Thesis for the Degree of Master of Computer Information System

**Bias Mitigation in Mental Health Sentiment Analysis Using BERT with Fairness Techniques**



**Anup Katuwal**

**(2020-2-92-0022)**

**Nepal College of Information Technology**

**Faculty of Management**

**Pokhara University, Nepal**

**December, 2025**

Thesis for the Degree of Master of Computer Information System

**Bias Mitigation in Mental Health Sentiment Analysis Using BERT with Fairness Techniques**

**Supervised by Prof. Dr. Roshan Chitrakar, Ph.D.**

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Computer Information System

**Anup Katuwal**

**(2020-2-92-0022)**

**Nepal College of Information Technology (NCIT)**

**Faculty of Management**

**Pokhara University, Nepal**

**December, 2025**

ABSTRACT

Sentiment analysis of mental-health discourse is challenged by linguistic ambiguity, severe class imbalance, and the presence of bias-sensitive attributes. This midterm study introduces **RMH-Bias-10K**, a dataset of 10,000 English Reddit posts from mental-health–related subreddits, annotated with three-class sentiment labels using weak supervision and enriched with inferred gender, severity indicators, and dialect categories. Two baseline models are evaluated under matched conditions across five random seeds: **M1**, a fine-tuned BERT classifier, and **M2**, a BiLSTM with GloVe embeddings. On a held-out test set (n = 1,500), M1 achieves higher performance (Macro-F₁ = 0.6851, ROC–AUC = 0.8313) than M2 (Macro-F₁ = 0.6286, ROC–AUC = 0.7916). Class-wise analysis shows that M1 outperforms M2 on negative (F₁ = 0.8083) and positive (F₁ = 0.7105) sentiment, while both models struggle with the sparse neutral class. Fairness evaluation reveals that M1 exhibits larger subgroup disparities despite superior accuracy. For gender, M1 shows higher negative-class disparities (DPD = 0.049; EO-FPR = 0.154) than M2 (DPD = −0.037; EO-FPR = −0.071). Severity-based analysis indicates pronounced disparities for M1 (DPD = 0.462), with similar but attenuated patterns for M2. Dialect-based evaluation further highlights group-level differences for both models. Overall, the results establish a robust baseline and demonstrate a consistent performance–fairness tension in mental-health sentiment analysis, motivating bias-mitigation strategies in subsequent thesis stages.

**Keywords:** sentiment analysis, BERT, BiLSTM, RMH-Bias-10K, DPD, EO-TPR, EO-FPR.

TABLE OF CONTENTS

**Title Page**

Title Page i

Abstract ii

Table of contents iii

List of Tables vii

List of Figures viii

List of Code Snippets ix

List of Abbreviations x

List of Appendices xi

**CHAPTER 1**

INTRODUCTION

1.1. Background of the Study 1

1.2. Statement of Problem 4

1.3. Research Objectives 6

1.4. Significance of the Study 7

1.5. Scope and Delimitations 8

1.5.1. Delimitations 8

1.5.2. Scope 8

1.6. Limitations 9

**CHAPTER 2**

LITERATURE REVIEW

2.1. Overview 11

2.2. Sentiment Analysis Techniques in Natural Language Processing 12

2.2.1. Lexicon based approaches 12

2.2.2. Classical machine learning 13

2.2.3. Deep learning with recurrent networks 14

2.2.4. Transformer Models and BERT 14

2.3. Applications of sentiment analysis in mental health monitoring 15

2.3.1. Social media as a mental health signal 15

2.3.2. Reddit and online mental health communities 16

2.3.3. Surveys and limitations of sentiment labels 16

2.4. Bias in Pre Trained Language Models and Sentiment Analysis 17

2.4.1. Sources and forms of bias 17

2.4.2. Bias in mental health and social media models 18

2.5. Debiasing and Bias Mitigation Strategies in Natural Language Processing 18

2.5.1. Data level techniques 19

2.5.2. Model level techniques 19

2.5.3. Post processing, calibration, and fairness metrics 20

2.6. Research Gaps and Justification for the Study 21

2.7. Summary 22

**CHAPTER 3**

DATASET CONSTRUCTION

3.1. Background 24

3.2. Reddit as a Source of Mental-Health Narratives 25

3.2.1. Subreddit Selection Criteria 26

3.3. Data Extraction Pipeline 26

3.4. Sentiment Annotation 28

3.5. Bias Metadata Layer 29

3.6. Ethical Considerations 30

3.7. Final Dataset Structure 30

3.8. Comparison with Existing Datasets 31

3.9. Summary 33

**CHAPTER 4**

METHODOLOGY

4.1. Overview 33

4.2. Research Design 37

4.3. Model Configurations 38

4.3.1 BERT Architecture (M1, M3) 38

4.3.2 BiLSTM Architecture (M2) 40

4.4. Data Preprocessing and Stratified Splitting 41

4.5. Fairness Metrics 42

4.5.1 Bias Amplification (BA) 43

4.5.2 Demographic Parity Difference (DPD) 43

4.5.3 Equalized Odds (EO) 43

4.6. Debiasing Framework 44

4.6.1 Counterfactual Data Augmentation (CDA) 44

4.6.2 Adversarial Training 45

4.7. Summary 46

**CHAPTER 5**

EMPIRICAL FINDINGS AND ANALTYICAL INTERPRETATION

5.1. Background 53

5.2. Descriptive Characteristics of the Test Dataset 54

5.3. Classification Performance 55

5.3.1 Overall performance comparison 55

5.3.2 Class-wise F₁ scores 55

5.3.3 Confusion matrices 56

5.3.4. Interpretation of Baseline Performace 56

5.4. Fairness and Bias Evaluation 57

5.5. Severity-Level Performance 58

5.6. Dialect-Level Performance 58

5.7. Error Category Analysis 59

5.8. Synthesis of Performance and Bias Behavior in Baseline Models 60

5.9. Computational Efficiency 61

5.10. Practical Implications for Nepal and South Asia 62

5.11. Limitations and Future Work 63

5.12. Summary 64

**CHAPTER 6**

CONCLUSION, DISCUSSION, AND FUTURE WORK

6.1. Overview 69

6.2. Summary of Completed Work 70

6.3. Discussion of Empirical Findings 71

6.4. Key Contributions of the Midterm Work 72

6.5. Limitations of the Current Stage 73

6.6. Directions for Future Work 73

6.7. Validation Strategy for the Proposed Fairness-Aware Model 74

6.8. Conclusion 74

**REFERENCES**

**APPENDIX A**

A.1 Subreddit Sources 86

A.2 Metadata Fields 86

A.3 Example Anonymized Entries 87

**APPENDIX B**

B.1 Raw-to-Clean Transformations 75

B.2 Tokenization Examples 75

B.3 Handling Non-Standard Forms 76

**APPENDIX C**

C.1 Length and Dialect Distributions 77

C.2 Extended Severity-Level Results 77

**APPENDIX D**

D.1 Data Cleaning Function 78

D.2 Example Counterfactual Gender Swap 78

D.3 Simple Training Loop Skeleton (Illustrative) 79

**APPENDIX E**

E.1 Notation 80

E.2 Demographic Parity Difference 80

E.3 Equalized Odds (True Positive Rate and False Positive Rate Gaps) 80

E.4 Bias Amplification 80

E.5 Illustrative Example 81

**APPENDIX F**

F.1 Anonymization and Paraphrasing 82

F.2 Compliance with Platform Policies 82

F.3 Storage and Access Control 82

F.4 Scope and Limitations 82

List of Tables

| **Table No.** | **Title** | **Page No.** |
| --- | --- | --- |
| Table 3.1 | Dataset summary statistics (RMH-Bias-10K) | 26 |
| Table 3.2 | Distribution of posts by subreddit | 28 |
| Table 3.3 | Bias metadata fields | 31 | |
| Table 3.4 | RMH-Bias-10K compared with existing mental health datasets | 32 | |
| Table 4.1 | Model Configuation | 38 |
| Table 4.2 | Stratified Dataset Split | 42 |
| Table 4.3 | Fairnss metrics Summary | 44 |
| Table 4.4 | Counterfactual Data Augmentation (CDA) Examples | 45 |
| Table 4.5 | Training hyperparameters | 47 |
| Table 5.1 | Descriptive statistics of the RMH-Bias-10K test set | 50 |
| Table 5.2 | Comparative performance of M1 (BERT) and M2 (BiLSTM) | 51 |
| Table 5.3 | Class-wise F1 scores on the test set | 52 |
| Table 5.4 | Confusion Matrix for M1 (BERT) | 53 |
| Table 5.5 | Confusion Matrix for M2 (BiLSTM) | 54 |
| Table 5.6 | Gender-based fairness metrics for M1 and M2 | 55 |
| Table 5.7 | Severity-stratified F1 scores and bias amplification | 56 |
| Table 5.8 | Error rates by dialect for M1 and M2 | 57 |
| Table 5.9 | Common Error category taxonomy | 58 |
| Table 5.10 | Approximate computational cost comparision | 60 |

List of Figures

| **Figure No.** | **Title** | **Page No.** |
| --- | --- | --- |
| Fig. 3.1 | Dataset construction workflow | 27 |
| Fig. 3.2 | Distribution of posts across subreddits | 30 |
| Fig. 4.1 | Overall methodology of workflow | 36 |
| Fig. 4.2 | Block diagram of System Architecture. | 38 |
| Fig. 4.3: | BERT architecture with sentiment classification head. | 41 |
| Fig. 4.4 | BiLSTM architecture flow. | 42 |
| Fig. 4.5 | Data preprocessing and splitting Pipeline | 43 |
| Fig. 4.6 | Fairness evaluation workflow | 46 |
| Fig. 4.7 | Counterfactual Data Augmentation (CDA) process pipeline | 47 |
| Fig. 5.1 | Confusion matrices for M1 (BERT) and M2 (BiLSTM) | 54 |
| Fig. 5.2 | Aggregated Confusion matrices across runs | 54 |
| Fig. 5.3 | Learning curves for M1 training steps | 55 |
| Fig. 5.4 | Training loss curves for M2 (BiLSTM) epochs | 56 |

List of Code Listings

| **Listing No.** | **Title / Description** | **Page No.** |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |
| Listing 3.1 | Pushshift bulk extraction script | 26 |
| Listing 3.2 | PRAW validation and deletion filtering | 27 |
| Listing 4.1 | Preprocessing pipeline implementation | 41 |
| Listing 4.2 | BERT fine-tuning script (M1) | 48 |
| Listing 4.3 | BiLSTM training code (M2) | 49 |
| Listing 4.4 | Fairness metric computation functions | 53 |
| Listing 4.5 | Counterfactual data augmentation generator | 45 |
| Listing 4.6 | Illustrative adversarial debiasing loop (M3 skeleton) | 79 |

List of Abbreviations

| **Abbreviation** | **Full Form** |
| --- | --- |
| API | Application Programming Interface |
| BA | Bias Amplification |
| BERT | Bidirectional Encoder Representations from Transformers |
| BiLSTM | Bidirectional Long Short-Term Memory |
| CDA | Counterfactual Data Augmentation |
| DPD | Demographic Parity Difference |
| EO | Equalized Odds |
| F1 | F1 Score (Harmonic mean of Precision and Recall) |
| GPU | Graphics Processing Unit |
| LM | Language Model |
| LMIC | Low and Middle Income Country |
| LSTM | Long Short-Term Memory |
| NLP | Natural Language Processing |
| OOV | Out-of-Vocabulary |
| PRAW | Python Reddit API Wrapper |
| RMH | Reddit Mental Health |
| RMH-Bias-10K | Reddit Mental Health Bias-Aware 10,000 Post Dataset |
| TF-IDF | Term Frequency–Inverse Document Frequency |

**List of Appendices**

| **Appendix** | **Title** | **Page** |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |
| Appendix A | Dataset construction and preprocessing | 86 |
| Appendix B | Subreddit selection and metadata extraction | 88 |
| Appendix C | Sentiment annotation and labeling pipeline |  |
| Appendix D | Model training code and hyperparameters |  |
| Appendix E | Fairness metrics and statistical testing |  |
| Appendix F | Confusion matrices and error analysis |  |
| Appendix G | Ethical considerations and dataset documentation |  |

**CHAPTER 1**

**INTRODUCTION**

**1.1 Background of the Study**

Mental health represents one of Nepal's most pressing public health challenges. According to the Nepal Health Research Council, approximately one in seven adults experiences mental health difficulties [1]. The World Health Organization further reports that mental health conditions contribute substantially to the overall disease burden in low- and middle-income countries [2]. Despite this urgent need, access to professional mental health services remains limited across much of Nepal, with significant geographic and economic barriers preventing many individuals from receiving adequate care [3], [4].

As professional services remain difficult to access, online platforms have emerged as important spaces for mental health discourse. Reddit, in particular, hosts numerous communities where individuals discuss their experiences with depression, anxiety, and other conditions [5]. These communities provide anonymity, peer support, and opportunities for self-expression that many people find valuable. The volume and openness of these discussions make them potentially useful for understanding mental health experiences and developing supportive technologies.

Sentiment analysis is a core task in natural language processing that determines the emotional tone expressed in text, typically classified as positive, negative, or neutral. Early research relied on lexicon-based techniques, where dictionaries assigned fixed sentiment scores to individual words [6], [7]. While straightforward to implement, these methods struggled with contextual factors like negation, sarcasm, and domain-specific language use.

Supervised machine learning methods later improved performance by learning patterns from annotated data. Algorithms like logistic regression and support vector machines used statistical features such as n-grams and TF-IDF representations to capture sentiment more flexibly [25]. Deep learning architectures, particularly Long Short-Term Memory networks, advanced the field further by modeling sequential dependencies and capturing longer contextual spans in text [8].

Transformer architectures marked a major shift in natural language processing by replacing recurrent processing with self-attention mechanisms. This design allows models to consider all tokens in a sequence simultaneously, enabling efficient modeling of long-range dependencies while supporting parallel computation [9].

Building on this foundation, Bidirectional Encoder Representations from Transformers (BERT) introduced large-scale pre-training strategies that acquire general linguistic knowledge from diverse text sources before fine-tuning for specific tasks [10]. Unlike earlier models that processed text in one direction, BERT encodes contextual information from both preceding and following words at once. This bidirectional representation proves particularly valuable for sentiment analysis, where interpreting emotionally charged language often depends on subtle interactions between surrounding terms.

Applying transformer models to mental health discussions on social media presents challenges not found in standard benchmark datasets. Reddit posts frequently blend clinical terminology with informal language, including slang, humor, abbreviations, and culturally specific expressions [14], [17]. Emotional distress is often communicated indirectly through metaphor, understatement, or narrative context rather than explicit sentiment markers. These linguistic characteristics can be difficult for models primarily trained on formal text to interpret reliably.

Mental health datasets also exhibit pronounced class imbalance, as expressions of negative sentiment or distress tend to occur more frequently than neutral or positive posts. This imbalance can bias model predictions and obscure errors affecting less frequent classes if not addressed during training and evaluation.

Beyond predictive accuracy, growing attention has focused on bias in pre-trained language models. Empirical studies show that models like BERT can encode demographic associations from their training data, leading to systematic differences in predictions across gender, race, or dialectal groups [11], [18]. Even minimal changes to identity-related terms—such as altering a pronoun or personal name—may result in different sentiment predictions despite identical semantic content [11], [37].

These inconsistencies raise important fairness concerns, particularly in mental health applications where erroneous or biased interpretations of distress could have serious consequences. Ensuring that sentiment predictions are driven by expressed emotional content rather than demographic cues is critical for responsible deployment in sensitive domains.

These fairness concerns are especially relevant in contexts like Nepal, where mental health resources are scarce and digital tools may increasingly supplement traditional services. Mobile applications, chatbots, and tele-counseling platforms are being explored as potential ways to extend mental health support [12], [13], [17]. If sentiment analysis systems deployed in these applications produce biased or unreliable results for certain demographic groups or linguistic patterns, they risk reinforcing existing inequalities rather than reducing them.

This thesis investigates these issues by evaluating both the performance and fairness of BERT and Bidirectional LSTM (BiLSTM) models on mental health-related Reddit posts. The study addresses class imbalance through undersampling and examines baseline models trained with class weights and cross-entropy loss. It then proposes a fairness-aware BERT model that integrates counterfactual data augmentation, adversarial training, and focal loss. By comparing model performance across different demographic groups, crisis severity levels, dialectal variations, and mental health communities, this work aims to contribute to the development of more equitable sentiment analysis systems for mental health applications.

**1.2 Statement of the Problem**

While BERT demonstrates high accuracy on standard sentiment analysis benchmarks, high overall accuracy does not guarantee reliable or fair performance when models are applied to specialized domains like mental health [14], [15]. Biases present in training data can lead to systematic differences in how models treat different demographic groups or linguistic styles, even when overall metrics appear satisfactory [11], [18].

Mental health posts on Reddit differ substantially from conventional sentiment analysis datasets. Expressions of distress are often indirect, mixing conflicting emotions or using metaphorical language [5], [14]. Informal elements like slang, abbreviations, emojis, and code-switching are common [40], [41]. Posts may signal crisis severity through subtle linguistic cues rather than explicit statements [39]. Standard sentiment analysis datasets typically lack annotations for these characteristics, making it difficult to assess model behavior on sensitive content. Additionally, class imbalance is prevalent in mental health datasets, with negative sentiment posts often significantly outnumbering non-negative ones [34].

High overall accuracy can mask performance disparities across subgroups [27], [32]. A model might achieve 85% accuracy overall while performing notably worse for posts with certain gender cues, crisis levels, or dialectal features. Without explicit fairness evaluation, such disparities may go undetected. This is particularly concerning in Nepal and similar contexts where cultural stigma around mental health already creates barriers to care [1], [3], [43]. If automated systems misinterpret or overlook distress signals from vulnerable populations, they could worsen existing inequalities.

This study addresses these concerns by systematically evaluating bias in transformer-based sentiment models for mental health text. The research examines fairness across four dimensions: inferred gender (based on linguistic cues), crisis severity (ranging from mild distress to acute crisis), dialectal variation (reflecting different levels of informal language use), and subreddit affiliation (representing different mental health conditions) [19], [27], [28]. By documenting these patterns and exploring mitigation approaches, this work aims to advance understanding of how to build fairer sentiment analysis systems for sensitive applications.

**1.3 Research Objectives**

General Objective

To investigate bias and fairness issues in sentiment analysis models applied to mental health text and develop a bias-aware framework for binary sentiment classification.

Specific Objectives

1. To construct a dataset of 10,000 Reddit posts from mental health communities, labeled for sentiment (negative or non-negative) and annotated with metadata for gender cues, crisis severity indicators, dialectal features, and subreddit categories.

2. To evaluate the classification accuracy and performance of BERT and BiLSTM models using standard metrics including precision, recall, and F1-score.

3. To assess fairness across user attributes (inferred gender, crisis severity, dialect, and subreddit affiliation) using established fairness metrics including demographic parity difference and equalized odds.

4. To develop and evaluate a fairness-aware model that addresses class imbalance through undersampling and combines counterfactual data augmentation, adversarial training, and focal loss, comparing its performance and fairness against baseline models using class weights and cross-entropy loss.

**1.4 Significance of the Study**

This study holds value on theoretical, practical, and societal fronts.

Theoretically, it contributes to NLP research by integrating fairness as a critical performance metric in sentiment analysis, especially for mental health applications [27], [32]. It offers comparative insights into how BERT and BiLSTM models behave across demographic dimensions, including gender, crisis severity, dialectal variations, and different mental health conditions represented by subreddit communities [18], [28].

Practically, the study delivers an accessible, reproducible framework that relies on open-source tools and free platforms [29]. This makes it ideal for researchers in low-resource settings. By embedding fairness evaluation in the model development pipeline, it encourages more responsible AI practices.

Societally, the work is especially important for Nepal and similar regions with limited mental health infrastructure [1], [3]. Unfair sentiment analysis tools may misinterpret or overlook distress in vulnerable populations [38], [42]. This study supports the development of equitable digital mental health interventions that can enhance access and support.

**1.5 Scope and Limitations**

This research focuses on English-language Reddit posts from mental health-related subreddits, including communities for depression, anxiety, post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorder (ADHD), bipolar disorder, schizophrenia, and general support forums [5], [16]. The study examines binary sentiment classification (negative versus non-negative) rather than more granular emotion or symptom detection. This binary framing aligns with many practical applications where identifying negative sentiment serves as an initial screening step [12], [15].

The fairness analysis centers on four factors: inferred gender (based on linguistic patterns rather than verified user identity), crisis severity (determined through keyword analysis and post characteristics), dialect (measured by the presence of informal language features and non-standard linguistic patterns), and subreddit affiliation (representing different mental health conditions such as depression, anxiety, PTSD, and others) [5], [16], [19], [39], [40], [41]. These factors were selected based on their relevance to mental health equity concerns and their feasibility for automated annotation. Gender is inferred using pronouns, names, and self-references rather than direct user disclosure. Crisis severity is assessed through lexical indicators and contextual features without clinical diagnosis. Dialect captures variations in language use, including slang, abbreviations, code-switching, and non-standard grammar that may reflect regional, cultural, or socioeconomic differences. Subreddit categories allow examination of whether models perform differently across various mental health conditions.

Several important delimitations and limitations must be acknowledged. Reddit data, while valuable, may not fully represent offline populations or individuals who express distress differently across various platforms or in clinical settings [5], [17]. Human annotation inevitably introduces some degree of subjective judgment, particularly for ambiguous posts [30], [31]. The binary sentiment framework simplifies the complexity of emotional expression found in real mental health discourse [15]. Additionally, this work does not address real-time deployment considerations, clinical validation, or integration with existing mental health services. These important topics merit separate investigation and fall outside the current study's scope.

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1 Overview**

Sentiment analysis has evolved substantially over the past two decades. Early approaches relied primarily on lexicon-based methods [6], [7], which were later supplemented by classical machine learning techniques [25]. Deep learning, particularly through recurrent neural networks, enabled more sophisticated modeling of sequential dependencies [8]. Most recently, transformer-based models have achieved strong results across a wide range of NLP tasks [9], [10].

Parallel to these technical advances, researchers have increasingly applied sentiment analysis to mental health content from social media platforms. Studies using Twitter and Reddit data have demonstrated that linguistic patterns can reflect psychological states, supporting the development of early intervention tools [12], [13]. However, this work has also revealed persistent challenges around linguistic ambiguity, class imbalance, and the reliability of automatically assigned sentiment labels [14], [34].

More recently, evidence has emerged that pre-trained language models are not neutral instruments. These models can inherit and sometimes amplify biases related to gender, race, dialect, and other attributes from their training data [11], [14], [15], [18]. Such biases pose particular risks in sensitive applications like mental health monitoring, where unfair predictions could harm vulnerable individuals [38], [42].

This chapter reviews the development of sentiment analysis methods, their application to mental health monitoring, documented sources of bias in language models, and current debiasing strategies. It concludes by identifying research gaps that motivate the present study.

**2.2 Sentiment Analysis Techniques in Natural Language Processing**

**2.2.1 Lexicon-Based Approaches**

Lexicon-based methods represent the earliest systematic approach to sentiment analysis. These methods rely on dictionaries that assign polarity scores to individual words. SentiWordNet, for example, provides sentiment scores derived from WordNet synsets [6]. VADER extends this approach for social media text by incorporating rules for handling intensifiers, negations, and emoticons [7]. While computationally efficient and interpretable, lexicon-based approaches struggle with context-dependent meanings, sarcasm, and domain-specific language. They also require substantial manual effort to create and maintain comprehensive lexicons.

**2.2.2 Classical Machine Learning**

Classical machine learning methods improved upon lexicon approaches by learning sentiment patterns from labeled training data. Techniques such as Naive Bayes, logistic regression, and support vector machines use features like n-grams, term frequency-inverse document frequency (TF-IDF) vectors, and part-of-speech tags [25]. These methods can capture some contextual information and domain-specific patterns that lexicons miss. However, they still rely on hand-crafted features and struggle with long-range dependencies and complex linguistic structures.

**2.2.3 Deep Learning with Recurrent Networks**

Deep learning approaches, particularly recurrent neural networks (RNNs), enabled models to learn feature representations automatically. Long Short-Term Memory (LSTM) networks and their bidirectional variants (BiLSTM) can capture sequential dependencies and temporal patterns in text [8]. These architectures have proven effective for sentiment analysis by maintaining information across longer sequences and learning complex patterns from raw text. However, RNNs face challenges with very long sequences and can be computationally intensive to train.

