0.921	0.921	0.918	0.915	0.915
		concat	**0.948**	**0.948**	**0.948**	0.926	0.925	0.925

Table A.6: Overview of the performance of the trimodal models trained on a combination of modalities using the validation and test split. The highest values per split and metric for the models are highlighted in **bold**. Abbreviations: M1: the features of the first modality. M2 + M3: the features of the second and third modalities.

A.5.4 Bias Mitigation Training Results

		Val performance			Test performance		
Bias mitigation	**Method**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
None	-	0.929	0.929	0.928	0.929	0.928	0.928
Pre: oversampling	-	0.934	0.933	0.933	0.926	0.925	0.925
Pre: MixFeat	Group	0.935	0.933	0.933	0.929	0.927	0.927
	Mixg	**0.939**	**0.938**	**0.938**	0.928	0.927	0.927
	Subgroup	0.920	0.917	0.916	0.931	0.928	0.928
	Synth	0.906	0.905	0.905	0.913	0.912	0.911
	Synth mixg	0.928	0.924	0.923	0.927	0.924	0.923
In: reweighing	-	0.930	0.929	0.928	**0.935**	**0.935**	**0.934**
Post: Equal odds	-	0.929	0.929	0.929	0.927	0.926	0.926

Table A.7: Overview of the performance of the sentence-level bias mitigation methods trained on the textual modality using the validation and test split. The highest values per split and metric for the models are highlighted in **bold**. Abbreviations: Mixg: mixgendered. Synth: synthetic.