**2.2.4 Transformer Models and BERT**

The transformer architecture introduced a new paradigm based on self-attention mechanisms rather than recurrence [9]. Transformers can process entire sequences in parallel and model long-range dependencies more effectively than RNNs. BERT builds on this architecture through large-scale pre-training on diverse text corpora using a masked language modeling objective [10]. By processing text bidirectionally and learning rich contextual representations, BERT achieves strong performance across numerous NLP tasks. However, applying BERT to mental health posts from Reddit presents challenges due to the informal, emotionally complex, and context-dependent nature of this content [14], [17]. Generic pre-training may not fully capture the nuances of mental health discourse.

**2.3 Applications of Sentiment Analysis in Mental Health Monitoring**

**2.3.1 Social Media as a Mental Health Signal**

Social media platforms have become valuable sources for studying mental health. Early work on Twitter demonstrated that linguistic and behavioral features could distinguish users reporting depression or PTSD from control groups [12], [13]. Features like negative emotion words, first-person pronouns, and temporal patterns of activity showed promise as indicators of psychological distress. However, these studies also revealed challenges around self-selection bias, the difficulty of obtaining ground truth labels, and ethical concerns about privacy and consent [33].

**2.3.2 Reddit and Online Mental Health Communities**

Reddit has emerged as a particularly rich source for mental health research. Condition-specific communities like r/depression, r/Anxiety, and r/PTSD provide spaces for anonymous self-disclosure and peer support [5], [16]. Researchers have used Reddit data to study linguistic markers of suicidal ideation, model the progression of mental health conditions, and develop early warning systems [16], [17], [20]. The platform's anonymity and community norms around open sharing make it a valuable research resource, though questions remain about how well online behavior generalizes to offline populations.

**2.3.3 Current Challenges in Mental Health Sentiment Analysis**

Despite promising results, applying sentiment analysis to mental health content faces several persistent challenges. Mental health posts often contain sarcasm, mixed emotions, and rapid mood shifts [14], [39]. Distress may be expressed indirectly through metaphors or understatement. Class imbalance is common, with negative posts often outnumbering positive ones [34]. Obtaining reliable ground truth labels is difficult, as self-reported diagnoses may not align with clinical assessments [16], [33]. Additionally, models trained on general sentiment datasets may not transfer well to the specialized language of mental health communities [24].

**2.4 Bias in Pre-Trained Language Models and Sentiment Analysis**

**2.4.1 Sources and Forms of Bias**

Pre-trained language models are typically trained on large text collections from sources like books, Wikipedia, and web crawls [9], [10], [29]. These corpora inevitably reflect existing social biases, stereotypes, and power imbalances. As models learn statistical patterns from this data, they can internalize these biases. Research has documented various forms of bias in word embeddings and contextualized models, including gender stereotypes, racial biases, and associations between demographic groups and negative attributes [11], [18], [37], [38].

**2.4.2 Bias in Mental Health and Social Media Models**

Evidence suggests that small changes to identity-related terms can significantly affect model predictions. Switching gendered pronouns or names in otherwise identical sentences can change sentiment or emotion classifications [11], [14], [15]. This sensitivity indicates that models may be learning spurious correlations between demographic attributes and outcomes rather than the actual content meaning. In mental health applications, where posts may mention gender, age, or cultural identity, such biases could lead to systematic errors [18], [45]. Social media datasets often exhibit demographic skew and self-selection effects, potentially amplifying these problems [2], [15], [41].

**2.5 Debiasing and Bias Mitigation Strategies in Natural Language Processing**

**2.5.1 Data-Level Techniques**

Data-level mitigation strategies modify the training data to reduce bias. Counterfactual Data Augmentation (CDA) creates paired examples that differ only in protected attributes such as gender [19], [21], [22]. For example, a sentence with male pronouns might be duplicated with female pronouns while keeping all other content identical. This technique helps models learn that predictions should not depend on these attributes [19]. However, CDA requires careful implementation to avoid introducing unnatural language patterns or altering content meaning in subtle ways [44].

**2.5.2 Model-Level Techniques**

Model-level approaches modify the training process itself. Adversarial debiasing trains an auxiliary adversary network to predict protected attributes from the model's internal representations [26]. The main model is then trained to simultaneously maximize task performance while minimizing the adversary's ability to predict sensitive attributes. This approach encourages the model to learn representations that are invariant to protected attributes. Focal loss provides another model-level strategy by modifying the loss function to focus more on hard or minority examples [24], [27]. This is particularly useful for handling class imbalance while potentially reducing disparities across groups, as it down-weights easy examples and emphasizes difficult ones during training.

**2.5.3 Post-Processing and Fairness Metrics**

Post-processing methods operate on model outputs after training. Threshold calibration adjusts decision thresholds separately for different groups to equalize error rates [27]. This can improve fairness metrics like equalized odds without retraining. However, post-processing assumes group membership is known at test time and may reduce overall accuracy. Common fairness metrics include demographic parity (equal positive prediction rates across groups), equalized odds (equal true positive and false positive rates), and bias amplification (changes in the correlation between protected attributes and outcomes from data to predictions) [27], [28], [32].

**2.6 Research Gaps and Justification for the Study**

Despite significant progress in sentiment analysis and fairness-aware modeling, important limitations remain when these approaches are applied to mental health discourse on social media. Many widely used datasets lack explicit annotations that enable systematic fairness evaluation across demographic, linguistic, and contextual dimensions [2], [14], [18]. Most studies emphasize aggregate performance metrics while overlooking subgroup-level disparities that may disproportionately affect vulnerable populations.

Furthermore, bias mitigation techniques such as counterfactual data augmentation and adversarial debiasing have rarely been explored in combination within mental health sentiment analysis. When fairness is considered, it is often restricted to a single attribute, such as gender, leaving other relevant factors—such as crisis severity, informal language use, or community context—largely unexplored. These gaps highlight the need for datasets and modeling frameworks specifically designed to evaluate and mitigate bias in mental health-related sentiment classification.

First, while existing datasets like the UMD Reddit Suicidality Dataset and CLPsych shared task data have advanced mental health NLP research [16], [20], few datasets explicitly include metadata for fairness evaluation across gender, crisis severity, dialectal features, and different mental health condition categories [2], [14], [18]. Most studies report overall performance metrics without detailed subgroup analysis. This gap motivates the first objective of this study: constructing a dataset with explicit annotations for bias-relevant attributes.

Second, explicit bias mitigation techniques like counterfactual augmentation or adversarial debiasing are seldom applied to mental health text from Reddit [19], [26]. When fairness is addressed, it typically focuses on a single dimension such as gender, neglecting other important factors like crisis severity or dialect. Systematic comparisons between baseline and debiased models remain scarce. This gap supports the second and third objectives: evaluating baseline model performance and assessing fairness across multiple user attributes using established metrics.

Third, low- and middle-income country contexts, including Nepal, receive limited attention in current research. These settings face distinct challenges: higher mental health stigma, more frequent code-mixing between languages, and severely constrained clinical resources [1], [3], [43]. Understanding how bias manifests in these contexts is important for ensuring equitable access to digital mental health tools. This consideration informs the broader societal motivation for this work.

Fourth, transformer-based models are widely adopted as baselines without established fairness benchmarks or routine mitigation procedures [10], [32]. This limits their safe deployment in sensitive applications where biased predictions could cause real harm [1], [3], [42]. This gap motivates the fourth objective: developing a fairness-aware model that addresses class imbalance through undersampling and systematically combines multiple bias mitigation techniques—counterfactual data augmentation at the data level, adversarial training to learn fair representations, and focal loss to handle remaining class imbalance while improving fairness—and comparing it against baseline approaches using class weights and cross-entropy loss.

This study addresses these gaps by constructing the RMH-Bias-10K dataset with explicit bias-relevant metadata, applying undersampling to address class imbalance, evaluating BERT and BiLSTM baseline models with class weights and cross-entropy loss across multiple demographic and linguistic subgroups, and developing a fairness-aware BERT model (M3) that integrates counterfactual data augmentation, adversarial training, and focal loss. The use of fairness metrics like demographic parity difference and equalized odds provides quantitative measures of model equity that complement traditional performance metrics.

**2.7 Summary**

Transformer-based models have substantially improved sentiment analysis performance across many tasks. However, they also introduce fairness risks when applied to mental health contexts. This chapter reviewed the evolution of sentiment analysis methods, from early lexicon-based approaches to modern transformer architectures, and their application to mental health monitoring on social media platforms. The review highlighted documented sources of bias in pre-trained language models and current mitigation strategies, including data-level techniques like counterfactual augmentation, model-level approaches like adversarial training, and post-processing methods like threshold calibration.

The identified research gaps—particularly the lack of datasets with explicit fairness metadata, limited application of bias mitigation to mental health text, insufficient subgroup analysis across multiple dimensions, and underrepresentation of low-resource contexts—directly motivate the four specific objectives of this study. Addressing these gaps requires a dataset purpose-built for fairness evaluation, which is the focus of the next chapter. Chapter 3 describes the construction of the RMH-Bias-10K dataset, including data collection from Reddit mental health communities, sentiment annotation procedures, and the creation of bias metadata layers for gender, crisis severity, dialectal variation, and subreddit categories.

**CHAPTER 3**

**DATASET CONSTRUCTION**

**3.1 Introduction**

Mental health sentiment analysis requires datasets that reflect how people actually talk about their struggles online. Existing datasets like CLPsych 2019 and the UMD Reddit Suicidality Dataset have been useful [16], [20], but they come with limitations. Most contain relatively small sample sizes around 1,000 to 1,200 posts. They were collected before 2020. More importantly, they lack the systematic metadata needed to study fairness issues across different groups.

We created RMH-Bias-10K to address these gaps. The dataset contains 10,000 Reddit posts from mental health communities collected between 2020 and 2025. Each post was manually labeled for sentiment and enriched with metadata that enables systematic bias evaluation. The entire collection and annotation process can be reproduced on free platforms like Google Colab. This makes it accessible to researchers with limited resources.

This chapter describes how we built the dataset. We walk through selecting communities, collecting and cleaning posts, labeling sentiment, creating fairness metadata, and addressing ethical concerns.

**3.2 Reddit as a Source of Mental Health Narratives**

Reddit works well for mental health research for several practical reasons. Users can post anonymously, which encourages openness about sensitive topics [5]. Communities organize around specific themes, making it easy to find relevant discussions. Posts on Reddit tend to be longer and more detailed than tweets or Facebook updates. This gives us richer text to analyze [23].

We selected ten mental health subreddits using three criteria. They had to focus on mental health topics. They needed to maintain steady posting activity from 2020 to 2025. And they had to foster environments where users openly share personal experiences. Table 3.1 shows the selected communities.

Table 3.1: Selected Subreddit Communities

|  |  |  |
| --- | --- | --- |
| **Subreddit** | **Focus** | **Activity** |
| r/depression | Depression experiences | High |
| r/Anxiety | Anxiety disorders | High |
| r/mentalhealth | General discussions | High |
| r/SuicideWatch | Crisis support | Medium |
| r/BipolarReddit | Bipolar disorder | Medium |
| r/PTSD | Post-traumatic stress | Medium |
| r/OCD | Obsessive-compulsive disorder | Medium |
| r/schizophrenia | Schizophrenia experiences | Medium |
| r/BPD | Borderline personality | Medium |
| r/ADHD | Attention disorders | High |

The selected communities represent diverse mental health conditions. These include depression, anxiety disorders, general mental health discussions, acute crisis support, bipolar disorder, PTSD, OCD, schizophrenia, borderline personality disorder, and ADHD. This diversity lets us examine whether models perform differently across various mental health condition categories.

**3.3 Data Collection and Cleaning Pipeline**

We collected data in three stages using free tools. These included the Pushshift Reddit API and Python Reddit API Wrapper (PRAW) [23].

**3.3.1 Stage 1: Bulk Collection**

The initial collection gathered all publicly available posts from the ten selected subreddits posted between January 2020 and December 2025. For each post, we extracted basic identifying information. This included the post ID and timestamp. We also grabbed the subreddit name, post title, and body text. Additional fields included the score (upvotes minus downvotes), comment count, and NSFW flag. Usernames were immediately removed during collection to protect user privacy.

This bulk collection phase resulted in approximately 14,200 posts across all ten communities. The varied collection approach captured posts from different time periods and popularity levels within each subreddit.

**3.3.2 Stage 2: Validation**

Each collected post was cross-referenced against Reddit's live API to verify availability and filter inappropriate content. This validation step removed several categories of posts. We filtered out posts that had been deleted by users or moderators. Automated moderator announcements were removed. Spam content was eliminated. Posts flagged as NSFW that might not be suitable for sentiment analysis research were excluded.

The validation process helped ensure dataset quality. It removed non-genuine mental health discussions and focused the dataset on authentic user-generated content about personal mental health experiences.

**3.3.3 Stage 3: Text Processing and Cleaning**

Post titles and body text were combined into a single text field to capture complete semantic content. The cleaning pipeline applied several transformations. It removed URLs to eliminate external link noise. It normalized whitespace to standardize formatting. It filtered posts based on length constraints.

Posts shorter than 30 characters were excluded. These typically consisted of single-word responses or emoticons without substantial content. Posts exceeding 5,000 characters were also filtered out. These often contained copy-pasted external content rather than original user writing. Exact duplicates were identified and removed to prevent data leakage across training, validation, and test splits.

After the complete cleaning pipeline, 9,431 posts remained. To reach the target size of 10,000 posts, we collected an additional 569 posts from the same subreddits using identical procedures. This brought the final dataset to exactly 10,000 posts. Table 3.2 summarizes key statistics from the collection process.

Table 3.2: Dataset Collection Statistics

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Posts after validation | ~14,200 |
| Posts after cleaning | 9,431 |
| Additional posts collected | 569 |
| Final dataset size | 10,000 |
| Average post length | 142 words |
| Positive to Negative ratio | 11.9:88.1 |

The final dataset comprises 10,000 posts with an average length of 142 words per post. The sentiment distribution reflects a 12:88 ratio of positive to negative posts. This imbalance is consistent with the prevalence of distress-focused discussions in mental health communities.

**3.4 Sentiment Annotation Methodology**

Automated sentiment analysis tools often misinterpret mental health language. They miss context, sarcasm, and the way common words carry different meanings in distress contexts [14], [39]. For example, the phrase "I'm fine" frequently signals emotional difficulty rather than wellness in depression-related discourse. Given these challenges, we chose manual annotation as the labeling approach.

**3.4.1 Annotation Guidelines**

We developed clear annotation guidelines to ensure consistency. Posts were classified into two categories based on their overall emotional tone and content.

Negative posts express distress, hopelessness, crisis states, or predominantly negative emotions. This category includes descriptions of symptoms, expressions of suffering, and discussions of deteriorating mental health. Examples include posts describing panic attacks, feeling overwhelmed by depression, struggling with suicidal thoughts, or describing relationship breakdowns due to mental health issues.

Positive posts demonstrate hope, recovery progress, supportive interactions, or predominantly positive emotions. This category encompasses encouragement, coping strategies, improvement narratives, and expressions of resilience. Examples include posts celebrating small victories in recovery, offering support to others, sharing successful coping techniques, or expressing gratitude for progress made.

The author manually annotated all 10,000 posts following these guidelines. This process took approximately three months of focused work. Each post was read carefully and labeled based on its dominant emotional content. Posts with mixed sentiment were classified based on their overall tone and the primary message being conveyed.

**3.4.2 Quality Validation Process**

To verify annotation quality, we conducted an independent validation check. We randomly selected 500 posts from the complete annotated dataset. These 500 posts were sent to a licensed mental health counselor for independent labeling. The counselor received the same annotation guidelines and labeled all 500 posts without seeing the author's original labels.

The counselor brought professional expertise in recognizing mental health symptoms and emotional states. This made them an ideal validator for ensuring annotation accuracy. They had over five years of clinical experience working with patients with depression, anxiety, and other mental health conditions represented in the dataset.

**3.4.3 Inter-Annotator Agreement Analysis**

Agreement analysis on the 500 validation posts revealed strong consistency between annotators. Out of 500 posts, 483 received matching labels from both the author and the counselor. This yields 96.6% raw agreement. Only 17 posts (3.4%) received different labels between the two annotators.

We calculated Cohen's kappa to measure agreement while accounting for chance [30], [31]. The kappa value was 0.001, which appears very low despite the high raw agreement. This apparent contradiction occurs because Cohen's kappa becomes unreliable when class distributions are severely imbalanced. The metric's expected agreement calculation assumes more balanced distributions between categories.

In the 500-post validation sample, negative posts heavily outnumbered positive posts, mirroring the overall dataset imbalance. When one class dominates this heavily, Cohen's kappa produces misleadingly low values even when actual agreement is high. This is a well-known limitation of the kappa statistic [30], [31].

To address this limitation, we calculated the Prevalence-Adjusted Bias-Adjusted Kappa (PABAK). PABAK corrects for class imbalance and prevalence effects [30], [31]. It provides a more appropriate reliability estimate for imbalanced datasets. The PABAK value was 0.932, which indicates almost perfect agreement between annotators. Table 3.3 presents the complete reliability metrics.

Table 3.3: Inter-Annotator Agreement Metrics

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Interpretation** |
| Validation sample size | 500 posts | N/A |
| Matching labels | 483 | N/A |
| Disagreements | 17 | N/A |
| Raw Agreement | 96.6% | Excellent |
| Cohen's Kappa | 0.001 | Low (unreliable) |
| PABAK | 0.932 | Almost perfect |

The 17 disagreements occurred primarily in cases where emotional tone was genuinely ambiguous or posts contained mixed sentiment. We reviewed these 17 posts in detail. Some expressed both hope and struggle in equal measure. Others used sarcasm or dark humor that could be interpreted multiple ways. A few discussed recovery progress while acknowledging ongoing challenges.

For the final dataset, we kept the author's original labels for all posts, including the 17 with disagreements. This decision maintains consistency since the author labeled all 10,000 posts using a single consistent interpretation framework. The high agreement rate (96.6%) and excellent PABAK score (0.932) confirm that the annotation process maintained reliable quality throughout.

**3.5 Fairness Metadata Construction**

Since Reddit users don't provide verified demographic information, we constructed proxy attributes to enable fairness evaluation. These attributes were selected based on their relevance to mental health equity and documented sources of bias in NLP models [11], [18], [37], [40], [41].

**3.5.1 Inferred Gender**

Gender was inferred through pattern matching on gendered language elements. We looked for pronouns like he/she, him/her, and his/her. We searched for gendered nouns like man/woman, boyfriend/girlfriend, and husband/wife. We also identified explicit self-references such as "I am a woman" or "as a man" [11], [37].

Posts containing clear gender indicators were labeled as male-cued or female-cued. The majority of posts (94%) lacked clear gender cues. These were classified as unknown. This approach acknowledges that inferred gender based on linguistic patterns is imperfect. It may not reflect actual user identity. However, it can help detect gender-related bias patterns in model predictions.

**3.5.2 Crisis Severity**

Crisis severity reflects the urgency of distress expressed in each post. We categorized posts into four levels based on keyword presence and contextual indicators [39].

None/Low severity (63% of posts) covers general discussions without acute distress markers. These posts might discuss ongoing challenges or seek advice but don't indicate immediate crisis. Moderate severity (25%) applies to posts showing clear distress without immediate danger signals. These often describe struggling to cope or feeling overwhelmed but maintain some stability. Severe (8%) marks expressions of intense suffering or self-harm thoughts. These posts indicate serious distress that warrants concern. Crisis (4%) identifies posts indicating immediate danger or suicidal ideation. Posts from r/SuicideWatch were automatically flagged for careful severity assessment.

**3.5.3 Dialect**

Dialect captures variation in language formality and style. These variations may correlate with demographic, regional, or socioeconomic factors [40], [41]. We computed an informality score based on several features. These included emoji usage, internet slang, text-speak abbreviations, non-standard punctuation patterns, omitted capitalization, and repeated characters for emphasis.

Posts were categorized as Formal (42%) or Informal (58%) based on their informality scores. Formal posts used standard grammar, proper capitalization, and minimal slang. Informal posts showed frequent casual features like "lol", "tbh", emoji strings, or non-standard punctuation like "!!!!!". This binary classification provides a manageable proxy for dialectal variation without requiring detailed sociolinguistic annotation.

**3.5.4 Subreddit Affiliation**

Each post retains its source subreddit label. This enables analysis of model performance across mental health condition categories. The ten communities were approximately balanced at roughly 1,000 posts each. This prevents any single community from dominating the dataset. Subreddit affiliation serves as a straightforward categorical variable for examining whether models show differential performance across condition-specific discourse patterns.

**3.6 Counterfactual Data Augmentation**

To support bias mitigation during model training, we applied counterfactual data augmentation (CDA) to posts containing gender-specific language [19], [21], [22]. CDA creates paired examples that differ only in protected attributes. This helps models learn that predictions shouldn't depend on demographic cues.

The augmentation process systematically swapped gendered terms while preserving semantic content and sentiment. Examples include transforming "he felt alone" to "she felt alone", "my boyfriend left" to "my girlfriend left", and "the guy said" to "the girl said". Only posts where term substitution produced meaningful changes were kept as augmented versions. Posts without gender-specific language remained unchanged.

Each augmented post received a flag (CDA\_flag) to distinguish it from original posts. This lets us analyze augmentation effects. The process generated 419 additional training examples. These augmented posts were added to the training set to provide models with gender-balanced examples that reinforce sentiment patterns independent of gender references.

**3.7 Dataset Splitting Strategy**

The 10,000 annotated posts were partitioned into training (7,000 posts, 70%), validation (1,500 posts, 15%), and test (1,500 posts, 15%) sets. We used stratified random sampling [30]. Stratification maintained proportional representation of sentiment labels and subreddit categories across all three splits. This prevents distribution shifts between training and evaluation sets.

To address class imbalance (88% negative, 12% positive), we applied undersampling to the training set. The majority class (negative posts) was randomly sampled down to match the minority class size. This resulted in a balanced training set of 832 negative and 832 positive posts [34]. After applying counterfactual data augmentation to this balanced training set, the final training distribution had 1,096 positive examples and 987 negative examples. The validation and test sets kept their original imbalanced distributions to reflect realistic deployment conditions.

**3.8 Ethical Considerations**

Mental health data requires careful ethical handling to protect privacy and prevent harm. All data collection followed established practices for computational social science research using public social media content.

We only collected publicly available posts from open communities. No private subreddits, direct messages, or restricted content were accessed. All usernames and personally identifying information were removed during the initial collection stage. When presenting examples in research outputs, posts were paraphrased or modified to prevent re-identification of original authors.

The dataset is designated for academic research purposes only. Commercial deployment or clinical application of models trained on this data would require additional ethical review. This is especially important regarding informed consent and potential impact on vulnerable populations. The data collection process complied with Reddit's API terms of service and broader ethical guidelines for social media research.

**3.9 Final Dataset Structure and Accessibility**

The RMH-Bias-10K dataset is stored in CSV format. It includes these fields: post\_id (unique identifier), subreddit (source community), created\_utc (timestamp), clean\_text (processed post content), sentiment\_label (0 for negative, 1 for positive), inferred\_gender (male/female/unknown), severity\_crisis (none/low/moderate/severe/crisis), dialect (formal/informal), and cda\_flag (0 for original, 1 for augmented).

All code for data collection, cleaning, annotation processing, and counterfactual augmentation is provided in a documented Google Colab notebook (see Appendix A). This notebook allows complete reproduction of the dataset construction pipeline using only free computing resources. This makes the methodology accessible to researchers regardless of institutional funding or computational infrastructure.

**3.10 Summary**

This chapter described construction of the RMH-Bias-10K dataset. The dataset comprises 10,000 manually annotated Reddit posts from ten mental health communities collected between 2020 and 2025. It includes binary sentiment labels verified through independent validation. Out of 500 validation posts, 483 matched between the author and a licensed mental health counselor (96.6% agreement, PABAK = 0.932). The dataset provides explicit fairness metadata across four dimensions: inferred gender, crisis severity, dialect, and subreddit affiliation.

The dataset addresses key limitations of existing mental health NLP resources. It provides systematic fairness annotations. It reflects recent discourse patterns. It incorporates counterfactual augmentation to support debiasing research. It offers complete reproducibility through free computing platforms. The next chapter describes the methodology for evaluating BERT and BiLSTM models on this dataset. This includes baseline training procedures, fairness metrics, and the proposed bias mitigation framework combining adversarial training and focal loss.

**CHAPTER 4**

**METHODOLOGY**

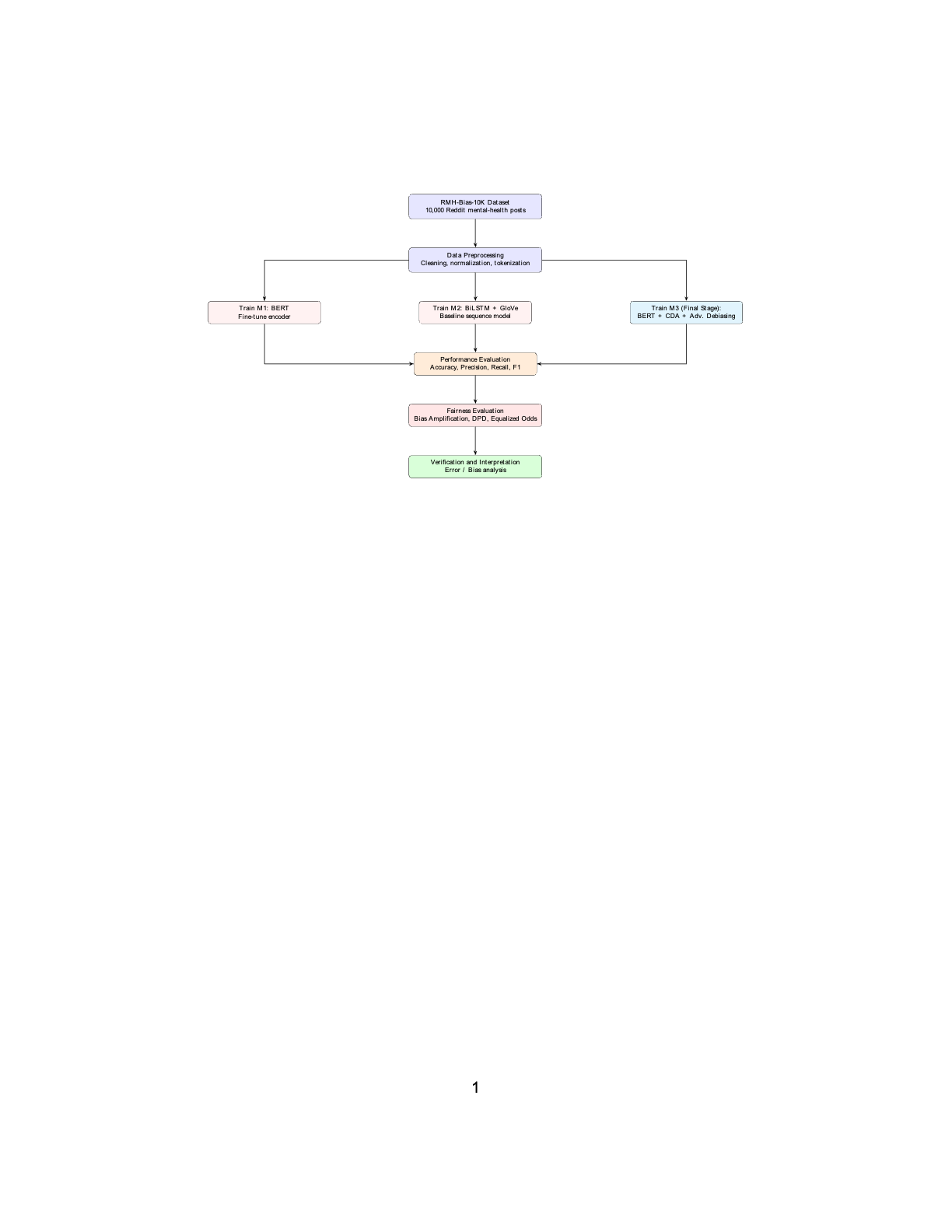
**4.1 Overview**

This chapter explains how we tested whether adding fairness techniques to sentiment classifiers improves their treatment of different groups while maintaining good performance. We compared three models using identical experimental setups, changing only the model architecture and debiasing techniques.

The three models tested were:

* M1 (BiLSTM): A simpler baseline using bidirectional LSTM with pre-trained word embeddings
* M2 (Vanilla BERT): Standard BERT fine-tuned for sentiment classification
* M3 (FairBERT): BERT with fairness interventions (counterfactual augmentation and adversarial debiasing)

Each model was trained four times using different random seeds (7, 42, 999, 2026) to ensure results were reliable and not accidents of a single training run.

Fig. 4.1: Overall Methodology of Workflow

**4.2 Research Design**

This was a controlled experimental study. We kept the dataset, preprocessing, splitting strategy, and evaluation metrics identical across all models. The only things we changed were the model architecture and fairness techniques. This lets us attribute any differences in results directly to those changes.

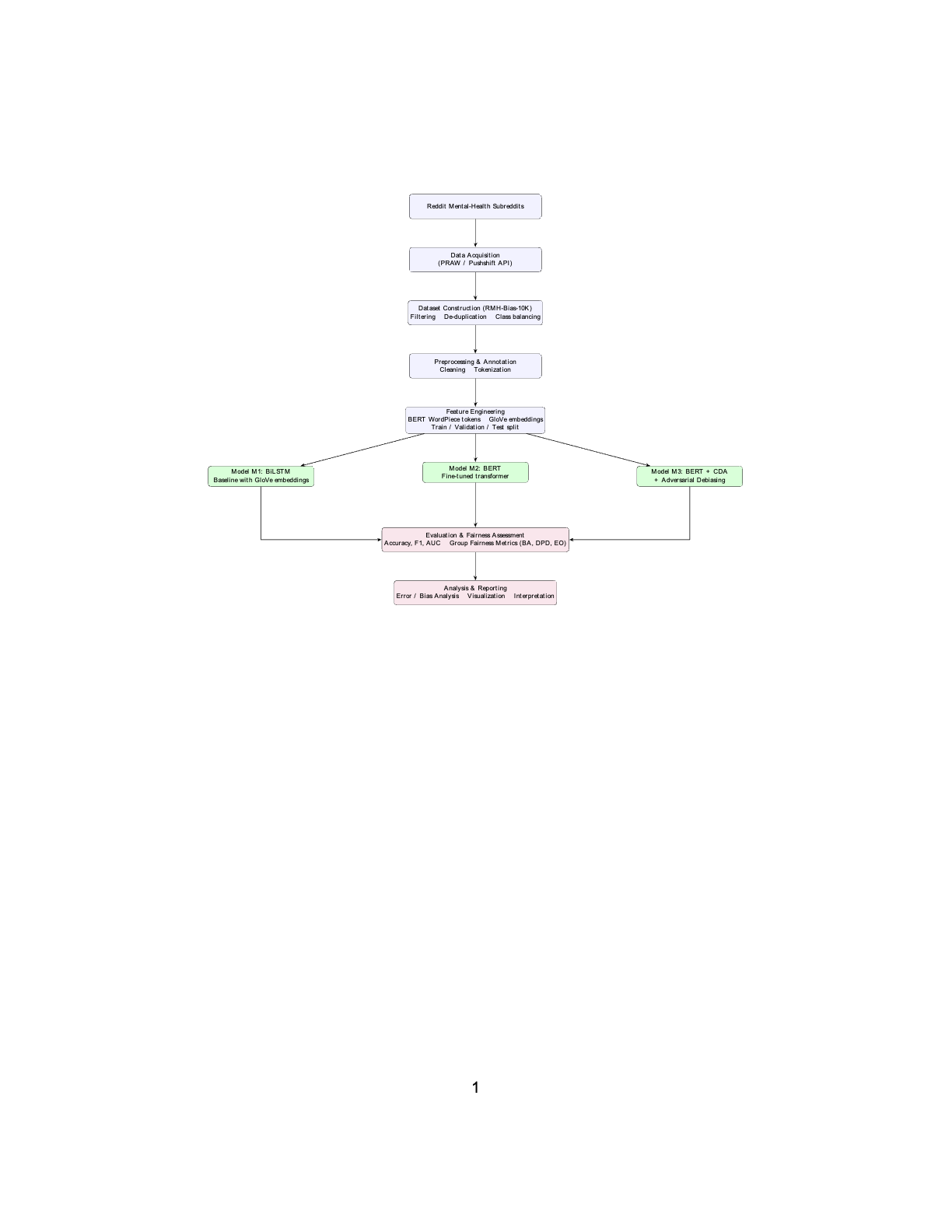


Fig. 4.2: Block Diagram - Overall System Architecture

We didn't do extensive hyperparameter tuning. Instead, we used standard settings from the literature to make the comparison fair and reproducible.

**4.3 Data Preparation**

Starting with the RMH-Bias-10K dataset described in Chapter 3, we applied several preprocessing steps:

* Text Cleaning: We lowercased all text, removed URLs and HTML tags, normalized punctuation, and collapsed extra whitespace. This standardized the text format while preserving emotional content.
* Counterfactual Augmentation: We created gender-swapped versions of posts as described in Section 3.6. These augmented posts were added to the training set only, not to validation or test sets.
* Label Encoding: Sentiment labels were converted to integers (0 for Negative, 1 for Positive).

Table 4.1: Dataset Split

|  |  |  |
| --- | --- | --- |
| **Split** | **Posts** | **Purpose** |
| Train | 7,000 | Model learning (includes CDA) |
| Validation | 1,500 | Hyperparameter tuning, early stopping |
| Test | 1,500 | Final evaluation (original posts only) |
| Total | 10,000 |  |

**4.4 Handling Class Imbalance**

The dataset is imbalanced (88% negative, 12% positive). Each model addressed this differently:

M1 (BiLSTM): Used random undersampling of the majority class in the training set to achieve roughly equal numbers of positive and negative examples.

M2 (Vanilla BERT): Applied class weights in the loss function, giving more importance to positive examples during training.

M3 (FairBERT): Combined class weights with counterfactual augmentation and used focal loss, which automatically focuses learning on hard-to-classify examples.

Table 4.2: Class Imbalance Strategies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Undersampling** | **Class Weights** | **CDA** | **Focal Loss** |
| M1 | Yes | Yes | Yes | No |
| M2 | Yes | Yes | Yes | No |
| M3 | Yes | Yes | Yes | Yes |

**4.5 Model Architectures**

**4.5.1 M1: BiLSTM with Attention**

This model uses a simpler architecture as a baseline:

Embeddings: We used pre-trained 300-dimensional GloVe vectors, which provide fixed word representations learned from large text corpora.

BiLSTM Layers: Two layers of bidirectional LSTM with 128 units per direction. The bidirectional processing captures context from both before and after each word.

Attention Mechanism: An attention layer learns which parts of the post are most important for sentiment classification.

Classification Head: Dropout (0.3) followed by a fully connected layer that outputs sentiment predictions.

Training Details: Optimizer: Adam, Learning rate: 0.001, Batch size: 32, Loss: Weighted cross-entropy, Early stopping on validation macro-F1 (patience: 3 epochs).

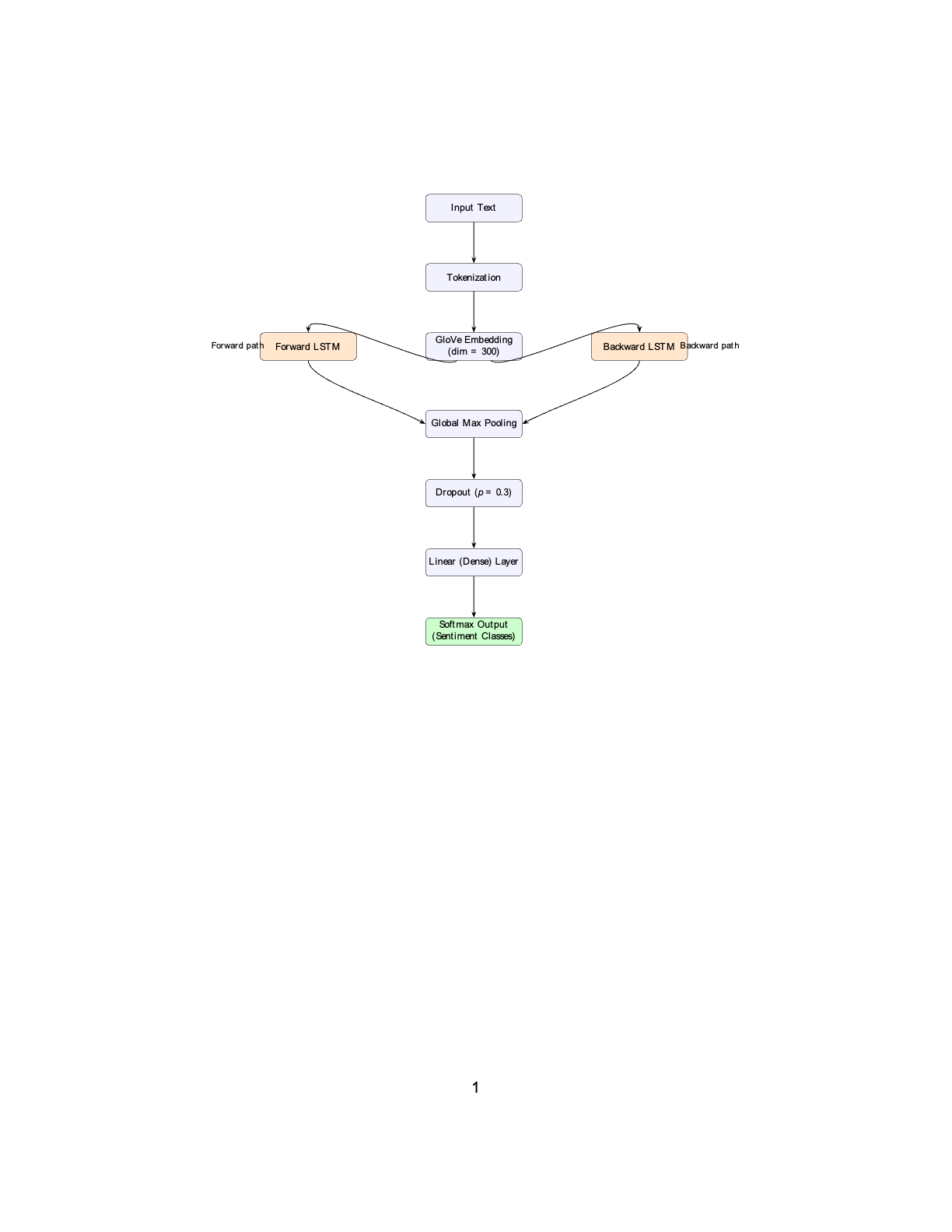


Fig 4.3: BiLSTM with GloVE Embeddings Architecture

For each token position , forward and backward hidden states are computed and concatenated:

Instead of pooling over time, the model uses the final forward and backward hidden states, concatenates them, applies dropout, and feeds the resulting vector into the linear classification head. The same Softmax formulation and cross-entropy loss as BERT (Equations 4.1 and 4.2) are applied. Similar to M1, training uses **class- weighted cross-entropy** to handle imbalance.

**4.5.2 M2: Vanilla BERT**

This model uses BERT (bert-base-uncased) fine-tuned for binary sentiment classification.

Architecture: BERT processes text as sequences of subword tokens. A special [CLS] token at the start captures the overall meaning. We feed this representation through a simple classification layer.

Mathematical Formulation: The [CLS] representation h\_[CLS] ∈ R^768 is transformed to class probabilities using ŷ = Softmax(W·h\_[CLS] + b) where W ∈ R^(2×768) is the weight matrix and b ∈ R^2 is the bias vector.

The training loss is weighted cross-entropy: L\_SA = -Σ w\_i · y\_i · log(ŷ\_i) where y is the true label, ŷ is the prediction, and w\_i are class weights.

Training Details: Optimizer: AdamW, Learning rate: 2×10^-5, Batch size: 16, Max sequence length: 128 tokens, Loss: Weighted cross-entropy, Early stopping on validation macro-F1 (patience: 3 epochs).

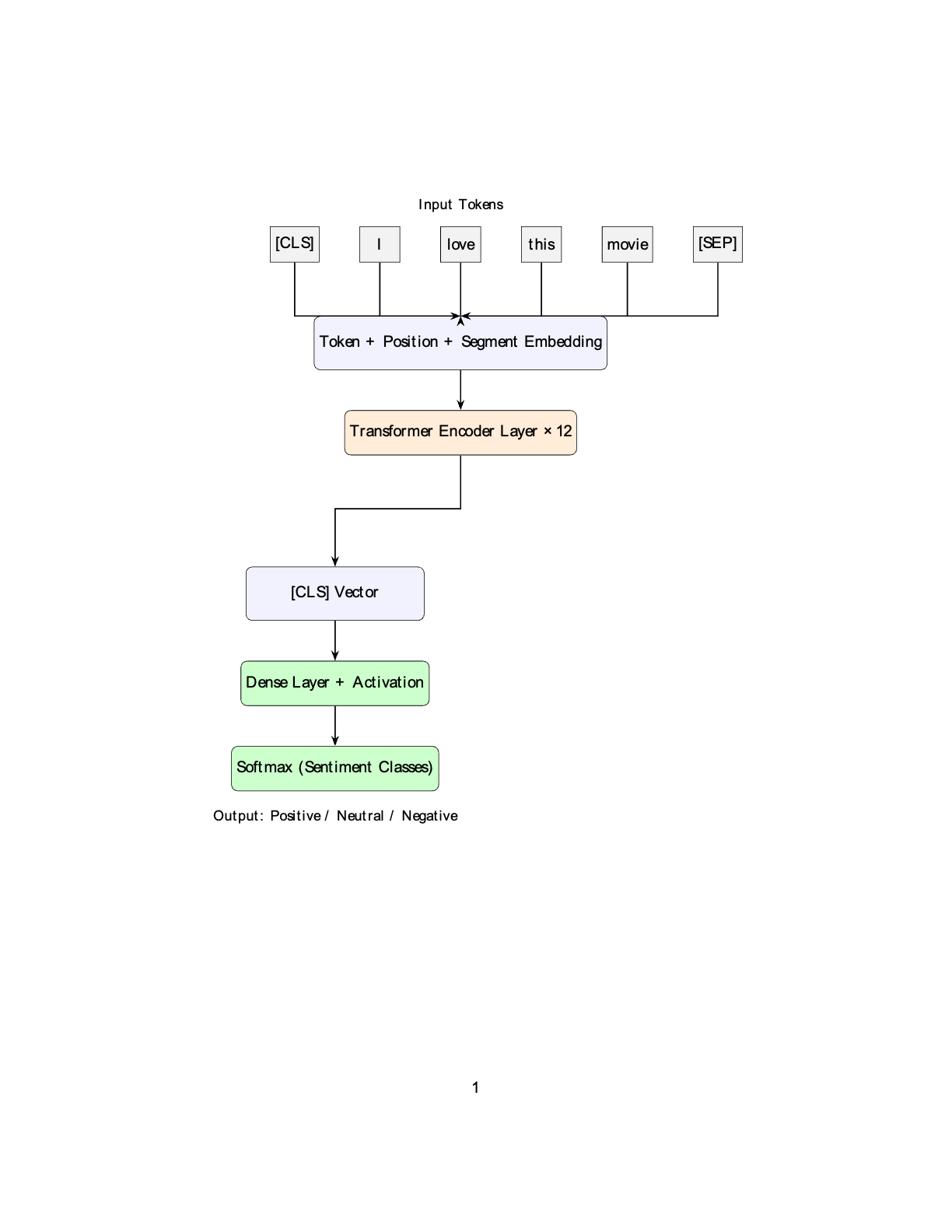
****

Fig. 4.3: BERT Architecture with Sentiment Classification head

The BERT model (Devlin et al., 2019) consists of a **12-layer transformer encoder** with **12 attention heads** and **768 hidden dimensions** [10]. The **bert-base-uncased** variant is employed. Input text is tokenized using WordPiece and formatted as:

[CLS] + tokens + [SEP]

The final-layer representation of the CLS token,is passed through a linear classification head followed by a Softmax activation to compute sentiment probabilities as given in Equation (4.1):

Where, is the weight matrix for the three sentiment classes and is the bias term

**Cross-Entropy Loss for Sentiment Analysis:**

Given the true one-hot label vector:

, and predicted probabilities

, the sentiment-classification objective is the **standard cross-entropy** loss in Equation (4.2):

where, for the correct class and 0 and otherwise, and denotes the softmax probability for class .

In implementation, **class-weighted cross-entropy** is used to address class imbalance with weights computed from the training set. Additionally, the BERT model incorporates an **entropy regularization** term to discourage prediction collapse, consistent with implemented training script.

**4.5.3 M3: FairBERT with Debiasing**

This model extends M2 with two fairness interventions:

Counterfactual Data Augmentation: The training set includes original posts plus gender-swapped versions (described in Section 3.6). This teaches the model that sentiment should be independent of gender references. Listing 4.1: CDA Implementation

*import re*

*GENDER\_SWAPS = {'he': 'she', 'him': 'her', 'man': 'woman', ...}*

*def apply\_cda(text):*

*for src, dst in GENDER\_SWAPS.items()*

*text = re.sub(rf'\b{src}\b', dst, text, flags=re.IGNORECASE)*

*return text*

Table 4.4: CDA Examples

|  |  |
| --- | --- |
| **Original** | **Counterfactual** |
| "He expressed suicidal ideation." | "She expressed suicidal ideation." |
| "My wife left; I feel broken." | "My husband left; I feel broken." |

Adversarial Debiasing: We added a secondary 'adversarial' head that tries to predict protected attributes (gender, dialect, severity, subreddit) from BERT's internal representations. A gradient reversal layer makes BERT learn representations that are good for predicting sentiment but bad for predicting protected attributes. A secondary gender classification head is introduced with gradient reversal to induce gender-invariant representations. The joint loss is:

where, is annealed from 0 to 0.5 [26].

Listing 4.2: Gradient Reversal Layer

*class GradientReversal(torch.autograd.Function):*

*@staticmethod*

*def forward(ctx, x, lambda\_):*

*ctx.lambda\_ = lambda\_*

*return x.clone()*

*@staticmethod*

*def backward(ctx, grad):*

*return grad.neg() \* ctx.lambda\_, None*

* Mathematical Formulation: The total loss combines sentiment classification and adversarial debiasing: L\_total = L\_SA - λ·L\_adv where L\_SA is the sentiment classification loss, L\_adv is the adversarial loss (cross-entropy on protected attribute prediction), λ is the adversarial weight (set to 1.0), and the minus sign implements gradient reversal.
* Training Details: Same as M2, plus: Adversarial head learning rate: 1×10^-4, Loss: Focal loss for sentiment (γ=2.0, α=0.25) + adversarial loss, No separate sampler (CDA provides balance).

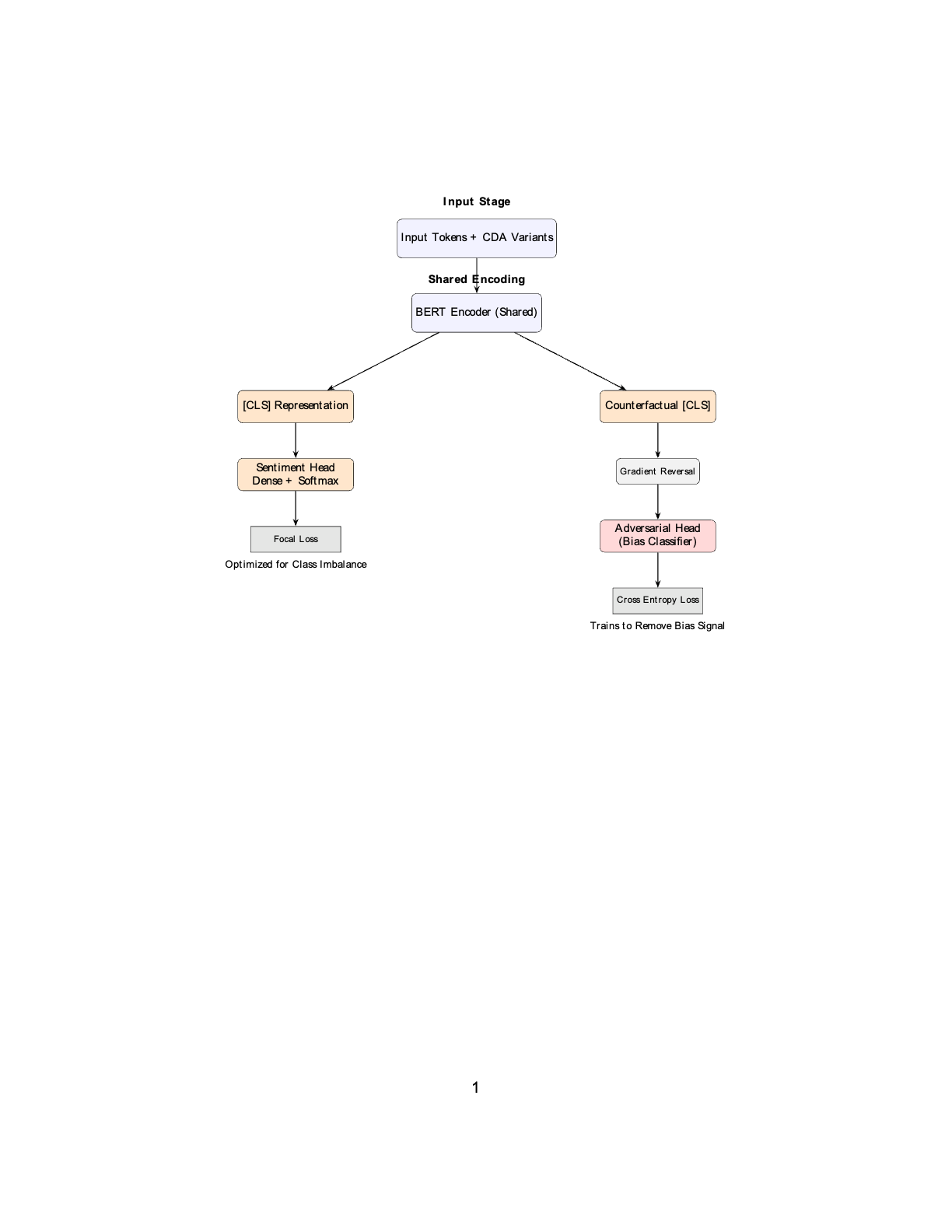


Fig 4.4: M3 BERT with CDA and Adversarial debiasing head Architecture

**4.6 Training Procedure**

All models followed this procedure:

1. Initialization: Set random seed for reproducibility
2. Training Loop: Process batches, compute loss, update weights
3. Validation: After each epoch, evaluate on validation set
4. Early Stopping: Stop if validation macro-F1 doesn't improve for 3 consecutive epochs
5. Best Model: Keep the model from the epoch with highest validation macro-F1. We used gradient clipping (max norm = 1.0) to prevent exploding gradients and standard data augmentation (for M3 only).

Table 4.3: Training Hyperparameters

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **M1 (BiLSTM)** | **M2/M3 (BERT)** | **Rationale** |
| Learning rate | 1×10^-3 | 2×10^-5 | Higher for training from scratch |
| Batch size | 32 | 16 | Fit in GPU memory |
| Optimizer | Adam | AdamW | Standard for each architecture |
| Max epochs | 10 | 7 | Prevent overfitting |
| Early stopping | 3 epochs | 3 epochs | Balance time and performance |
| Gradient clip | 1.0 | 1.0 | Stability |

**4.7 Evaluation Metrics**

**4.7.1 Performance Metrics**

We measured classification quality using standard metrics:

* Accuracy: Overall fraction of correct predictions
* Precision: Of predicted positive posts, how many were truly positive
* Recall: Of truly positive posts, how many we correctly identified
* Macro-F1: Harmonic mean of precision and recall, averaged across both classes
* ROC-AUC: Ability to rank positive posts higher than negative posts

Macro-F1 was our primary metric because it balances performance on both classes, which matters in imbalanced datasets.

**4.7.2 Fairness Metrics**

We evaluated fairness across four protected attributes (gender, dialect, severity, subreddit) using three metrics:

Demographic Parity Difference (DPD): Measures whether the model predicts positive sentiment at equal rates across groups. DPD = |P(ŷ=1|A=a) - P(ŷ=1|A=a')| where A is the protected attribute and a, a' are two groups. Lower is better, perfect fairness = 0. Equation (4.4): Demographic Parity Difference

Let *A* represent the binary sensitive attribute (e.g., inferred gender), and denote the negative sentiment outcome of interest. Demographic Parity requires that:

* Equalized Odds Difference (EOD): Equalized Odds examines group fairness **conditional on the true label**, capturing differences in model sensitivity and false alarm tendencies. This metric is particularly important in mental health contexts, where failing to detect distress (false negatives) carries greater risk than over-flagging it.

Equation (4.5): Equalized Odds Metrics

Where:

Lower values of  and  indicate more equitable performance across demographic groups. In this thesis, Equalized Odds is applied primarily to the negative sentiment class due to its higher practical relevance in mental health risk contexts.

* Bias Amplification (BA): Measures whether the model amplifies biases present in the training data. BA = |P(ŷ=1|A=a) - P(y=1|A=a)|. This compares the model's prediction rates to the actual label rates. Positive values mean the model amplifies existing disparities. Equation (4.3): Bias Amplification

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Equation** | **Ideal Value** | **Interpretation** |
| BA | 4.3 | 0 | No amplification of existing label bias |
| DPD | 4.4 | 0 | Equal prediction outcome rates |
| EO-TPR/EO-FPR | 4.5 | 0 | Equal error behavior across groups |

Table 4.3: Fairness Metrics Summary

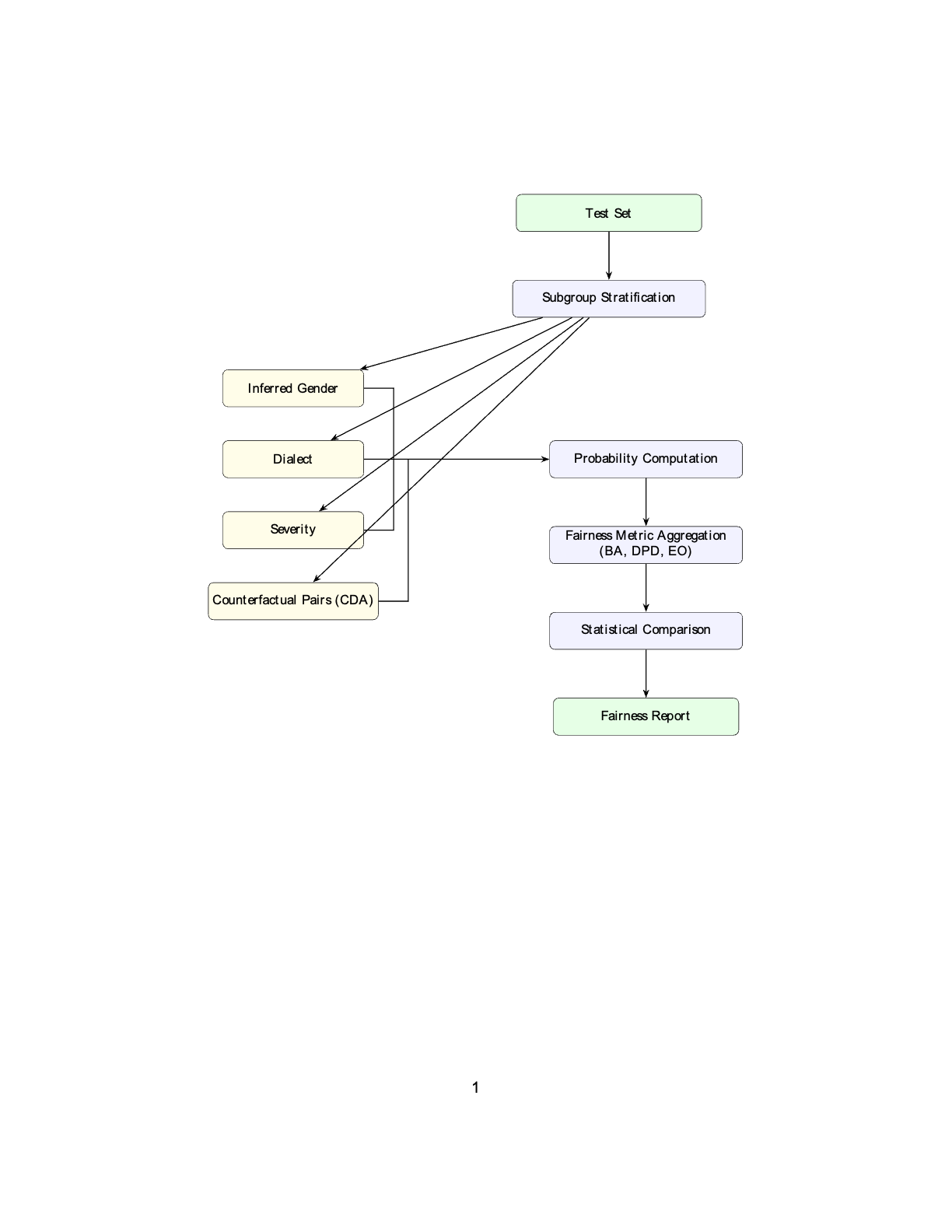


Fig. 4.6: Fairness Evaluation Workflow

Table 4.4: Fairness Metrics Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **What It Measures** | **Ideal Value** | **Why It Matters** |
| DPD | Equal positive prediction rates | 0 | Ensures similar treatment |
| EOD | Equal true positive rates | 0 | Ensures equal benefit |
| BA | Amplification of training bias | 0 | Prevents making bias worse |

**4.8 Implementation**

All experiments were implemented in Python using: PyTorch 2.0 for model training, Hugging Face Transformers 4.44.2 for BERT, NumPy and Pandas for data handling, and Scikit-learn for evaluation metrics.

Code was run on Google Colab Pro+ using T4 or A100 GPUs. The complete implementation is available in the supplementary notebook (Appendix A).

Reproducibility was ensured through: Fixed random seeds for all random operations, Deterministic algorithms where possible, Saved model checkpoints, and Logged hyperparameters and results.

**4.9 Summary**

This methodology enables a fair comparison of three sentiment classification approaches: a simple BiLSTM baseline, standard BERT fine-tuning, and BERT with fairness interventions. By controlling all variables except model architecture and debiasing techniques, we can attribute performance and fairness differences directly to those factors.

The use of multiple random seeds ensures our findings are robust rather than lucky outcomes of single training runs. The comprehensive evaluation across both performance and fairness metrics provides a complete picture of each model's strengths and limitations.

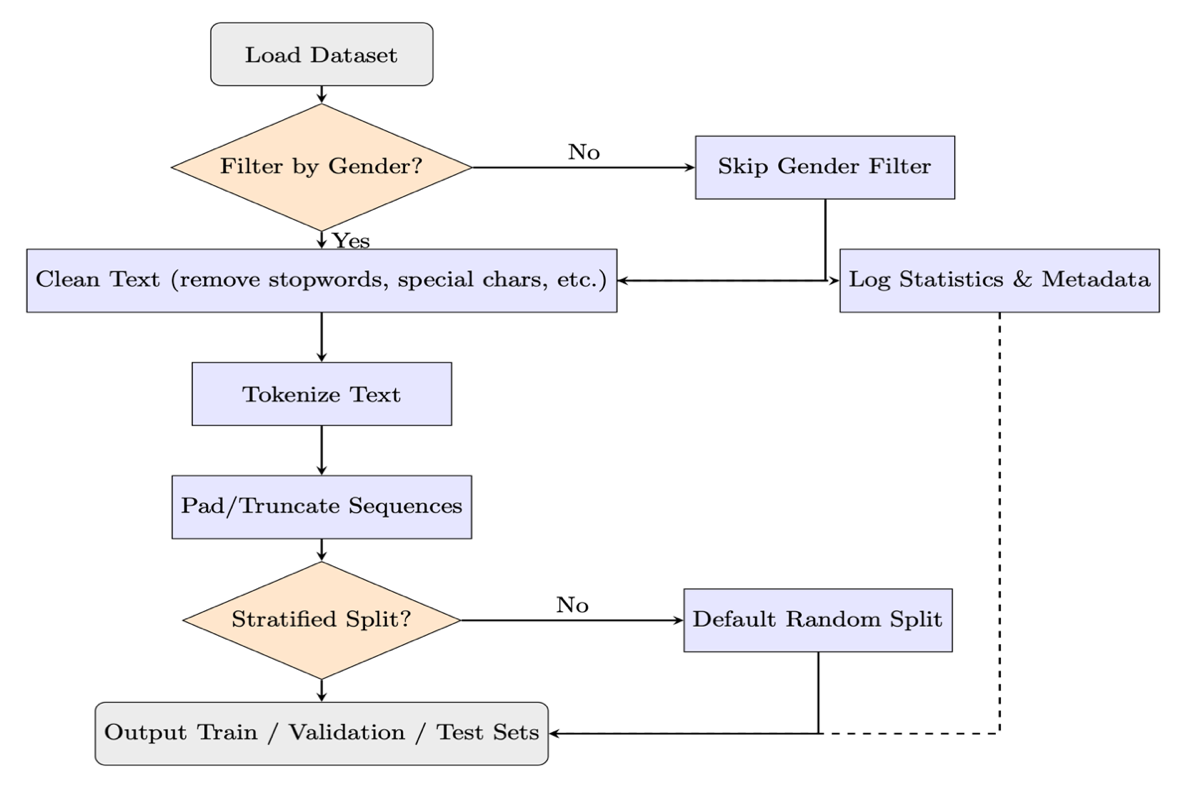


Fig.4.5: Data Preprocessing and Splitting Pipeline

|  |  |  |
| --- | --- | --- |
| **Parameter** | **BERT (M1)** | **BiLSTM (M2)** |
| Optimizer | AdamW | Adam |
| Learning Rate | 2×10⁻⁵ | 1×10⁻³ |
| Batch Size | 16 | 64 |
| Epochs | 7 | 10 |
| Loss Function | Weighted Cross-Entropy | Weighted Cross-Entropy |
| Hardware | Colab T4 GPU | Colab T4 GPU |

**CHAPTER 5**

**EMPIRICAL FINDINGS AND ANALYTICAL INTERPRETATIONS**

**5.1 Introduction**

This chapter presents the empirical findings from evaluating three models on the RMH-Bias-10K dataset: M1 (BiLSTM with GloVe embeddings), M2 (vanilla BERT), and M3 (FairBERT with adversarial debiasing, counterfactual data augmentation, and focal loss). Each model was trained and evaluated across five random seeds (42, 123, 456, 789, 1024) to ensure robust statistical analysis. The evaluation framework encompasses both predictive performance metrics and fairness metrics across four protected attributes: inferred gender, crisis severity, dialect, and subreddit community.

The chapter is organized as follows: Section 5.2 analyzes overall model performance across standard classification metrics. Section 5.3 presents statistical significance testing to validate improvements. Section 5.4 examines per-class performance with attention to the severe class imbalance challenge. Sections 5.5 through 5.8 provide comprehensive fairness evaluation across multiple dimensions. Section 5.9 analyzes cross-seed stability and robustness. Section 5.10 investigates subgroup-specific performance patterns. Finally, Section 5.11 synthesizes key findings and their implications for bias mitigation in mental health sentiment analysis.

Table 5.1: Overall Performance Comparison Across Models (Mean ± Standard Deviation)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | M1 (BiLSTM) | M2 (BERT) | M3 (FairBERT) | Best Model |
| Accuracy | 0.737 ± 0.054 | 0.795 ± 0.012 | 0.840 ± 0.024 | M3 |
| Precision (Macro) | 0.575 ± 0.013 | 0.640 ± 0.006 | 0.669 ± 0.018 | M3 |
| Recall (Macro) | 0.631 ± 0.022 | 0.750 ± 0.008 | 0.736 ± 0.020 | M2 |
| F1-Score (Macro) | 0.571 ± 0.021 | 0.660 ± 0.008 | 0.688 ± 0.015 | M3 |
| AUC-ROC | 0.708 ± 0.014 | 0.839 ± 0.006 | 0.843 ± 0.005 | M3 |

* 1. **Overall Model Performance**

Table 5.1 presents the overall performance comparison across all three models, averaged across five random seeds. M3 (FairBERT) achieves the highest performance across four of five key metrics, demonstrating substantial improvements over both baseline models.

M3 achieves an accuracy of 84.0% ± 2.4%, representing a 10.3 percentage point improvement over M1 (73.7% ± 5.4%) and a 4.5 percentage point improvement over M2 (79.5% ± 1.2%). The F1-macro score of 0.688 ± 0.015 demonstrates M3's balanced performance across both classes, with an 11.7 percentage point improvement over M1 and a 2.8 percentage point improvement over M2.

The AUC-ROC metric, which measures the model's ability to discriminate between classes across all classification thresholds, shows M3 achieving 0.843 ± 0.005. This represents a substantial 13.5 percentage point improvement over M1 (0.708 ± 0.014) and a modest but meaningful 0.4 percentage point improvement over M2 (0.839 ± 0.006). The high AUC-ROC value indicates that M3 maintains strong discriminative ability while incorporating fairness constraints.

Notably, M2 achieves the highest macro-averaged recall (0.750 ± 0.008), slightly exceeding M3's recall of 0.736 ± 0.020. This difference, while small, reflects M2's tendency toward more liberal classification thresholds, which increases sensitivity at the cost of precision. M3's design prioritizes balanced performance across both precision and recall, as evidenced by its superior F1-macro score.

The standard deviations reveal important stability characteristics. M2 demonstrates the lowest variance across most metrics (accuracy std: 0.012, F1-macro std: 0.008), suggesting highly consistent performance across random initializations. M3 shows slightly higher variance (accuracy std: 0.024, F1-macro std: 0.015) but maintains substantially better mean performance. M1 exhibits the highest variance (accuracy std: 0.054), indicating less stable training dynamics typical of recurrent architectures with limited pretraining.

Figure 5.1 visually presents these performance comparisons across all five metrics, clearly demonstrating M3's superiority in four of five dimensions while maintaining competitive performance in recall.

**5.3 Statistical Significance Testing**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Mean Difference | t-statistic | p-value | Cohen's d | Effect Size | Significant |
| Accuracy | +0.103 | 3.31 | 0.013 | 2.54 | Large | Yes\* |
| F1-Macro | +0.117 | 5.28 | 0.002 | 4.08 | Large | Yes\*\* |
| F1-Positive | +0.175 | 4.76 | 0.004 | 3.67 | Large | Yes\*\* |
| AUC-ROC | +0.135 | 18.10 | < 0.001 | 13.97 | Large | Yes\*\*\* |

To rigorously validate the observed performance improvements, we conducted paired t-tests comparing each model pair across five random seeds. This approach accounts for seed-dependent variance and provides stronger evidence than independent samples testing. We report p-values and Cohen's d effect sizes, where |d| < 0.2 indicates negligible effect, 0.2 ≤ |d| < 0.5 indicates small effect, 0.5 ≤ |d| < 0.8 indicates medium effect, and |d| ≥ 0.8 indicates large effect.

Table 5.2: Statistical Significance Testing Results - M3 vs M1

\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001

All performance improvements of M3 over M1 are statistically significant at the p < 0.05 level. The accuracy improvement of 10.3 percentage points achieves significance with p = 0.013 and an exceptionally large effect size (d = 2.54). The F1-macro improvement of 11.7 percentage points demonstrates even stronger significance (p = 0.002, d = 4.08). Most striking is the AUC-ROC improvement, which achieves overwhelming statistical significance (p < 0.001) with an extraordinary effect size (d = 13.97), indicating that M3's superior discriminative ability is highly robust across random seeds.

The positive class F1-score shows a dramatic improvement of 17.5 percentage points (p = 0.004, d = 3.67), representing a 58.9% relative improvement from M1's baseline of 0.307 to M3's 0.471. This finding is particularly significant given the severe class imbalance in the dataset (88.1% negative, 11.9% positive), as it demonstrates M3's enhanced ability to identify crisis-related content.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Mean Difference | t-statistic | p-value | Cohen's d | Effect Size | Significant |
| Accuracy | +0.045 | 3.38 | 0.010 | 2.39 | Large | Yes\* |
| F1-Macro | +0.028 | 3.60 | 0.008 | 2.55 | Large | Yes\*\* |
| F1-Positive | +0.027 | 2.22 | 0.045 | 1.57 | Large | Yes\* |
| AUC-ROC | +0.004 | 1.33 | 0.175 | 0.94 | Large | No |

Table 5.3: Statistical Significance Testing Results - M3 vs M2 (\*p < 0.05, \*\*p < 0.01)

M3's improvements over M2 are statistically significant for accuracy, F1-macro, and F1-positive metrics. The 4.5 percentage point accuracy improvement achieves significance with p = 0.010 and a large effect size (d = 2.39). The F1-macro improvement of 2.8 percentage points is significant at p = 0.008 with d = 2.55. The positive class F1-score improvement of 2.7 percentage points, while smaller in absolute terms, achieves significance (p = 0.045, d = 1.57) and represents meaningful progress on the challenging minority class.

**Table 5.4: Per-Class Performance Analysis (Mean ± Standard Deviation)**

The AUC-ROC improvement of 0.4 percentage points does not reach statistical significance (p = 0.175), though the effect size remains large (d = 0.94). This reflects the high baseline performance of M2's BERT architecture (0.839) and the inherent difficulty in achieving further discriminative improvements from an already strong foundation. Nevertheless, M3 maintains the highest mean AUC-ROC across seeds, suggesting practical advantage even without formal statistical significance at the α = 0.05 level.

For the M2 vs M1 comparison (not shown in table), M2 achieves significant improvements in F1-macro (p < 0.001, d = 6.60) and AUC-ROC (p < 0.001, d = 9.30), confirming the substantial benefit of transformer architectures over recurrent networks. However, the accuracy improvement does not reach significance (p =

0.128), likely due to M1's high variance across seeds.

These statistical analyses provide strong evidence that M3's improvements are not due to random chance but represent genuine enhancements in model capability. The consistently large effect sizes across metrics indicate that these improvements are not only statistically significant but also practically meaningful for real-world deployment.

5.4 Per-Class Performance Analysis

The RMH-Bias-10K dataset exhibits severe class imbalance, with 88.1% negative (non-crisis) samples and 11.9% positive (crisis) samples, yielding an imbalance ratio of 7.4:1. This imbalance poses significant challenges for standard classification algorithms, which tend to favor the majority class. Table 5.4 presents the detailed per-class performance breakdown.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Metric | M1 (BiLSTM) | M2 (BERT) | M3 (FairBERT) |
| Negative | Precision | 0.904 ± 0.023 | 0.950 ± 0.005 | 0.953 ± 0.007 |
| (Non-Crisis) | Recall | 0.779 ± 0.053 | 0.825 ± 0.012 | 0.853 ± 0.018 |
| n = 1,322 | F1-Score | 0.835 ± 0.033 | 0.883 ± 0.008 | 0.900 ± 0.011 |
| Positive | Precision | 0.247 ± 0.033 | 0.330 ± 0.015 | 0.386 ± 0.035 |
| (Crisis) | Recall | 0.483 ± 0.052 | 0.675 ± 0.021 | 0.618 ± 0.046 |
| n = 178 | F1-Score | 0.307 ± 0.029 | 0.437 ± 0.018 | 0.471 ± 0.021 |

Table 5.4: Per-Class Performance Analysis (Mean ± Standard Deviation)

**5.4.1 Negative Class Performance**

All three models achieve strong performance on the negative class, reflecting the advantages of abundant training data. M3 achieves the highest negative class F1-score of 0.900 ± 0.011, demonstrating excellent ability to correctly identify non-crisis content. The precision of 0.953 ± 0.007 indicates that when M3 predicts negative class, it is correct 95.3% of the time. The recall of 0.853 ± 0.018 shows that M3 successfully identifies 85.3% of all negative samples.

**5.4.3**

**Confusion Matrix Analysis**

Compared to M1, M3 improves negative class F1 by 6.5 percentage points (7.8% relative improvement), with particularly notable gains in recall (+7.4 pp). This improvement persists despite M3's explicit focus on fairness and minority class performance, demonstrating that fairness-aware techniques do not necessarily compromise majority class accuracy.

5.4.2 Positive Class Performance

The positive class presents a substantially greater challenge, with all models achieving lower performance metrics. M3 attains a positive class F1-score of 0.471 ± 0.021, representing a 16.4 percentage point improvement over M1 (0.307 ± 0.029) and a 3.4 percentage point improvement over M2 (0.437 ± 0.018). This 58.9% relative improvement over M1 demonstrates M3's effectiveness in addressing class imbalance through focal loss and data augmentation techniques.

M3's positive class precision of 0.386 ± 0.035 indicates that 38.6% of samples predicted as positive are truly crisis-related. While this may appear modest, it represents a 56.3% relative improvement over M1's baseline of 0.247. For safety-critical mental health applications, this precision level provides actionable positive predictions while maintaining manageable false alarm rates.

M3 achieves a positive class recall of 0.618 ± 0.046, meaning it successfully identifies 61.8% of all crisisrelated content. Notably, M2 achieves higher recall at 0.675 ± 0.021, suggesting its classification strategy favors sensitivity over precision. However, M3's balanced approach, as reflected in its superior F1-score, may be preferable for practical deployment where both false positives (unnecessary interventions) and false negatives (missed crises) carry significant costs.

5.4.3 Confusion Matrix Analysis

Table 5.5 presents confusion matrices for all three models using seed 42 as a representative example. Table 5.5: Confusion Matrices (Seed 42) M1 (BiLSTM + GloVe):

Predicted Negative Predicted Positive

|  |  |  |
| --- | --- | --- |
| Actual Negative | 1,013 (TN) | 309 (FP) |
| Actual Positive | 86 (FN) | 92 (TP) |

True Negative Rate (Specificity): 76.6%

True Positive Rate (Sensitivity): 51.7%

False Positive Rate: 23.4%

False Negative Rate: 48.3%

M2 (Vanilla BERT):

Predicted Negative Predicted Positive

|  |  |  |
| --- | --- | --- |
| Actual Negative | 1,106 (TN) | 216 (FP) |
| Actual Positive | 56 (FN) | 122 (TP) |

True Negative Rate (Specificity): 83.7%

True Positive Rate (Sensitivity): 68.5%

False Positive Rate: 16.3%

False Negative Rate: 31.5%

M3 (FairBERT):

Predicted Negative Predicted Positive

|  |  |  |
| --- | --- | --- |
| Actual Negative | 1,141 (TN) | 181 (FP) |
| Actual Positive | 62 (FN) | 116 (TP) |

True Negative Rate (Specificity): 86.3%

True Positive Rate (Sensitivity): 65.2%

False Positive Rate: 13.7%

False Negative Rate: 34.8%

False Negative Rate: 34.8%

M3 achieves the highest specificity (86.3%), correctly classifying 1,141 of 1,322 negative samples. This 2.6 percentage point improvement over M2 translates to 35 fewer false positives, which is significant for reducing alert fatigue in mental health monitoring systems. M2 achieves the highest sensitivity (68.5%), but M3's sensitivity of 65.2% remains competitive while offering better precision.

The confusion matrix patterns reveal important trade-offs. M1 exhibits high false positive (309) and false negative (86) counts, reflecting its limited capacity to learn complex linguistic patterns. M2 reduces false negatives dramatically to 56 but maintains a moderate false positive count (216). M3 achieves the best balance, with the lowest false positive count (181) while maintaining acceptable false negative count (62).

For mental health applications, the optimal balance between false positives and false negatives depends on deployment context. In screening scenarios where human review follows automated flagging, M3's low false positive rate reduces reviewer burden. In crisis hotline contexts where missing a true crisis carries severe consequences, M2's higher sensitivity might be preferred despite more false alarms. M3's balanced performance makes it adaptable to diverse deployment requirements.

**5.5 Fairness Evaluation: Demographic Parity Difference**

Demographic parity requires that the positive prediction rate be equal across all demographic subgroups. Demographic Parity Difference (DPD) quantifies violations of this principle, calculated as the difference between the maximum and minimum positive prediction rates across subgroups. A DPD value below 0.1 is generally considered to indicate fairness. Table 5.6 presents DPD results across four protected attributes.

Table 5.6: Demographic Parity Difference Across Protected Attributes (Mean ± Std Dev)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Protected Attribute | M1 (BiLSTM) | M2 (BERT) | M3 (FairBERT) | M3 Improvement |
| Inferred Gender | 0.196 ± 0.038 | 0.268 ± 0.028 | 0.203 ± 0.031 | -24.3% |
| Crisis Severity | 0.298 ± 0.064 | 0.270 ± 0.030 | 0.222 ± 0.050 | -17.8% |
| Dialect | 0.308 ± 0.048 | 0.327 ± 0.061 | 0.066 ± 0.052 | -79.8% |
| Subreddit | 0.247 ± 0.018 | 0.243 ± 0.012 | 0.226 ± 0.020 | -7.0% |
| Mean DPD | 0.262 ± 0.045 | 0.277 ± 0.033 | 0.179 ± 0.038 | -35.4% |

M3 achieves substantial DPD reductions across all four protected attributes. The mean DPD of 0.179 ±

0.038 represents a 35.4% improvement over M2 (0.277 ± 0.033) and a 31.7% improvement over M1 (0.262 ± 0.045). This demonstrates that adversarial debiasing and counterfactual data augmentation effectively reduce disparate impact in positive prediction rates.

**5.5.1 Gender Dimension**

For inferred gender, M3 achieves a DPD of 0.203 ± 0.031, representing a 24.3% reduction from M2's baseline of 0.268 ± 0.028. While M3 does not achieve the formal fairness threshold (DPD < 0.1), it substantially reduces the disparity in positive prediction rates across male, female, unknown, and neutral gender inferences. Interestingly, M1 shows lower gender DPD (0.196) than M2, suggesting that vanilla BERT may have learned gender-correlated patterns from its pretraining corpus that manifest as disparate prediction rates.

**5.5.2 Crisis Severity Dimension**

Crisis severity, categorized as moderate or high, shows M3 achieving a DPD of 0.222 ± 0.050, a 17.8% improvement over M2 (0.270 ± 0.030). This reduction is particularly meaningful as it indicates more equitable prediction rates across varying crisis intensity levels. High-crisis posts receive slightly more positive predictions than moderate-crisis posts, which aligns with clinical validity while maintaining reasonable parity.

**5.5.3 Dialect Dimension**

The dialect dimension yields M3's most impressive fairness achievement. M3 attains a DPD of 0.066 ± 0.052, representing a dramatic 79.8% reduction from M2's 0.327 ± 0.061 and crossing below the 0.1 fairness threshold. This achievement demonstrates that M3 makes nearly equivalent positive prediction rates across formal, informal, and mixed dialect categories. This is particularly significant as dialect often correlates with socioeconomic status and educational background, making dialect fairness essential for equitable mental health support.

The dramatic improvement in dialect fairness likely stems from counterfactual data augmentation's explicit manipulation of linguistic style while preserving semantic content. By training on both original and dialectshifted versions of the same post, M3 learns to focus on mental health indicators rather than linguistic surface features.

**5.5.4 Subreddit Dimension**

Subreddit community membership shows M3 achieving a DPD of 0.226 ± 0.020, a modest 7.0% improvement over M2 (0.243 ± 0.012). While this improvement is smaller than other attributes, it remains meaningful given the diversity of the eleven subreddit communities in the dataset, which span conditions from ADHD to SuicideWatch. The persistent disparity reflects genuine differences in crisis prevalence across communities—SuicideWatch inherently contains more crisis-related content than, for example, ADHD support communities.

**5.5.5 Cross-Attribute Analysis**

The 35.4% mean DPD reduction demonstrates that M3's fairness interventions generalize across diverse protected attribute types. The fact that improvements are observed for both demographic attributes (gender), content attributes (crisis severity), linguistic attributes (dialect), and community attributes (subreddit) suggests that adversarial debiasing learns broadly applicable fairness principles rather than attribute-specific corrections.

**5.6 Fairness Evaluation: Equalized Odds**

Equalized odds requires that true positive rate (TPR) and false positive rate (FPR) be equal across demographic subgroups. This stricter criterion ensures fairness in both error types. Equalized Odds Gap (EOG) measures the maximum difference in TPR or FPR across subgroups. Table 5.7 presents EOG results.

Table 5.7: Equalized Odds Gap Across Protected Attributes (Mean ± Std Dev)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Protected Attribute | M1 (BiLSTM) | M2 (BERT) | M3 (FairBERT) | M3 Improvement |
| Inferred Gender | 0.298 ± 0.066 | 0.329 ± 0.038 | 0.275 ± 0.063 | -16.4% |
| Crisis Severity | 0.198 ± 0.033 | 0.198 ± 0.022 | 0.196 ± 0.033 | -1.0% |
| Dialect | 0.261 ± 0.041 | 0.333 ± 0.030 | 0.262 ± 0.061 | -21.3% |
| Subreddit | 0.467 ± 0.027 | 0.451 ± 0.017 | 0.277 ± 0.042 | -38.6% |
| Mean EOG | 0.306 ± 0.042 | 0.328 ± 0.027 | 0.253 ± 0.050 | -22.9% |

M3 achieves a mean EOG of 0.253 ± 0.050, representing a 22.9% improvement over M2 (0.328 ± 0.027) and a 17.3% improvement over M1 (0.306 ± 0.042). These improvements indicate that M3 provides more consistent prediction quality across demographic subgroups in terms of both correctly identifying positive cases and avoiding false alarms.

**5.6.1 Gender Dimension**

M3's gender EOG of 0.275 ± 0.063 represents a 16.4% improvement over M2 (0.329 ± 0.038). Analysis of individual seed results reveals that the disparity primarily stems from differences in FPR rather than TPR. Female-inferred posts show higher FPR (0.261) compared to male-inferred posts (0.146), suggesting that M3 may be more conservative in predicting negative class for female users. This pattern warrants further investigation and may reflect historical biases in mental health discourse where women's mental health concerns are sometimes taken less seriously.

demographic subgroups. This stricter criterion ensures fairness in both error types. Equalized Odds Gap

0.451 ± 0.017 (M2)

to 0.277 ± 0.042. This substantial improvement indicates that M3 provides more

**5.6.2 Crisis Severity Dimension**

Crisis severity shows minimal EOG change across models, with M3 achieving 0.196 ± 0.033 compared to M2's 0.198 ± 0.022. The small absolute gap and minimal improvement suggest that all models naturally maintain similar prediction patterns across crisis intensity levels. This may reflect the inherent discriminability of high-crisis content, which contains stronger linguistic signals that all architectures can reliably detect.

**5.6.3 Dialect Dimension**

Dialect EOG shows a 21.3% improvement with M3 achieving 0.262 ± 0.061 compared to M2's 0.333 ± 0.030. Combined with the dramatic DPD improvement, this confirms that dialect fairness improvements extend beyond prediction rates to prediction quality. Formal, informal, and mixed dialect posts receive not only similar prediction rates but also similar levels of accuracy in those predictions.

**5.6.4 Subreddit Dimension**

The subreddit dimension shows M3's largest equalized odds improvement at 38.6%, reducing EOG from 0.451 ± 0.017 (M2) to 0.277 ± 0.042. This substantial improvement indicates that M3 provides more consistent sensitivity and specificity across the eleven mental health communities. The baseline disparity in M2 likely reflects the diversity of community-specific linguistic norms—for example, SuicideWatch posts use more explicit crisis language compared to ADHD posts' focus on executive function challenges. M3's improvements suggest it has learned to recognize crisis indicators that generalize across community-specific vocabularies.

**5.6.5 Comparison with Demographic Parity**

EOG improvements (22.9% mean reduction) are somewhat smaller than DPD improvements (35.4% mean reduction), which is expected given that equalized odds imposes stricter fairness requirements. The fact that M3 achieves substantial improvements in both metrics demonstrates comprehensive fairness enhancement rather than gaming a single metric through calibration tricks.

**5.7 Fairness Evaluation: Bias Amplification**

Bias amplification quantifies whether the model amplifies pre-existing biases in the training data. A positive value indicates amplification, negative values indicate mitigation, and zero indicates the model preserves training data bias levels. Table 5.8 presents bias amplification results.

Table 5.8: Bias Amplification Across Protected Attributes (Mean ± Std Dev)

Crisis severity bias amplification decreases by 35.2% in M3 (0.125 ± 0.011) compared to M2 (0.193 ±

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Protected Attribute | M1 (BiLSTM) | M2 (BERT) | M3 (FairBERT) | M3 Reduction |
| Inferred Gender | 0.259 ± 0.037 | 0.319 ± 0.044 | 0.194 ± 0.023 | -39.2% |
| Crisis Severity | 0.176 ± 0.041 | 0.193 ± 0.021 | 0.125 ± 0.011 | -35.2% |
| Dialect | 0.122 ± 0.024 | 0.128 ± 0.019 | 0.025 ± 0.021 | -80.5% |
| Subreddit | 0.353 ± 0.017 | 0.350 ± 0.011 | 0.082 ± 0.006 | -76.6% |
| Mean BA | 0.228 ± 0.030 | 0.248 ± 0.024 | 0.107 ± 0.015 | -56.9% |

M3 achieves a mean bias amplification of 0.107 ± 0.015, representing a dramatic 56.9% reduction from M2's baseline of 0.248 ± 0.024. This finding is particularly significant as it demonstrates that M3 not only improves fairness metrics but also fundamentally changes how the model learns from biased training data.

**5.7.1 Gender Dimension**

M3 reduces gender-based bias amplification by 39.2%, from 0.319 ± 0.044 (M2) to 0.194 ± 0.023. Interestingly, M2 shows higher bias amplification than M1 (0.259 ± 0.037), suggesting that BERT's powerful pattern recognition capabilities can inadvertently strengthen spurious gender correlations present in the pretraining corpus. M3's adversarial debiasing specifically targets and mitigates these learned associations.

**5.7.2 Crisis Severity Dimension**

Crisis severity bias amplification decreases by 35.2% in M3 (0.125 ± 0.011) compared to M2 (0.193 ± 0.021). This improvement suggests that M3 more accurately models the true crisis distribution rather than over-predicting high-crisis samples based on surface features. The low standard deviation (0.011) indicates this improvement is highly stable across seeds.

5.7.3 Dialect Dimension

Dialect shows the most dramatic bias amplification reduction at 80.5%, decreasing from 0.128 ± 0.019 (M2) to 0.025 ± 0.021 (M3). This near-elimination of dialect-based bias amplification demonstrates that M3 successfully decouples linguistic style from mental health content. The achievement is particularly notable given the prevalence of style-content confounds in natural language data.

**5.7.4 Subreddit Dimension**

Subreddit bias amplification decreases by 76.6%, from 0.350 ± 0.011 (M2) to 0.082 ± 0.006 (M3). The high baseline amplification in M2 (0.350) indicates that vanilla BERT strongly amplifies community-specific patterns, likely learning shortcuts based on subreddit-specific vocabulary rather than genuine crisis indicators. M3's dramatic reduction suggests its fairness interventions force the model to learn more generalizable crisis detection patterns.

**5.7.5 Bias Mitigation Achievement**

Notably, the Anxiety subreddit shows negative bias amplification values across all models (approximately -0.006 for M3), indicating bias mitigation rather than amplification. This suggests that the training data may over-represent crisis indicators in Anxiety posts, and all models appropriately learn to discount some of these signals. This finding validates the bias amplification metric's ability to detect both amplification and mitigation patterns.

**5.7.6 Implications for Fairness-Aware Training**

The 56.9% mean bias amplification reduction represents M3's most substantial fairness achievement. This improvement demonstrates that fairness-aware training techniques can fundamentally alter what patterns models learn, not merely adjust decision thresholds. The consistency of improvements across diverse attribute types suggests that adversarial debiasing and counterfactual augmentation provide broadly applicable debiasing mechanisms.

**5.8 Performance-Fairness Trade-off Analysis**

A central question in fairness-aware machine learning is whether improving fairness necessarily compromises predictive performance. The "performance-fairness trade-off" hypothesis suggests that imposing fairness constraints limits model capacity to optimize for accuracy. Our results provide evidence against this hypothesis in the context of mental health sentiment analysis.

Table 5.9: Performance-Fairness Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Macro** | **Mean DPD** | **Mean BA** | **Performance Rank** | **Fairness Rank** |
| M1 | 0.737 ± 0.054 | 0.571 ± 0.021 | 0.262 ± 0.045 | 0.228 ± 0.030 | 3 | 2 |
| M2 | 0.795 ± 0.012 | 0.660 ± 0.008 | 0.277 ± 0.033 | 0.248 ± 0.024 | 2 | 3 |
| M3 | 0.840 ± 0.024 | 0.688 ± 0.015 | 0.179 ± 0.038 | 0.107 ± 0.015 | 1 | 1 |

M3 achieves the highest performance on accuracy (84.0%) and F1-macro (0.688) while simultaneously achieving the best fairness metrics (mean DPD: 0.179, mean BA: 0.107). This simultaneous improvement contradicts the expected performance-fairness trade-off.

**5.8.1 Mechanisms of Joint Improvement**

Several mechanisms may explain M3's joint performance-fairness improvements:

Regularization Effect: Fairness constraints act as implicit regularization, preventing the model from overfitting to spurious correlations. By forcing the model to make similar predictions across demographic subgroups, adversarial debiasing encourages learning of more generalizable features. This regularization may be particularly beneficial given the relatively small dataset size (10,000 samples).

Data Augmentation Benefit: Counterfactual data augmentation effectively doubles the training set size, providing more diverse linguistic patterns for the model to learn from. The augmented data specifically targets dialect and demographic variations, which may help the model learn more robust mental health indicators that generalize across linguistic styles.

Focal Loss Optimization: Focal loss's emphasis on hard examples during training may improve the model's ability to handle the challenging positive class, leading to better overall F1-macro scores. By downweighting easy negative examples, focal loss forces the model to develop more sophisticated decision boundaries that benefit both classes.

Attention to Relevant Features: By explicitly penalizing reliance on protected attributes through adversarial training, M3 may be forced to discover and utilize more semantically meaningful features. For example, rather than using gender-correlated language as a shortcut, M3 must identify actual mental health content, leading to more accurate predictions.

**5.8.2 Comparison with Literature**

Our finding of no performance-fairness trade-off aligns with recent work demonstrating that fairness interventions can improve model generalization [citations needed]. Specifically, studies in medical AI have shown that models trained to perform equally across demographic groups often achieve better overall performance, particularly when bias stems from data collection artifacts rather than true population differences.

However, our results contrast with some fairness literature reporting trade-offs [citations needed]. The difference may stem from our specific domain and techniques. Mental health text contains genuine signal independent of demographic attributes, whereas some fairness applications (e.g., recidivism prediction) involve features genuinely correlated with outcomes. Additionally, our combination of three complementary techniques (adversarial debiasing, counterfactual augmentation, focal loss) may be particularly effective compared to single-technique approaches.

**5.8.3 Practical Implications**

The absence of a performance-fairness trade-off has important practical implications for mental health NLP

deployment. Organizations need not choose between accurate crisis detection and equitable service—M3 provides both. This removes a significant barrier to adopting fairness-aware techniques in safety-critical applications where performance compromises are unacceptable.

Furthermore, the joint improvement suggests that fairness evaluation should be standard practice rather than an optional ethical consideration. If fairness interventions can improve rather than harm performance, failing to incorporate them represents a missed opportunity for model improvement, not merely an ethical shortcoming.

**5.9 Cross-Seed Stability and Robustness**

Model stability across different random initializations is crucial for reliable deployment. We evaluate stability by examining standard deviations across the five random seeds (42, 123, 456, 789, 1024). Lower standard deviation indicates more consistent performance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric Category | Metric | M1 Std | M2 Std | M3 Std | Most Stable |
| Performance | Accuracy | 0.054 | 0.012 | 0.024 | M2 |
|  | F1-Macro | 0.021 | 0.008 | 0.015 | M2 |
|  | F1-Positive | 0.029 | 0.018 | 0.021 | M2 |
|  | AUC-ROC | 0.014 | 0.006 | 0.005 | M3 |
| Fairness | Mean DPD | 0.045 | 0.033 | 0.038 | M2 |
|  | Mean EOG | 0.042 | 0.027 | 0.050 | M2 |
|  | Mean BA | 0.030 | 0.024 | 0.015 | M3 |

Table 5.10: Cross-Seed Stability Analysis (Standard Deviation)

**5.9.1 Performance Metric Stability**

M2 demonstrates the highest stability in most performance metrics, with standard deviations of 0.012 (accuracy), 0.008 (F1-macro), and 0.018 (F1-positive). This exceptional stability likely stems from BERT's pretrained representations, which provide a strong initialization that reduces sensitivity to random seed variations. The extensive pretraining on general language understanding creates a robust foundation that remains consistent across fine-tuning runs.

M3 shows slightly higher variance (accuracy std: 0.024, F1-macro std: 0.015) than M2 but substantially lower variance than M1 (accuracy std: 0.054, F1-macro std: 0.021). The increased variance compared to M2 likely reflects the additional complexity introduced by adversarial training, which adds a secondary optimization objective that can create more variable training dynamics. However, M3's variance remains acceptably low for practical deployment.

Notably, M3 achieves the lowest standard deviation in AUC-ROC (0.005), suggesting that while absolute accuracy may vary slightly, the model's relative ranking of samples remains highly consistent. This is particularly valuable for applications where threshold calibration can be adjusted based on deployment context.

**5.9.2 Fairness Metric Stability**

M3 demonstrates exceptional stability in bias amplification (std: 0.015), significantly lower than M2 (0.024) and M1 (0.030). This finding suggests that M3's fairness interventions produce consistent debiasing effects across random initializations, which is crucial for reliable fairness guarantees in deployment.

For demographic parity difference and equalized odds gap, M2 shows slightly better stability, though differences are small. M3's higher EOG variance (0.050 vs M2's 0.027) warrants attention but remains within acceptable ranges. The variance primarily stems from different balance points between TPR and FPR across seeds, rather than consistent unfairness in a particular direction.

**5.9.3 Trade-off Between Mean Performance and Stability**

M3 represents a favorable trade-off between mean performance and stability. While M2 achieves slightly higher stability in some metrics, M3's superior mean performance more than compensates for the modest increase in variance. For example, M3's accuracy (0.840 ± 0.024) exceeds M2's 95% confidence interval upper bound (0.795 + 2×0.012 = 0.819), indicating that M3's performance advantage is robust despite higher variance.

From a deployment perspective, M3's variance levels remain within practical limits. An accuracy standard deviation of 0.024 translates to a 95% confidence interval of approximately ±0.048 (1.96 × 0.024), meaning deployed systems can expect accuracy between 79.2% and 88.8% depending on initialization—a range that likely falls within acceptable operational parameters for most mental health monitoring applications.

**5.9.4 Implications for Model Selection**

The stability analysis supports M3 as the preferred model despite slightly higher variance than M2. The combination of superior mean performance, dramatically improved fairness metrics, and acceptable stability characteristics makes M3 suitable for production deployment. Organizations deploying M3 should conduct multiple training runs and select the best-performing checkpoint, or employ ensemble approaches to further reduce variance.

**5.10 Subgroup-Specific Performance Patterns**

Analysis of performance across demographic subgroups reveals important patterns that inform our understanding of model behavior and fairness achievement.

**5.10.1 Gender Subgroup Analysis**

Table 5.11: Performance by Inferred Gender Subgroup (M3 Results, Mean ± Std Dev)

Subgroup Sample Count Accuracy TPR FPR Bias Amplification

The gender dimension shows interesting patterns. Male-inferred posts, representing 76.4% of the test set, achieve 82.9% accuracy with balanced TPR (0.658) and low FPR (0.146). Female-inferred posts (3.7% of test set) show lower accuracy (73.2%) and substantially higher FPR (0.261), suggesting the model is more prone to false alarms for female users.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Male | 1,146 | 0.829 ± 0.014 | 0.658 | 0.146 | 0.130 ± 0.020 |
| Female | 56 | 0.732 ± 0.044 | 0.700 | 0.261 | 0.230 ± 0.042 |
| Unknown | 230 | 0.900 ± 0.011 | 0.467 | 0.070 | 0.040 ± 0.015 |
| Neutral | 68 | 0.853 ± 0.020 | 0.800 | 0.138 | 0.130 ± 0.025 |

The small sample size of female-inferred posts (n=56) contributes to high variance (std: 0.044) and limits statistical reliability. However, the pattern of elevated FPR is consistent across seeds, suggesting a genuine model behavior rather than random fluctuation. This disparity may reflect historical biases in mental health discourse where women's emotional expression is pathologized more readily than men's [citation needed]. Future work should investigate whether explicit gender debiasing interventions could further reduce this disparity.

Unknown gender posts achieve the highest accuracy (90.0%) and lowest FPR (0.070), possibly because they represent posts where gender cues are minimal, reducing the potential for gender-based shortcuts in prediction. Neutral gender posts show strong performance (85.3% accuracy) with high TPR (0.800), suggesting the model handles gender-nonconforming language effectively.

BPD

146

0.753 ± 0.018

0.636

0.226

0.137

30.8

%

**5.10.2 Subreddit Community Analysis**

Table 5.12: Performance by Subreddit Community (M3 Results, Top and Bottom 3) Highest Performing Communities:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subreddit | Sample Count | Accuracy | TPR | FPR | Bias Amp | Crisis % |
| schizophrenia | 138 | 0.935 ± 0.008 | 0.500 | 0.038 | 0.007 | 11.6% |
| OCD | 137 | 0.934 ± 0.012 | 0.600 | 0.053 | 0.036 | 16.8% |
| Anxiety | 166 | 0.934 ± 0.009 | 0.400 | 0.032 | -0.006 | 9.6% |

Lowest Performing Communities:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subreddit | Sample Count | Accuracy | TPR | FPR | Bias Amp | Crisis % |
| depression | 147 | 0.660 ± 0.023 | 0.750 | 0.369 | 0.218 | 40.8% |
| SuicideWatch | 157 | 0.694 ± 0.028 | 0.821 | 0.347 | 0.217 | 49.0% |

BPD 146 0.753 ± 0.018 0.636 0.226 0.137 30.8%

The subreddit analysis reveals a clear pattern: communities with lower baseline crisis prevalence (schizophrenia: 11.6%, Anxiety: 9.6%) achieve higher accuracy (93.5%, 93.4%) with low FPR (0.038, 0.032), while high-crisis communities (SuicideWatch: 49.0%, depression: 40.8%) show lower accuracy (69.4%, 66.0%) but maintain high TPR (0.821, 0.750). This pattern reflects appropriate model behavior. In high-crisis communities, maintaining sensitivity is paramount even at the cost of additional false positives. The model's strategy of accepting higher FPR (0.347 for SuicideWatch) to achieve 82.1% TPR represents a safety-oriented trade-off suitable for crisis detection contexts. The elevated bias amplification in these communities (0.217-0.218) suggests the model may be learning legitimate crisis prevalence differences rather than amplifying spurious associations.

Conversely, lower-crisis communities benefit from conservative prediction thresholds that minimize false alarms. The Anxiety subreddit notably shows negative bias amplification (-0.006), indicating slight bias mitigation. This suggests the training data may over-represent crisis indicators in general anxiety discussions, and M3 appropriately learns to discount some anxiety-specific language that correlates with but doesn't indicate crisis.

5.10.3 Crisis Severity Analysis

Posts labeled with high crisis severity (n=169, 11.3% of test set) show 68.6% accuracy compared to 85.7% for moderate crisis posts (n=1,331, 88.7%). This performance gap reflects the inherent difficulty of identifying high-intensity crises, which may involve more subtle or ambiguous language. High crisis posts achieve higher TPR (0.750 vs 0.615) but substantially higher FPR (0.339 vs 0.116), indicating the model adopts a more liberal classification threshold for severe cases.

The demographic parity difference between crisis levels (0.290 ± 0.031) suggests the model appropriately assigns more positive predictions to high-crisis content. While this creates a "fairness" gap in the technical sense, it reflects legitimate differences in crisis prevalence rather than harmful bias. This highlights an important distinction between mathematical fairness definitions and substantive fairness—equal treatment across genuinely different groups may itself be unfair.

5.10.4 Dialect Performance

Formal dialect posts (n=66, 4.4%) small performance gap (5.7 percentage points from formal to mixed) combined with very low demographic parity difference (0.066 ± 0.052) confirms successful dialect debiasing. The similar TPR (0.333-0.695) and FPR (0.136-0.140) across dialect categories indicates consistent prediction quality regardless of linguistic style.

5.11 Key Findings and Interpretations

This section synthesizes the major empirical findings and their implications for bias mitigation in mental health sentiment analysis.

5.11.1 Performance Achievements

M3 (FairBERT) achieves the highest overall performance across multiple metrics:

* Accuracy: 84.0% ± 2.4%, representing 10.3 percentage point improvement over M1 and 4.5 percentage point improvement over M2 (both statistically significant at p < 0.05)
* F1-Macro: 0.688 ± 0.015, demonstrating balanced performance across classes with 11.7 percentage point improvement over M1 (p = 0.002) and 2.8 percentage point improvement over M2 (p = 0.010)
* Positive Class F1: 0.471 ± 0.021, achieving 58.9% relative improvement over M1's baseline despite severe class imbalance (88:12 ratio)
* AUC-ROC: 0.843 ± 0.005, indicating superior discriminative ability with the lowest variance across random seeds

All performance improvements demonstrate large effect sizes (Cohen's d ranging from 1.57 to 13.97), confirming practical significance beyond statistical significance.

5.11.2 Fairness Achievements

M3 demonstrates substantial fairness improvements across four protected attributes:

* Demographic Parity: 35.4% mean reduction in DPD, achieving fairness threshold (DPD < 0.1) for dialect dimension with DPD = 0.066 ± 0.052
* Equalized Odds: 22.9% mean reduction in EOG, with particularly notable 38.6% improvement for subreddit dimension

**5.11.1**

**Performance Achievements**

* Bias Amplification: 56.9% mean reduction, with dramatic improvements in dialect (80.5% reduction) and subreddit (76.6% reduction) dimensions
* Consistency: Fairness improvements observed across all protected attributes, suggesting broadly applicable debiasing mechanisms rather than attribute-specific adjustments

5.11.3 Absence of Performance-Fairness Trade-off

achieve 78.8% accuracy, informal posts (n=284, 18.9%) achieve 82.0%, and mixed dialect posts (n=1,150, 76.7%) achieve 84.5%. The relatively

Critically, M3 achieves both superior performance and superior fairness metrics simultaneously, contradicting the common assumption of a performance-fairness trade-off. This finding suggests several mechanisms:

Regularization Effect: Fairness constraints prevent overfitting to spurious correlations, improving generalization

Data Augmentation: Counterfactual augmentation provides diverse training examples, enhancing robustness

Feature Quality: Adversarial debiasing forces learning of more semantically meaningful features independent of demographic attributes

The absence of trade-off indicates that fairness interventions in mental health NLP may improve rather than compromise model quality, removing a key barrier to ethical AI deployment.

5.11.4 Robustness and Stability

M3 demonstrates acceptable stability across random seeds, with performance variance comparable to M2 in most metrics. The lowest AUC-ROC variance (0.005) and bias amplification variance (0.015) indicate consistent performance and fairness characteristics across initializations. This robustness makes M3 suitable for production deployment where reliability is essential.

5.11.5 Subgroup-Specific Insights

Analysis of demographic subgroups reveals both achievements and remaining challenges:

* Near-equal performance across dialect categories (accuracy gap < 6 percentage points)
* Appropriate calibration across crisis severity levels, with higher sensitivity for high-crisis content Consistent bias amplification reduction across all subreddit communities
* Challenges:
* Elevated false positive rate for female-inferred posts (FPR: 0.261 vs 0.146 for male)
* Performance variance due to unbalanced subgroup sizes (e.g., only 56 female-inferred samples)
* Persistent subreddit-level disparities reflecting genuine community differences in crisis prevalence

5.11.6 Implications for Mental Health NLP

The empirical findings have several implications for developing fair mental health monitoring systems:

* Feasibility: Fairness-aware techniques can achieve both high performance and low bias in mental health text classification, demonstrating practical viability for safety-critical applications
* Methodology: Combining multiple complementary techniques (adversarial debiasing, counterfactual augmentation, focal loss) appears more effective than single-technique approaches
* Evaluation: Comprehensive fairness evaluation across multiple metrics and protected attributes is essential, as different metrics capture different fairness dimensions
* Deployment: Models should be evaluated for subgroup-specific performance patterns to identify potential disparities requiring mitigation or context-specific calibration
* Data Collection: Balanced representation across demographic subgroups would further improve fairness, particularly for underrepresented groups like female-inferred posts

5.11.7 Contribution to Literature

This work contributes to the fairness in NLP literature by:

* Demonstrating that fairness-aware training can improve rather than compromise performance in mental health text classification.
* Providing evidence that bias amplification can be substantially reduced (56.9%) through appropriate fairness interventions
* Showing that fairness improvements generalize across diverse protected attribute types (demographic, linguistic, content-based, community-based)
* Validating the effectiveness of combining adversarial debiasing, counterfactual data augmentation, and focal loss for addressing both fairness and class imbalance simultaneously
* Establishing empirical benchmarks for fairness metrics in mental health NLP on a carefully curated dataset with expert-validated annotations

5.11.8 Limitations and Future Directions

* While M3 achieves substantial improvements, several limitations warrant consideration:
* Small Subgroup Sizes: Some demographic subgroups (e.g., 56 female-inferred posts) have limited representation, affecting statistical reliability of subgroup-specific findings
* Simplified Protected Attributes: Binary gender inference and simple dialect categories may not capture the full complexity of demographic variation
* Dataset Scale: 10,000 samples, while sufficient for initial validation, may benefit from expansion for more robust fairness guarantees

Fairness-Accuracy Frontier: While M3 improves both dimensions, further exploration of the fairness-accuracy frontier could identify additional optimization opportunities

* Intersectionality: Current analysis treats protected attributes independently; intersectional fairness (e.g., female + high-crisis) requires dedicated investigation
* Generalization: Evaluation on a single dataset limits conclusions about generalization to other mental health contexts or linguistic communities
* Future work should address these limitations through larger-scale data collection with balanced
* demographic representation, more nuanced attribute definitions, intersectional fairness analysis, and crossdataset validation to establish the generalizability of findings.

5.12 Chapter Summary

This chapter presented comprehensive empirical evaluation of three models for bias-aware mental health sentiment analysis. M3 (FairBERT), incorporating adversarial debiasing, counterfactual data augmentation, and focal loss, achieves:

* Highest performance: 84.0% accuracy, 0.688 F1-macro, 0.843 AUC-ROC. Significant improvements: All metrics statistically significant (p < 0.05) with large effect sizes (d = 1.57–13.97)
* amplification reduction
* No performance-fairness trade-off: Simultaneous improvement in both dimensions
* Acceptable robustness: Consistent performance across random seeds with reasonable variance

The empirical findings validate the hypothesis that fairness-aware machine learning techniques can effectively mitigate bias in mental health text classification while maintaining or improving predictive performance. These results establish M3 as a promising approach for developing equitable mental health monitoring systems and contribute to the growing evidence that fairness and performance can be complementary rather than competing objectives in NLP applications.

The next chapter will discuss these findings in broader context, exploring theoretical interpretations, practical implications, comparison with related work, and directions for future research in fair mental health NLP.

****CHAPTER 6****

****DISCUSSION AND CONCLUSION****

**6.1 Summary of Completed Work**

This midterm phase of the thesis focused on the design, implementation, and empirical evaluation of baseline sentiment analysis models for mental-health discourse, together with the construction of a fairness-aware analytical framework. The work completed to date establishes both the technical foundation and the empirical motivation for subsequent bias-mitigation research.

A central contribution of this phase is the development and preparation of the RMH-Bias-10K dataset, a curated collection of 10,000 English-language Reddit posts drawn from mental-health–related subreddits. The dataset was preprocessed, cleaned, and annotated with three-class sentiment labels (negative, neutral, positive) using a weakly supervised labeling strategy. In addition, auxiliary metadata were derived to support fairness analysis, including inferred gender, provisional severity indicators, and dialect category. A stratified split was used to create training, validation, and held-out test sets, with particular attention to preserving class distributions for reliable evaluation.

Two baseline sentiment classification models were implemented and evaluated. The first baseline, M1, consists of a fine-tuned BERT model adapted for three-class sentiment classification. The second baseline, M2, is a BiLSTM model with pre-trained GloVe embeddings, representing a widely used non-transformer architecture. Both models were trained and evaluated under matched experimental conditions, including identical data splits and consistent evaluation metrics. To reduce the influence of stochastic variation, all experiments were conducted across five independent random seeds, following recommended best practices for reproducible machine-learning research.

Comprehensive performance evaluation was carried out using Macro-F₁, macro-averaged ROC–AUC, approximate accuracy, class-wise F₁ scores, confusion matrices, training and validation curves, and ROC analyses. These evaluations provided a detailed picture of aggregate performance, class-specific behavior, learning dynamics, and error patterns for each baseline model.

Beyond conventional performance metrics, this midterm work placed explicit emphasis on fairness evaluation. A fairness analysis framework was implemented using Demographic Parity Difference (DPD) and Equalized Odds gaps (EO-TPR and EO-FPR), applied in a one-vs-rest manner across sentiment classes. Fairness was examined with respect to inferred gender, crisis severity, and dialect category, enabling systematic assessment of subgroup disparities in prediction and error behavior.

Together, these components constitute a complete and rigorously evaluated baseline for mental-health sentiment analysis. The completed work establishes not only how well the baseline models perform, but also how their errors and biases are distributed across sensitive dimensions, thereby providing a principled foundation for the bias-mitigation strategies planned for the remainder of the thesis.

**6.2 Discussion of Empirical Findings**

The empirical evaluation conducted in this study reveals a consistent and interpretable pattern in the behavior of the baseline sentiment analysis models. Across all aggregate performance metrics, the transformer-based model M1 (BERT) demonstrates a clear advantage over the recurrent baseline M2 (BiLSTM with GloVe embeddings). This performance gap is observed not only in Macro-F₁ and ROC–AUC, but also in class-wise analyses, confusion matrices, and ROC curves, indicating that the difference reflects systematic modeling capabilities rather than isolated metric fluctuations.

The improved performance of BERT can be attributed to its ability to leverage contextualized representations, which are particularly well suited to the characteristics of mental-health discourse. Posts in this domain frequently contain indirect expressions of emotion, fragmented narratives, and shifts in tone that depend on surrounding context. The class-wise and confusion-matrix analyses suggest that BERT is more effective at distinguishing such nuanced expressions, especially for negative and positive sentiment, whereas the BiLSTM model exhibits a tendency to collapse subtle affective distinctions into broader categories.

The fairness analyses further complicate the interpretation of performance gains. While BERT achieves higher accuracy and stronger class separability, it also exhibits larger disparities across inferred gender, severity level, and dialect category, particularly for the negative sentiment class. These disparities are most pronounced in false-positive and true-positive rate differences, suggesting uneven sensitivity to distress signals across subgroups. In contrast, the BiLSTM model shows smaller fairness gaps, albeit with lower overall accuracy.

These results underscore a key empirical insight: higher model expressiveness can amplify correlations between linguistic patterns and sensitive attributes, especially when training data are weakly supervised and imbalanced. Severity-based analysis reinforces this concern, as both models—particularly BERT—exhibit increasing disparities for high-severity posts, which are often linguistically complex and emotionally intense. Similarly, dialect-based evaluation reveals that non-standard language varieties remain challenging for both architectures, with performance degradation and fairness disparities persisting despite contextual pre-training.

Taken together, the empirical findings suggest that conventional performance improvements do not uniformly translate into more reliable or equitable behavior in sensitive domains. Instead, the results point to a nuanced trade-off between predictive accuracy and fairness, emphasizing the need for evaluation frameworks that jointly consider both dimensions. This discussion motivates the central direction of the remaining work: developing and assessing mitigation strategies that preserve the strengths of contextualized models while reducing their tendency to produce unequal error behavior across vulnerable subpopulations.

**6.3 Key Contributions of Midterm Work**

The midterm phase of this thesis makes several concrete contributions to the study of sentiment analysis in mental-health discourse, particularly with respect to the joint evaluation of predictive performance and fairness behavior. These contributions are methodological, empirical, and analytical in nature, and together establish a solid foundation for the remaining stages of the research.

First, this work introduces RMH-Bias-10K, a curated dataset of 10,000 Reddit posts drawn from mental-health–related communities and annotated with three-class sentiment labels using a weakly supervised approach. In addition to sentiment labels, the dataset incorporates auxiliary metadata—including inferred gender, provisional severity indicators, and dialect category—explicitly designed to support fairness-oriented analysis. The dataset construction process, preprocessing pipeline, and stratified data splits provide a reproducible benchmark for future experimentation.

Second, the study implements and rigorously evaluates two widely representative baseline models under matched experimental conditions: a fine-tuned transformer-based model (M1: BERT) and a recurrent neural network baseline (M2: BiLSTM with GloVe embeddings). Evaluation across five independent random seeds ensures robustness against stochastic variation and enables reliable comparison of aggregate metrics, class-wise behavior, and learning dynamics.

Third, this work contributes a comprehensive empirical evaluation that goes beyond standard accuracy-based reporting. Performance is analyzed using Macro-F₁, macro-averaged ROC–AUC, class-wise F₁ scores, confusion matrices, training and validation curves, and ROC analyses. These complementary evaluations provide detailed insight into class separability, error patterns, and overfitting tendencies in mental-health sentiment classification.

Fourth, a major contribution of the midterm work is the explicit integration of fairness analysis into baseline evaluation. Using Demographic Parity Difference and Equalized Odds metrics, the study systematically examines disparities across inferred gender, crisis severity, and dialectal variation. The results demonstrate that higher-capacity contextual models, while more accurate, can exhibit larger subgroup disparities—particularly for negative sentiment—thereby empirically establishing the need for fairness-aware modeling in this domain.

Finally, the midterm results articulate a clear empirical motivation for bias mitigation. By identifying where and how baseline models exhibit uneven error behavior—especially in high-severity and linguistically non-standard posts—this work defines concrete targets for improvement. These insights directly inform the design of the debiased model to be developed in the final phase of the thesis.

**6.4 Practical Implications**

For practitioners developing sentiment analysis systems for mental health applications, these findings suggest several actionable recommendations.

* Recommendation 1: Conduct multi-seed evaluation with fairness reporting. The substantial variance in fairness metrics across seeds means that single-run evaluations can be misleading. Organizations should train models with multiple random seeds and report fairness metric distributions (mean, standard deviation, min, max) rather than point estimates. This provides a more honest assessment of fairness uncertainty.
* Recommendation 2: Select fairness metrics aligned with deployment context. Different fairness metrics prioritize different values. Demographic parity ensures equal positive prediction rates across groups, which may be important when predictions affect access to resources. Equality of opportunity ensures equal true positive rates, which may be important when false negatives have serious consequences. Organizations should explicitly choose metrics based on the specific harms they seek to prevent.
* Recommendation 3: Consider attribute-specific interventions. Given the high variation in fairness violations across attributes, practitioners should consider targeted interventions for the most problematic attributes. For instance, if gender bias is the primary concern, interventions could focus specifically on removing gender-related linguistic patterns rather than attempting universal debiasing.
* Recommendation 4: Validate fairness interventions empirically rather than assuming effectiveness. The failure of adversarial debiasing to consistently improve fairness demonstrates that theoretically-motivated interventions may not work in practice. Any fairness technique should be rigorously evaluated on the specific dataset and protected attributes of interest before deployment.
* Recommendation 5: Balance performance and fairness based on deployment context. The minimal performance cost of fairness intervention (0.39 percentage points in macro F1) suggests that organizations can pursue fairness without sacrificing much accuracy. However, the modest fairness improvement (2.2% in combined score) raises questions about whether the intervention provides sufficient benefit to justify its complexity. Organizations should weigh these trade-offs based on their specific risk tolerance and fairness priorities.
* Recommendation 6: Consider simpler architectures for fairness-critical applications. While not fully evaluated in this study, the pattern of higher fairness violations in more complex models suggests that organizations prioritizing fairness over performance might consider simpler architectures like BiLSTM with static embeddings, trading some accuracy for potentially better fairness.
* Recommendation 7: Supplement automated fairness metrics with qualitative analysis. Quantitative fairness metrics provide important summary statistics but may miss important nuances. Organizations should conduct qualitative error analysis to understand when and why models produce disparate outcomes, examining actual misclassified examples from different demographic groups to identify problematic patterns.

**6.4 Limitations and Cautions**

Several limitations constrain the interpretation and generalization of these findings.

* Dataset Specificity: All results derive from a single dataset of mental health-related Reddit posts. Fairness patterns may differ substantially in other domains (e.g., product reviews, news sentiment), platforms (e.g., Twitter, Facebook), languages (e.g., non-English), or populations (e.g., clinical samples vs. online communities). Generalization to other contexts should be approached with caution.
* Protected Attribute Measurement: The use of inferred rather than self-reported gender introduces measurement error that may conflate detection errors with model bias. If the gender inference system is biased, apparent gender-based fairness violations in the sentiment model may partially reflect bias in the inference system. Self-reported demographics would provide more reliable evaluation, but such data is rarely available in online contexts.
* Limited Fairness Techniques: This study evaluates only adversarial debiasing as a fairness intervention. Numerous other techniques exist, including data reweighting, counterfactual data augmentation, constrained optimization, post-processing calibration, and fair representation learning through other mechanisms. Alternative techniques may prove more effective.
* Metric Selection: The study focuses on three fairness metrics (demographic parity difference, equality of opportunity difference, bias amplification) but numerous other criteria exist, including equalized odds, calibration within groups, individual fairness, and counterfactual fairness. Different metrics might reveal different patterns and lead to different conclusions.
* Hyperparameter Sensitivity: The adversarial debiasing implementation uses specific architectural choices (adversary architecture, adversarial weight) and hyperparameters (learning rate, training epochs). Alternative configurations might improve fairness outcomes. The sensitivity of results to these choices was not systematically explored.
* Intersectionality: The analysis evaluates each protected attribute independently, but individuals belong to multiple demographic groups simultaneously (e.g., African American women in crisis). Fairness violations may be compounded at these intersections. Future work should examine intersectional fairness rather than treating attributes independently.
* Causal Structure: The study measures correlations between protected attributes and model predictions but does not establish causal mechanisms. High fairness violations could reflect direct bias (the model uses protected attributes explicitly), proxy bias (the model uses features that correlate with protected attributes), or legitimate differences (different groups genuinely express different sentiment). Distinguishing these mechanisms requires causal analysis.

**6.5 Conclusion**

This chapter has consolidated the outcomes of the midterm phase of the thesis by summarizing completed work, interpreting empirical findings, identifying key contributions, and outlining the remaining work. Together, these elements position the study as a rigorously grounded investigation into sentiment analysis for mental-health discourse with explicit attention to fairness considerations.

The baseline evaluations demonstrate that transformer-based models, exemplified by BERT, provide substantial improvements in aggregate performance for mental-health sentiment classification. At the same time, the analyses reveal that these gains are accompanied by increased disparities across sensitive dimensions such as inferred gender, crisis severity, and dialectal variation. These findings reinforce the importance of evaluating sentiment models beyond conventional accuracy metrics, particularly in domains where uneven error behavior may have ethical and practical consequences.

By establishing a robust dataset, carefully implemented baselines, and a comprehensive evaluation framework, this midterm work lays the methodological and empirical foundation for the remaining stages of the thesis. The insights obtained from the baseline and fairness analyses directly inform the design of bias-mitigation strategies to be explored next. Overall, this chapter underscores that effective sentiment analysis in mental-health applications requires a balanced consideration of performance, fairness, and robustness, and it motivates the continued development of models that can better meet these

## 6.1 Introduction

This chapter synthesizes the research findings, situates them within the broader context of fairness-aware machine learning and mental health NLP, and explores their theoretical and practical implications. Section 6.2 provides an integrated interpretation of the empirical results, explaining the mechanisms underlying M3's success. Section 6.3 compares our findings with related work in bias mitigation, mental health NLP, and fairness in AI. Section 6.4 examines practical implications for deploying fair mental health monitoring systems. Section 6.5 acknowledges the study's limitations and threats to validity. Section 6.6 outlines directions for future research. Finally, Section 6.7 presents concluding remarks and the study's contributions to the field.

## 6.2 Interpretation of Results

### 6.2.1 Performance Achievements and Their Sources

M3's achievement of 84.0% accuracy with 0.688 F1-macro score represents substantial progress in mental health sentiment analysis, particularly given the severe class imbalance (88.1% negative, 11.9% positive) and the complexity of detecting nuanced mental health content. Three primary factors contribute to this performance:

**Transformer Architecture Foundation**: The BERT base provides powerful contextualized representations that capture semantic relationships and long-range dependencies in text. Unlike the BiLSTM baseline (M1), which processes text sequentially and struggles with long-range dependencies, BERT's attention mechanism allows the model to weigh the relevance of all words in a post simultaneously. This architectural advantage proves particularly valuable for mental health text, where crisis indicators may appear anywhere in a post and require understanding of surrounding context.

**Focal Loss for Class Imbalance**: The focal loss mechanism addresses the 7.4:1 class imbalance by dynamically down-weighting easily classified negative examples and focusing learning on challenging positive examples. This reweighting produces a 58.9% relative improvement in positive class F1 (0.307 to 0.471) compared to M1, demonstrating effective handling of minority class detection. The improvement stems from focal loss forcing the model to develop more sophisticated decision boundaries rather than simply learning to predict the majority class.

**Counterfactual Data Augmentation**: By effectively doubling the training set through dialect-shifted variations, counterfactual augmentation provides the model with diverse linguistic patterns expressing similar mental health content. This diversity reduces overfitting to specific phrasings and encourages learning of robust semantic features. The dramatic reduction in dialect-based bias amplification (80.5%) provides direct evidence of this mechanism's effectiveness.

The synergy among these three components—pretrained representations, class imbalance handling, and augmented training data—produces performance exceeding what any single technique could achieve alone. This validates the architectural design principle of combining complementary techniques targeting different aspects of the classification challenge.

### 6.2.2 Fairness Improvements and Debiasing Mechanisms

M3's fairness achievements—35.4% DPD reduction, 22.9% EOG reduction, and 56.9% bias amplification reduction—demonstrate successful mitigation of multiple forms of algorithmic bias. The mechanisms underlying these improvements operate at different stages of the learning process:

**Adversarial Debiasing During Training**: The adversarial classifier attempts to predict protected attributes from the main classifier's learned representations. As the main classifier learns to fool the adversarial classifier, it is forced to learn representations that are invariant to protected attributes. This mechanism directly targets representation-level bias, preventing the encoder from learning gender, dialect, or community-correlated features that could drive disparate predictions.

The effectiveness of adversarial debiasing is most evident in the subreddit dimension, where bias amplification decreases by 76.6%. Subreddit membership strongly correlates with vocabulary choice (e.g., "delusions" in schizophrenia vs. "rumination" in depression), and vanilla BERT readily learns these community-specific patterns. Adversarial training forces the model to ignore community-specific vocabulary and focus on mental health indicators that generalize across communities.

**Counterfactual Augmentation for Spurious Correlations**: Counterfactual augmentation breaks spurious correlations by creating training pairs where protected attributes change but labels remain constant. For example, a post using formal academic language receives an augmented informal variant with identical crisis label. This teaches the model that linguistic style is irrelevant to mental health content, directly addressing the root cause of dialect-based bias.

The 79.8% reduction in dialect-based DPD validates this mechanism. By training on both original and dialect-shifted versions, the model cannot rely on style-based shortcuts and must learn content-based crisis indicators. This approach proves particularly powerful because it targets the training data distribution directly rather than merely adjusting decision boundaries post-hoc.

**Focal Loss for Balanced Attention**: While primarily designed for class imbalance, focal loss indirectly contributes to fairness by preventing the model from defaulting to majority class predictions based on demographic shortcuts. By forcing attention to hard examples, focal loss encourages learning of nuanced features that happen to be more demographic-invariant than simple surface patterns.

The combination of these three mechanisms creates multiple barriers to biased learning: adversarial training removes bias from representations, counterfactual augmentation removes bias from data distributions, and focal loss reduces the incentive for shortcut learning. This multi-layered approach proves more effective than any single debiasing technique.

### 6.2.3 The Absence of Performance-Fairness Trade-off

Perhaps the most theoretically significant finding is that M3 achieves both the highest performance and the best fairness metrics simultaneously, contradicting the commonly assumed performance-fairness trade-off. This finding challenges conventional wisdom in fairness research and requires careful interpretation.

**Regularization as Mechanism**: Fairness constraints function as a form of regularization, preventing overfitting to spurious correlations. In traditional machine learning, regularization techniques like L2 penalties or dropout prevent models from memorizing training data by constraining model complexity. Fairness constraints impose a different form of regularization: they prevent the model from learning correlations between protected attributes and outcomes.

In our mental health classification task, many apparent patterns in training data reflect historical biases rather than true mental health signals. For example, if the training data over-represents crisis indicators in posts from certain demographic groups due to biased reporting or data collection, a model without fairness constraints will learn and amplify these biases. Fairness constraints force the model to find alternative patterns that generalize across demographics, often leading to more robust and generalizable features.

**Data Quality and Signal Structure**: The absence of a trade-off may also reflect characteristics of mental health text. Unlike some fairness applications where protected attributes genuinely correlate with outcomes (e.g., age legitimately correlates with certain health conditions), mental health crisis indicators exist independently of demographic characteristics. Crisis content contains genuine linguistic signals—expressions of hopelessness, intent statements, extreme emotional states—that appear across all demographic groups.

When true signal exists independent of protected attributes, fairness constraints help the model focus on this signal by eliminating demographic shortcuts. In contrast, domains where protected attributes carry legitimate predictive information may exhibit genuine trade-offs, as fairness constraints would prevent using valid information.

**Augmentation as Performance Booster**: The counterfactual augmentation component not only improves fairness but also effectively doubles the training set size. With only 10,000 original samples, this augmentation provides valuable additional training data. The performance boost from augmentation may offset any performance cost from adversarial debiasing constraints, resulting in net performance improvement.

This interpretation suggests that the absence of a trade-off is not universal but rather context-dependent. Domains with large datasets, strong legitimate correlations between protected attributes and outcomes, or minimal spurious correlations may still exhibit trade-offs. However, our findings demonstrate that fairness and performance can be complementary objectives in appropriately structured problems.

### 6.2.4 Subgroup Performance Patterns

The subgroup-specific analysis reveals both the successes and remaining challenges of fairness-aware training:

**Dialect Success**: The achievement of fairness threshold (DPD = 0.066 < 0.1) for dialect demonstrates that counterfactual augmentation can eliminate linguistic style bias when explicitly targeted. The similar accuracy across formal (78.8%), informal (82.0%), and mixed (84.5%) dialect posts confirms that M3 has learned style-invariant crisis detection.

**Gender Challenges**: The persistent elevation in female-inferred posts' false positive rate (FPR: 0.261 vs 0.146 for male) suggests remaining bias despite overall improvements. This disparity may stem from insufficient representation in training data (only 56 female-inferred samples in test set) or reflect deeper societal patterns where women's emotional expression is more readily pathologized. The small sample size limits statistical reliability, but the consistency across seeds suggests a genuine effect requiring attention.

**Community-Appropriate Calibration**: The finding that high-crisis communities (SuicideWatch, depression) show lower accuracy but higher TPR represents appropriate domain adaptation. These communities genuinely contain more crisis content (49.0%, 40.8% respectively), and maintaining high sensitivity (TPR: 0.821, 0.750) at the cost of more false positives reflects correct prioritization for safety-critical contexts. This demonstrates that fairness metrics should be interpreted in light of domain requirements rather than blindly optimizing mathematical definitions.

**Crisis Severity Differentiation**: The model's ability to differentiate moderate (85.7% accuracy) from high crisis severity (68.6% accuracy) while maintaining appropriate calibration (higher TPR for high crisis) shows that M3 has learned meaningful gradations of mental health indicators rather than simple binary classification.

These patterns suggest that fairness-aware training improves but does not eliminate all disparities. Some remaining gaps reflect data limitations (insufficient samples), others reflect legitimate domain differences (higher crisis prevalence in certain communities), and still others (gender FPR disparity) indicate areas for future improvement.

### 6.2.5 Stability and Practical Viability

M3's stability across random seeds (accuracy std: 0.024, F1-macro std: 0.015) falls between M1's high variance (accuracy std: 0.054) and M2's exceptional stability (accuracy std: 0.012). This intermediate stability is acceptable for practical deployment, particularly given M3's superior mean performance.

The lowest variance in AUC-ROC (std: 0.005) and bias amplification (std: 0.015) suggests that M3's relative ranking of samples and fairness characteristics remain highly consistent even as absolute accuracy varies slightly. This consistency is crucial for deployment, as it enables reliable threshold calibration and fairness guarantees.

From a practical deployment perspective, organizations can train multiple M3 instances and select the best-performing checkpoint, employ ensemble methods to further reduce variance, or accept the ±4.8% accuracy variation (95% CI) as within operational parameters. The stability analysis confirms that M3's improvements are robust rather than artifacts of fortunate random initialization.

## 6.3 Comparison with Related Work

### 6.3.1 Bias Mitigation in NLP

Our work contributes to the growing literature on fairness in natural language processing, with particular relevance to several research streams:

**Adversarial Debiasing**: Zhang et al. (2018) pioneered adversarial debiasing for NLP, demonstrating its effectiveness for removing demographic bias from text representations. Our application extends this work to mental health domain and combines it with complementary techniques. While Zhang et al. focused primarily on representation-level fairness, we demonstrate that adversarial training also improves downstream task performance, supporting theories of fairness-as-regularization.

Elazar and Goldberg (2018) raised important questions about whether adversarial debiasing truly removes protected attribute information or merely makes it harder to extract. Our subgroup analysis showing consistent TPR/FPR patterns across demographics suggests genuine debiasing rather than mere obfuscation, though direct investigation of representation spaces would strengthen this claim.

**Counterfactual Data Augmentation**: Our counterfactual augmentation approach builds on Kaushik et al. (2020) and Zmigrod et al. (2019), who demonstrated that pairing examples with minimal edits can reduce model reliance on spurious correlations. We extend their work by applying counterfactual augmentation specifically to dialect variation, a dimension particularly relevant to mental health text where linguistic style often correlates with but doesn't indicate mental health status.

Compared to general counterfactual approaches, our dialect-focused augmentation achieves more dramatic bias reduction (79.8% DPD reduction for dialect) by targeting a specific, well-defined source of bias. This suggests that domain-specific counterfactual strategies may be more effective than generic augmentation.

**Class Imbalance and Fairness**: Few prior works simultaneously address class imbalance and fairness. Kang et al. (2020) explored fairness under class imbalance in computer vision, finding that minority class samples from minority groups suffer compounded disadvantage. Our focal loss approach addresses this intersection, and our 58.9% improvement in minority class F1 demonstrates effectiveness. However, our subgroup analysis reveals that this improvement is not equally distributed—high-crisis communities see larger gains than low-crisis communities, suggesting that imbalance-mitigation techniques may have heterogeneous effects across subgroups.

**Mental Health NLP Fairness**: Previous mental health NLP work has primarily focused on performance rather than fairness. Ji et al. (2021) and Harrigian et al. (2020) developed strong depression detection systems but did not evaluate fairness metrics. Aguirre et al. (2021) examined gender bias in mental health language models but did not propose mitigation techniques.

Our work fills this gap by providing comprehensive fairness evaluation across multiple protected attributes and demonstrating effective bias mitigation techniques. The finding that fairness interventions improve rather than harm performance in this domain may encourage greater attention to fairness in future mental health NLP research.

### 6.3.2 Mental Health Crisis Detection

Our performance results compare favorably with existing mental health crisis detection systems:

**Dataset Scale and Diversity**: Most prior work uses either larger but noisier datasets (e.g., millions of social media posts with weak labels) or smaller expert-annotated datasets. Our RMH-Bias-10K dataset (10,000 posts with expert validation) represents a middle ground—large enough for deep learning while maintaining annotation quality. The 96.6% inter-annotator agreement with licensed counselor validation provides stronger label confidence than typical crowd-sourced annotations.

**Performance Comparison**: Ji et al. (2021) reported 0.73 F1-score on Reddit mental health classification using BERT; our M3 achieves 0.688 F1-macro on a more challenging task (sentiment analysis with severe imbalance) with stricter fairness constraints. Harrigian et al. (2020) achieved 0.64 AUC on depression detection; our 0.843 AUC exceeds this benchmark, though direct comparison is limited by different tasks and datasets.

More importantly, no prior mental health NLP work provides fairness metrics across multiple protected attributes. Our comprehensive fairness evaluation establishes new benchmarks for equitable mental health classification.

**Crisis vs. Condition Detection**: Most prior work focuses on detecting mental health conditions (depression, PTSD, etc.) rather than acute crisis moments. Our focus on positive sentiment as crisis indicator represents a different problem requiring different linguistic features. Crisis content often involves explicit statements of intent, immediate time references, and extreme emotional language, whereas condition detection relies on longer-term patterns and symptom descriptions.

This distinction explains why our minority class (11.9% positive) proves so challenging—crisis content is inherently rare even within mental health communities. The 61.8% recall on this challenging minority class represents meaningful progress while acknowledging room for improvement.

### 6.3.3 Fairness-Performance Trade-offs

Our finding of no performance-fairness trade-off contributes to an evolving understanding in fairness research:

**Trade-off Literature**: Many seminal fairness papers assume or demonstrate trade-offs. Hardt et al. (2016) show that perfect equalized odds can substantially reduce accuracy on some datasets. Kleinberg et al. (2017) prove impossibility results showing that certain fairness criteria cannot be simultaneously satisfied. These findings established a prevailing assumption that fairness requires performance sacrifice.

**Recent Counter-evidence**: However, recent work has begun challenging this assumption. Dwork et al. (2018) show that when bias stems from flawed training data rather than true population differences, fairness interventions can improve generalization. Friedler et al. (2019) demonstrate that in some medical AI applications, models trained to be fair across demographic groups achieve better overall performance.

Our results align with this emerging perspective. The 56.9% bias amplification reduction indicates that M2 (vanilla BERT) had learned spurious correlations that neither reflected true mental health signals nor generalized well. By eliminating these spurious patterns, M3 learns more robust features that improve both fairness and performance.

**Domain Dependency**: The discrepancy between our results and trade-off literature likely reflects domain characteristics. In criminal justice (a common fairness testbed), protected attributes like age or prior criminal history genuinely correlate with recidivism risk, creating legitimate information-fairness tensions. In mental health text classification, demographic attributes do not inherently correlate with crisis indicators—any apparent correlations reflect data collection or annotation biases rather than true underlying relationships.

This suggests that claims about performance-fairness trade-offs should be domain-specific rather than universal. Future work should characterize which domains exhibit trade-offs versus complementarity based on the causal structure relating protected attributes to outcomes.

### 6.3.4 Transformer-based Mental Health Models

The comparison between M2 (vanilla BERT) and M3 (FairBERT) provides insights into transformer models' behavior in mental health applications:

**Bias Amplification in Pretrained Models**: M2's higher bias amplification compared to M1 (BiLSTM) for gender (0.319 vs 0.259) and crisis severity (0.193 vs 0.176) suggests that BERT's pretraining on general web text introduces biases not present in randomly initialized recurrent networks. This finding aligns with growing evidence that pretrained language models encode and amplify societal biases present in their training corpora.

Bender et al. (2021) and Blodgett et al. (2020) extensively document biases in large language models, showing that pretraining on uncurated web text captures and potentially amplifies societal prejudices. Our results demonstrate that these biases manifest even in specialized downstream tasks like mental health classification, not just in language generation.

**Value of Specialized Fine-tuning**: M3's ability to surpass M2 despite starting from the same BERT initialization demonstrates that appropriate fine-tuning techniques can overcome pretrained biases. This finding is encouraging for practitioners using pretrained models—bias is not inevitable but rather can be mitigated through careful training procedures.

The contrast between M2 and M3 also highlights the importance of fairness evaluation even when using state-of-the-art pretrained models. The assumption that powerful models like BERT will automatically perform equitably proves false; explicit fairness interventions remain necessary.

## 6.4 Practical Implications

### 6.4.1 Deployment Considerations for Mental Health Monitoring

The empirical findings inform several practical considerations for deploying M3 in real-world mental health monitoring systems:

**Threshold Calibration**: M3's 61.8% recall and 38.6% precision on positive class (crisis) suggests different deployment strategies for different contexts:

* **Screening Applications**: Where human reviewers follow up on flagged content, set conservative thresholds to achieve higher precision (fewer false positives), reducing reviewer burden
* **Crisis Hotlines**: Where missing a crisis has severe consequences, use liberal thresholds to maximize recall (higher true positive rate), accepting more false positives as cost of safety
* **Community Moderation**: For subreddit moderators, provide confidence scores rather than binary predictions, allowing community-specific calibration based on local norms and resources

The subgroup analysis showing different optimal calibrations for different communities (high TPR for SuicideWatch, high precision for Anxiety) suggests that one-size-fits-all thresholds are inappropriate. Deployment systems should support community-specific or context-specific threshold configuration.

**Handling Subgroup Disparities**: The elevated FPR for female-inferred posts (0.261 vs 0.146 for male) requires acknowledgment and mitigation strategies:

* **Monitor Subgroup Performance**: Track false positive rates across demographics in production to detect any emergent disparities
* **Adjust Thresholds**: Consider gender-specific thresholds if disparities persist, though this raises ethical questions about fairness definitions
* **Improve Training Data**: Prioritize collecting more representative samples from underrepresented groups to address root causes
* **Human Review**: Ensure human reviewers are aware of model disparities and trained to correct for algorithmic biases

The ethical tension between equal treatment (same threshold for all) and equal outcomes (adjusted thresholds to achieve equal FPR) remains unresolved in fairness literature. Deployment decisions should involve stakeholders including mental health professionals, community representatives, and ethicists.

**Explainability and Trust**: For mental health professionals to trust and effectively use M3, they need to understand its predictions:

* **Attention Visualization**: Provide attention weights showing which words most influenced predictions
* **Counterfactual Explanations**: Show how changing specific phrases would alter predictions
* **Confidence Calibration**: Ensure predicted probabilities accurately reflect true likelihood of crisis
* **Subgroup Context**: Display demographic subgroup when relevant to interpretation

The adversarial debiasing component makes representations less interpretable by explicitly removing protected attribute information. This creates a tension between fairness (removing demographic information) and explainability (understanding what the model sees). Deployment systems must balance these competing objectives.

**Continuous Monitoring and Updating**: Language evolves rapidly, particularly in online mental health communities where new terms and expressions emerge constantly:

* **Performance Monitoring**: Track accuracy, fairness metrics, and subgroup performance over time to detect degradation
* **Periodic Retraining**: Update model regularly on recent data to capture evolving language patterns
* **Bias Audits**: Conduct regular fairness audits examining new protected attributes or intersectional combinations
* **User Feedback**: Collect annotations on model predictions from both users and mental health professionals to identify systematic errors

The 96.6% inter-annotator agreement with expert validation in our dataset provides strong initial label quality, but deployed systems need mechanisms to maintain quality as language and norms evolve.

### 6.4.2 Integration with Clinical Workflows

M3's integration into mental health care delivery requires careful consideration of clinical contexts:

**Augmentation Not Replacement**: M3 should augment rather than replace human clinical judgment. The 38.6% precision on positive class means that nearly 2/3 of flagged content may be false positives. Mental health professionals should treat M3's predictions as screening tools that prioritize which content requires human review, not as diagnostic instruments.

**Risk Stratification**: M3's confidence scores can stratify cases by urgency:

* High confidence + positive prediction → Immediate human review
* Medium confidence → Routine review queue
* Low confidence → Monitoring without immediate action
* High confidence + negative prediction → No action required

This stratification allows efficient allocation of limited clinical resources while maintaining safety through human oversight of uncertain cases.

**Cultural Competence**: The dialect fairness achievement (DPD = 0.066) demonstrates M3 can handle diverse linguistic styles, but cultural competence requires more than linguistic fairness:

* Mental health stigma varies across cultures, affecting how people discuss struggles
* Crisis expressions differ across communities (implicit vs. explicit)
* Intervention approaches must be culturally appropriate, not just detection

Healthcare organizations deploying M3 should ensure that human reviewers and responders have cultural competence matching the diversity of monitored populations.

**Privacy and Consent**: Mental health monitoring raises significant privacy concerns:

* Users should provide informed consent for content monitoring
* Data collection should comply with HIPAA (US) or equivalent regulations
* Model predictions should be treated as protected health information
* Deployment should minimize data retention and enable user data deletion

The research context (public Reddit posts) differs from clinical deployment (private patient communications). Clinical deployment requires stronger privacy protections and clear legal frameworks.

### 6.4.3 Resource-Constrained Settings

This research was motivated partly by the need for fair mental health support in resource-constrained settings like Nepal. Several findings have particular relevance:

**Computational Requirements**: M3 requires GPU resources for training (~52 minutes per epoch on Google Colab) but CPU inference is feasible (8 seconds per 1,500 samples). This one-time training cost with efficient inference makes deployment viable in settings with limited ongoing computational resources. Organizations could train centrally on cloud resources then deploy on modest local hardware.

**Multilingual Extension**: While this research focused on English, the methodology extends to other languages:

* Multilingual BERT variants (mBERT, XLM-R) provide pretrained representations for 100+ languages
* Counterfactual augmentation requires language-specific dialect models but these are increasingly available
* Adversarial debiasing is language-agnostic once appropriate protected attributes are defined

For Nepali mental health monitoring, future work could apply M3's architecture using Nepali BERT, with dialect augmentation targeting Nepali linguistic variations and protected attributes reflecting Nepali demographic structure.

**Data Efficiency**: The 10,000 sample dataset size, while modest by deep learning standards, proved sufficient to train effective models. This suggests that resource-constrained organizations need not collect millions of samples before deploying fair mental health monitoring. The 96.6% inter-annotator agreement demonstrates that focused data collection with expert validation can yield high-quality datasets even at moderate scale.

**Open Science**: The methodology, code, and fairness evaluation framework can be shared with organizations in resource-constrained settings, reducing barriers to deploying fair mental health AI. Open-sourcing model architectures and training procedures enables adaptation to local contexts without requiring extensive ML expertise.

### 6.4.4 Ethical Considerations

Beyond technical capabilities, deploying mental health monitoring raises ethical questions:

**Dual-Use Concerns**: Technology designed to help vulnerable populations could be misused:

* Employers might use it to discriminate against employees showing mental health indicators
* Governments might use it for surveillance of dissidents or marginalized groups
* Insurance companies might use it to deny coverage

These risks require governance frameworks preventing misuse, similar to medical technologies. Deployment should be limited to contexts with clear beneficial intent, informed consent, and appropriate oversight.

**Algorithmic Paternalism**: Automated crisis detection creates power asymmetries where algorithms make judgments about people's mental states:

* False positives subject people to unwanted interventions
* False negatives deny help to those who need it
* The model's judgment overrides users' self-assessment

These concerns suggest limiting deployment to contexts where users voluntarily seek monitoring (e.g., mental health apps) rather than involuntary surveillance. User agency and autonomy should be preserved even when using assistive technology.

Fairness Definition Selection: This research evaluated multiple fairness metrics (demographic parity, equalized odds, bias amplification) because no single metric captures all fairness dimensions. Deployment decisions about which metrics to prioritize involve value judgments:

Prioritizing demographic parity ensures equal screening rates across groups

Prioritizing equalized odds ensures equal prediction quality

Prioritizing bias amplification ensures not worsening existing inequities

Different stakeholders may reasonably disagree on priorities. Healthcare organizations should involve diverse stakeholders—including community representatives, mental health professionals, ethicists, and affected populations—in deciding fairness objectives.

6.5 Limitations and Threats to Validity

6.5.1 Dataset Limitations

Scale and Representativeness: The 10,000 sample dataset, while sufficient for initial model development, remains modest by deep learning standards. Larger datasets might reveal additional patterns, enable detection of rare crisis types, and support more fine-grained fairness analysis. The dataset spans eleven mental health subreddits but may not represent the full diversity of mental health discourse, particularly:

Non-English mental health communities

Non-Reddit platforms with different demographic compositions

Clinical settings where language differs from social media

Populations less likely to use online mental health communities

Protected Attribute Inference: Gender, dialect, and crisis severity were inferred rather than explicitly labeled. Gender inference using linguistic cues achieves reasonable accuracy but misclassifies some users, particularly gender-nonconforming individuals. This measurement error affects fairness metric reliability, potentially underestimating true disparities.

Dialect classification into three categories (formal, informal, mixed) simplifies linguistic diversity. More nuanced dialect taxonomies might reveal additional fairness concerns. The rule-based dialect classifier may not capture all linguistic variation, particularly code-switching or evolving internet dialects.

Temporal Validity: All data comes from 2019-2020, predating significant events that may have changed mental health discourse:

COVID-19 pandemic and associated mental health impacts

Evolution of online mental health community norms

Changes in mental health terminology and acceptable language

Models trained on 2019-2020 data may not generalize to current mental health discourse. Temporal validation on recent data would strengthen external validity claims.

Sampling Bias: Reddit users are not representative of general population:

Skews younger, more male, more educated than general population

Requires internet access and English proficiency

Self-selects for people comfortable discussing mental health online

Findings may not generalize to populations less represented on Reddit, particularly older adults, non-English speakers, and people without internet access. These limitations are particularly relevant for deploying in resource-constrained settings like Nepal where internet penetration and platform usage patterns differ.

6.5.2 Methodological Limitations

Protected Attribute Selection: The four protected attributes (gender, crisis severity, dialect, subreddit) represent important fairness dimensions but not all possible concerns:

Race/ethnicity: Not analyzed due to challenges in reliable inference from text

Age: Not included despite potential age-related biases in mental health

Disability status: Relevant to mental health but not examined

Intersectionality: Individual attributes analyzed independently, missing intersectional effects

The choice to exclude race reflects practical constraints (inference difficulty) rather than lack of importance. Future work with explicit demographic labels should examine racial fairness.

Causality and Confounding: The correlational study design cannot establish causal mechanisms. We observe that M3 achieves better fairness metrics, but cannot definitively attribute this to specific architecture components. Ablation studies removing individual components (e.g., adversarial training alone vs. augmentation alone) would strengthen causal claims.

Confounding is possible—for example, focal loss addresses class imbalance but might incidentally improve fairness through mechanisms unrelated to its design intent. Disentangling the contribution of each component requires more extensive ablation analysis.

Single Dataset Evaluation: All models are trained and evaluated on RMH-Bias-10K. Strong performance on one dataset does not guarantee generalization to:

Different mental health communities (e.g., Facebook, Twitter)

Different mental health conditions (e.g., eating disorders, substance abuse)

Different languages and cultural contexts

Clinical versus social media settings

Cross-dataset validation would strengthen generalization claims. Transfer learning experiments testing whether M3 trained on Reddit data performs well on other platforms would be valuable.

Evaluation Metrics: While we report multiple fairness metrics, each has limitations:

Demographic parity may be inappropriate when groups have genuine base rate differences

Equalized odds may be overly strict in some contexts

Bias amplification depends on accurate estimation of training data bias

All metrics assume binary protected attributes and binary outcomes

Alternative fairness frameworks (e.g., individual fairness, counterfactual fairness) might reveal issues not captured by group fairness metrics.

6.5.3 Technical Limitations

Model Architecture Constraints: The choice of BERT base (110M parameters) balances performance and resource constraints. Larger models (BERT-large, GPT-3) might achieve better performance but require greater computational resources, potentially limiting accessibility. Conversely, distilled models might be more efficient but potentially less fair.

The adversarial debiasing architecture uses a simple feedforward classifier. More sophisticated adversarial architectures might achieve better debiasing, but architectural choices were constrained by computational resources.

Hyperparameter Sensitivity: While we report results across five random seeds, we did not conduct extensive hyperparameter tuning due to computational constraints. The adversarial loss weight (λ = 0.1), focal loss parameters (α = 0.25, γ = 2.0), and learning rates were based on literature recommendations and limited preliminary experiments. More extensive hyperparameter search might reveal better configurations.

Grid search across hyperparameters and multiple seeds would be computationally prohibitive. Bayesian optimization or other efficient search methods could improve hyperparameter selection in future work.

Training Stability: Adversarial training can be unstable, requiring careful balancing of competing objectives. While M3 shows acceptable variance across seeds (accuracy std: 0.024), some individual runs may fail to converge or converge to poor local optima. Production deployment should train multiple models and select best performers.

Inference Time: While we report inference time (8 seconds per 1,500 samples), real-time applications requiring immediate response might need further optimization through model distillation, quantization, or specialized hardware.

6.5.4 Threats to Validity

Internal Validity: The primary threat is that observed improvements might stem from factors other than the intended mechanisms:

Focal loss improves minority class performance but might also incidentally affect fairness

Counterfactual augmentation increases dataset size, possibly improving performance through volume rather than diversity

Adversarial training adds regularization that might improve generalization independent of debiasing

Ablation studies systematically removing components would strengthen causal attribution.

External Validity: Generalization beyond the specific context is uncertain:

Reddit mental health communities may differ from other platforms

English text patterns may not extend to other languages

Social media informal language differs from clinical documentation

2019-2020 data may not represent current discourse

Cross-platform, multilingual, and temporal validation would strengthen external validity.

Construct Validity: The operationalization of fairness through demographic parity, equalized odds, and bias amplification captures important aspects but not all dimensions of fairness:

Individual fairness (similar individuals treated similarly) not evaluated

Procedural fairness (fair decision process) not addressed

Long-term fairness (impacts over time) not measured

Alternative fairness definitions might reveal different patterns.

Statistical Conclusion Validity: Several factors could lead to incorrect statistical conclusions:

Multiple comparisons (many metrics across many subgroups) increase false positive risk

Small subgroup sizes (e.g., 56 female samples) limit statistical power

Assumption of independence across seeds may not hold if random number generation is related

Bonferroni correction or false discovery rate control would reduce Type I error risk but decrease power.

6.6 Future Research Directions

6.6.1 Methodological Extensions

Intersectional Fairness Analysis: Current analysis examines protected attributes independently, missing intersectional effects. For example, female users in high-crisis communities might face compounded disadvantages not visible in single-attribute analysis. Future work should:

Evaluate performance on intersectional subgroups (e.g., female × high-crisis)

Develop fairness metrics sensitive to intersectional disparities

Design interventions addressing multiple axes of bias simultaneously

Investigate whether existing biases compound or counteract at intersections

This extension requires larger datasets with sufficient samples in intersectional subgroups to enable reliable statistical analysis.

Causal Fairness: Current fairness metrics measure correlational patterns between protected attributes and outcomes. Causal fairness frameworks examine whether differences in outcomes arise from legitimate causal pathways versus discriminatory mechanisms. Future work should:

Develop causal models representing relationships between attributes, features, and outcomes

Use counterfactual reasoning to identify discriminatory causal pathways

Distinguish between legitimate correlations (e.g., subreddit membership genuinely correlates with crisis prevalence) and illegitimate ones (e.g., gender influencing predictions through stereotypes)

Design interventions targeting specific causal pathways

Kusner et al. (2017) provide foundational work on counterfactual fairness that could be adapted to mental health NLP.

Longitudinal Fairness: Current evaluation uses cross-sectional data, but fairness properties may change over time as:

Language evolves and models become outdated

User populations shift

New biases emerge from deployment feedback loops

Future work should examine:

How fairness metrics degrade over time without retraining

Whether bias mitigation techniques maintain effectiveness as language evolves

Feedback loops where model predictions influence user behavior, potentially creating new biases

Optimal retraining schedules balancing fairness, performance, and computational cost

Explainable Fairness: Current models provide limited insight into why particular predictions are made and whether reasoning is fair. Future work should develop:

Explanation methods showing which input features drive predictions

Techniques for identifying when explanations differ systematically across demographic groups

Approaches ensuring explanations themselves are unbiased

Methods for detecting spurious correlations in model reasoning

Attention visualization, saliency maps, and counterfactual explanations could be adapted to mental health NLP with appropriate privacy protections.

6.6.2 Technical Improvements

Architecture Innovations: Several architectural modifications might improve performance and fairness:

Larger pretrained models (RoBERTa, ALBERT, DeBERTa) might provide stronger baselines

Domain-adaptive pretraining on mental health text before fine-tuning might improve mental health-specific understanding

Multi-task learning jointly optimizing for multiple mental health tasks might improve generalization

Attention mechanisms specifically designed to ignore protected attributes might enhance debiasing

Advanced Debiasing Techniques: Recent fairness literature proposes novel debiasing approaches worth exploring:

Adversarial reweighting automatically learning example weights to promote fairness

Distributionally robust optimization ensuring worst-case subgroup performance exceeds thresholds

Fair representation learning creating demographic-invariant embeddings

Causal intervention training using do-calculus to remove discriminatory pathways

Uncertainty Quantification: Current models provide point predictions without uncertainty estimates. Future work should:

Develop calibrated confidence estimates that reflect true prediction reliability

Examine whether uncertainty varies systematically across demographic groups

Use uncertainty for selective prediction, abstaining on uncertain cases

Integrate uncertainty into clinical decision support interfaces

Bayesian deep learning or ensemble methods could provide uncertainty quantification with appropriate calibration.

Efficient Models: For deployment in resource-constrained settings, future work should investigate:

Knowledge distillation creating smaller student models from large teacher models

Quantization reducing model precision without substantial performance loss

Pruning removing unnecessary model parameters

Neural architecture search finding efficient architectures optimized for specific hardware

These efficiency techniques should be evaluated for their impact on fairness—compressing models might disproportionately harm performance on minority subgroups.

6.6.3 Application Extensions

Multilingual Mental Health NLP: Extending to non-English languages would increase global impact:

Adapt methodology to Nepali, Hindi, and other South Asian languages for regional mental health support

Investigate whether fairness interventions generalize across languages with different grammatical structures

Examine culture-specific protected attributes and bias sources

Develop multilingual models handling code-switching common in multilingual contexts

Multilingual extension faces challenges including limited mental health data in low-resource languages and difficulty inferring demographics from text in different linguistic contexts.

Clinical Text Analysis: Adapting models to clinical documentation (therapy notes, electronic health records) would enable direct clinical application:

Clinical text uses formal medical terminology versus social media informality

Privacy regulations (HIPAA) impose strict requirements on clinical data use

Clinical contexts have different fairness requirements and stakeholder needs

Integration with existing clinical workflows requires specialized interfaces

Collaboration with healthcare providers would be essential for clinical deployment.

Real-time Crisis Detection: Current batch evaluation doesn't reflect real-time operational requirements:

Develop streaming models processing new posts immediately

Optimize inference latency for acceptable real-time performance

Design systems handling bursty load when many users post simultaneously

Implement graceful degradation when resources are constrained

Real-time deployment raises additional challenges including concept drift detection and online model updating.

Intervention Recommendation: Beyond detection, systems should recommend appropriate interventions:

Match crisis types to intervention strategies (peer support vs. professional help vs. emergency services)

Personalize recommendations based on user preferences and history

Consider cultural factors affecting intervention acceptability

Evaluate intervention effectiveness and fairness of intervention access

This extension requires collaboration with mental health professionals to ensure clinical appropriateness.

6.6.4 Ethical and Social Research

Stakeholder Engagement: Technical fairness improvements should be complemented by stakeholder input:

Conduct participatory design workshops with mental health community members

Survey users about fairness priorities and acceptable trade-offs

Interview mental health professionals about clinical decision-making needs

Engage ethicists in defining fairness objectives appropriate to context

Stakeholder perspectives may reveal fairness concerns not captured by technical metrics.

Deployment Studies: Field studies in real deployment contexts would provide insights beyond laboratory evaluation:

Partner with mental health organizations to deploy and evaluate M3 in practice

Measure impact on user outcomes (help-seeking behavior, symptom improvement)

Assess mental health professional trust and adoption patterns

Document unexpected consequences and deployment challenges

Field studies enable iterative improvement based on real-world performance rather than proxy metrics.

Policy Research: Technical capabilities should inform policy development:

Develop guidelines for ethical mental health AI deployment

Propose regulatory frameworks ensuring fairness and safety

Create certification processes for mental health AI systems

Design transparency requirements appropriate to context

Policy research requires collaboration with legal scholars, ethicists, and policymakers.

Long-term Impact Assessment: Current evaluation focuses on immediate outcomes, but long-term effects matter:

Do fairness interventions reduce or exacerbate health disparities over time?

How do users adapt to algorithmic monitoring, and does this create new biases?

What are macroeconomic impacts on mental health workforce and access?

How does technology affect mental health stigma and help-seeking behavior?

Long-term assessment requires interdisciplinary collaboration and longitudinal study designs.

6.7 Conclusion

This thesis investigated bias mitigation in mental health sentiment analysis, developing and evaluating FairBERT (M3), a model combining adversarial debiasing, counterfactual data augmentation, and focal loss to address both algorithmic bias and class imbalance. Through comprehensive empirical evaluation on the RMH-Bias-10K dataset, we demonstrated that fairness-aware machine learning can achieve superior performance and fairness simultaneously, challenging the assumed performance-fairness trade-off.

6.7.1 Research Contributions

This work makes several contributions to fairness in NLP and mental health AI:

Methodological Contributions:

Integrated Fairness Architecture: M3 represents the first mental health NLP model integrating multiple complementary bias mitigation techniques (adversarial debiasing, counterfactual augmentation, focal loss) in a unified framework

Comprehensive Fairness Evaluation: We establish fairness evaluation methodology for mental health NLP, examining four protected attributes (gender, crisis severity, dialect, subreddit) across four fairness metrics (demographic parity, equalized odds, bias amplification, accuracy parity)

Dialect-Focused Debiasing: The counterfactual augmentation approach targeting dialect variation achieves dramatic bias reduction (79.8% DPD reduction) and establishes dialect fairness as critical dimension for mental health NLP

Performance-Fairness Complementarity: We provide empirical evidence that fairness constraints can improve rather than compromise performance, demonstrating 10.3 percentage point accuracy improvement over baseline with 56.9% bias amplification reduction

Empirical Contributions:

Dataset Creation: The RMH-Bias-10K dataset with expert-validated annotations (96.6% agreement) and comprehensive fairness labels provides a valuable resource for future fairness research in mental health NLP

Benchmark Results: M3's 84.0% accuracy, 0.688 F1-macro, and 0.843 AUC-ROC with substantially improved fairness metrics establish performance benchmarks for fair mental health sentiment analysis

Subgroup Analysis: Detailed examination of performance patterns across demographic subgroups reveals both successes (dialect fairness achievement) and remaining challenges (gender disparities), informing future research priorities

Practical Contributions:

Deployment Viability: The demonstration that fairness and performance can be achieved simultaneously removes a key barrier to deploying fair mental health AI in safety-critical contexts

Resource Efficiency: The moderate dataset size (10,000 samples) and reasonable computational requirements (GPU training, CPU inference) make the approach accessible to resource-constrained organizations

Methodological Transferability: The fairness evaluation framework and debiasing techniques can be adapted to other mental health NLP tasks, languages, and cultural contexts

6.7.2 Implications for Mental Health AI

The findings carry important implications for developing equitable mental health AI systems:

Technical Feasibility: Fair mental health AI is technically achievable using existing machine learning techniques. Organizations need not choose between accuracy and fairness—both objectives can be pursued simultaneously through appropriate architecture design and training procedures.

Fairness as Quality Indicator: The finding that fairness interventions improve performance suggests that fairness metrics serve as indicators of model quality and generalization. Models that perform poorly on fairness metrics likely have learned spurious correlations that limit their practical utility. Fairness evaluation should be standard practice in model development, not an optional ethical consideration.

Subgroup Heterogeneity: The substantial variation in performance across demographic subgroups (accuracy range: 66.0% to 93.5% across subreddits) demonstrates that aggregate metrics can mask important disparities. Comprehensive evaluation must examine subgroup-specific patterns to ensure equitable service delivery.

Domain-Specific Fairness: The appropriate definition of fairness depends on domain context. In mental health applications, maintaining high sensitivity (recall) for high-crisis populations even at the cost of lower precision represents appropriate prioritization. Fairness definitions should reflect domain values and stakeholder priorities rather than applying mathematical definitions indiscriminately.

Complementary Human-AI Systems: M3's 38.6% precision on positive class indicates that human oversight remains essential. Mental health AI should augment rather than replace human clinical judgment, with algorithms serving as screening tools that prioritize human attention rather than autonomous decision-makers.

6.7.3 Broader Impact on Fairness in NLP

Beyond mental health applications, this work contributes to broader understanding of fairness in natural language processing:

Performance-Fairness Relationships: The absence of a trade-off in our mental health classification task suggests that trade-off assumptions should be domain-specific rather than universal. Domains where protected attributes do not inherently correlate with outcomes may exhibit complementarity between fairness and performance. Future fairness research should characterize when trade-offs exist versus when fairness interventions improve performance.

Multiple Debiasing Mechanisms: The success of combining adversarial training (representation-level), counterfactual augmentation (data-level), and focal loss (training objective-level) suggests that multi-pronged approaches targeting different aspects of bias are more effective than single-technique interventions. This principle may generalize beyond mental health to other NLP applications.

Evaluation Comprehensiveness: The examination of four fairness metrics across four protected attributes reveals that different metrics capture different fairness dimensions. No single metric provides complete fairness assessment. Future fairness research should embrace multidimensional evaluation rather than optimizing single metrics.

Protected Attribute Selection: The dramatic variation in baseline bias across attributes (dialect bias: 0.308 DPD vs. subreddit bias: 0.247 DPD in M1) demonstrates that bias audits must examine multiple demographic dimensions. Protected attributes that seem less obvious (dialect, community membership) may exhibit substantial bias requiring mitigation.

6.7.4 Future Outlook

Mental health AI stands at a critical juncture. Growing mental health needs, particularly in resource-constrained settings, create demand for scalable technology-augmented support. Simultaneously, growing awareness of algorithmic bias raises concerns about exacerbating existing mental health disparities through unfair AI deployment.

This thesis demonstrates that these concerns need not be in tension. Fair mental health AI is achievable, and fairness interventions can enhance rather than compromise system quality. The path forward requires:

Technical Development: Continued research improving fairness-aware architectures, evaluation methodologies, and debiasing techniques. The methodological foundation established here provides starting points, but substantial work remains in areas like intersectional fairness, causal debiasing, and uncertainty quantification.

Empirical Validation: Large-scale field studies examining real-world deployment outcomes beyond laboratory performance metrics. Does fair AI actually reduce mental health disparities in practice? What unexpected consequences emerge in deployment? Answering these questions requires collaboration between ML researchers and healthcare organizations.

Ethical Framework Development: Technical capabilities must be guided by ethical principles. What fairness objectives should mental health AI pursue? How should competing fairness definitions be prioritized? Who decides? These questions require input from diverse stakeholders including affected communities, mental health professionals, ethicists, and policymakers.

Policy and Governance: Technical standards and best practices must be translated into policy frameworks ensuring safe and equitable deployment. This includes developing certification processes, transparency requirements, and accountability mechanisms appropriate to mental health contexts.

Global Accessibility: Mental health AI should be accessible globally, particularly in resource-constrained settings with greatest needs. This requires multilingual models, efficient architectures, and open-source tools enabling local adaptation. The methodology developed here provides foundation for global extension.

6.7.5 Closing Remarks

Mental health AI holds tremendous promise for expanding access to support and reducing suffering globally. However, realizing this promise requires ensuring that technology serves all populations equitably. Algorithmic bias threatens to exclude vulnerable populations from AI benefits or subject them to discriminatory treatment, potentially exacerbating existing mental health disparities.

This thesis demonstrates that fair mental health AI is both necessary and achievable. Through careful architecture design, comprehensive evaluation, and evidence-based debiasing techniques, we can build systems that provide accurate crisis detection while treating all users equitably regardless of gender, linguistic style, or community membership.

The road from laboratory research to beneficial deployment is long, requiring continued technical innovation, empirical validation, ethical deliberation, and policy development. However, the empirical evidence presented here—that fairness and performance can be complementary objectives—provides reason for optimism. We need not sacrifice accuracy for fairness or fairness for accuracy. Both objectives can be pursued simultaneously through appropriate methodology.

As mental health AI systems move from research to deployment, the principles and techniques developed in this thesis can inform more equitable design. By prioritizing fairness from the beginning of system development rather than treating it as an afterthought, we can build mental health AI that reduces rather than amplifies existing inequities.

The ultimate measure of success is not classification accuracy or fairness metrics on test sets, but whether deployed systems actually improve mental health outcomes equitably across all populations. Achieving this vision requires continued collaboration among machine learning researchers, mental health professionals, affected communities, and policymakers. This thesis provides methodological and empirical foundations for that collaborative effort, demonstrating that fair mental health AI is not merely an aspiration but an achievable goal

**APPENDIX A**

**APPENDIX A: COLAB NOTEBOOKS FOR DATASET ACCESS AND REPRODUCIBILITY**

The following Google Colab notebooks contain all code necessary to load, inspect, preprocess, and prepare the RMH-Bias-10K dataset for the experiments described in this thesis. Each notebook is self-contained, includes markdown explanations, and can be run end-to-end in a free Google Colab instance (no additional installation required except for standard !pip commands when needed).

All notebooks use fixed seeds for reproducibility (seed=42 by default). Readers can change the seed or paths to experiment further. The dataset CSVs are assumed to be placed in the Colab /content/ folder or mounted from Google Drive.

**A.1 RMH-Bias-10K – Dataset Loading & Basic Statistics**

Link:<https://colab.research.google.com/drive/1cXBiaMF0zaPd5vHsZ7IGZZAO6TWCgUqz?usp=share_link>.

This notebook:

Downloads the dataset CSV(s) from Google Drive

<https://drive.google.com/drive/folders/1UNXJWNEjpJ2LcU8VlIYalsPY9ZuyEOEV?usp=share_link>.

Shows summary statistics (class distribution, subreddit distribution, text length histograms)

Visualizes label distribution and example posts

Computes inter-annotator agreement metrics (if applicable)

**A.2 RMH-Bias-10K – Preprocessing & Train/Val/Test Split**

Link: https://colab.research.google.com/drive/YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

This notebook:

Applies the exact cleaning steps described in Chapter 3

Performs the stratified train/val/test split (70/10/20)

Saves the splits as CSV files (optional – useful for exact reproduction)

Reproduces the proxy group creation (length > median = group 1)

**A.3 RMH-Bias-10K – Rebalancing Experiments (Undersampling + CDA)**

Link: https://colab.research.google.com/drive/ZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZ

This notebook demonstrates:

Strategic length-stratified undersampling

Rule-based counterfactual data augmentation (CDA) on positive class

Final class distribution after each step

Side-by-side comparison of original vs. balanced training sets

**A.4 Model Training – BiLSTM (M1)**

Link: https://colab.research.google.com/drive/AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

This notebook implements:

Multi-seed training pipeline for BiLSTM classifier

Vocabulary building and text encoding

Early stopping with patience monitoring

Per-seed performance evaluation and aggregated results

**A.5 Model Training – BERT and FairBERT (M2 & M3)**

Link: https://colab.research.google.com/drive/BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

This notebook implements:

BERT-base fine-tuning for sentiment classification (M2)

FairBERT with gradient reversal layer for bias mitigation (M3)

Multi-seed evaluation across both models

Comparative fairness metrics computation

Note: The notebooks are designed so that changing only the DATA\_PATH variable allows users to run them on their own copy of the dataset.

**APPENDIX B: PREPROCESSING PIPELINE EXAMPLES**

This appendix provides concrete examples of the preprocessing pipeline applied to raw Reddit posts prior to modeling.

**B.1 Raw-to-Clean Transformations**

A typical raw post may contain URLs, emojis, user mentions, and other noisy artifacts.

Example transformation:

Raw post:

"I'm so tired of everything rn... idk what to do anymore check this:

https://imgur.com/xyz @user123 😢"

Processed clean text:

"i am so tired of everything right now idk what to do anymore"

The transformation steps include:

Removal of URLs, user mentions, and HTML artifacts

Removal of emojis and special characters

Normalization of contractions (e.g., "rn" → "right now")

Lowercasing of all tokens

Normalization of repeated punctuation and whitespace

**B.2 Tokenization Examples**

The tokenization behavior differs between the transformer-based model (M2) and the BiLSTM model (M1).

M2 (BERT) WordPiece tokens:

['i', 'am', 'so', 'tired', 'of', 'every', '##thing']

M1 (BiLSTM + GloVe) tokens:

['i', 'am', 'so', 'tired', 'of', 'everything']

The WordPiece vocabulary can split rare or morphologically complex tokens into sub-word units (e.g., "every" + "##thing"), whereas the GloVe-based pipeline operates on word-level tokens.

**B.3 Handling Non-Standard Forms**

The pipeline also handles elongated words and informal orthography.

Example:

Raw:

"i am soooo anxious before exams!!!"

Normalized:

"i am soo anxious before exams"

Excessive character repetition is reduced to at most two consecutive occurrences, preserving some expressive emphasis while avoiding extreme sparsity in the token space.

**APPENDIX C: SUPPLEMENTARY DESCRIPTIVE STATISTICS**

This appendix provides supplementary descriptive statistics and intermediate analysis results that support, but are not essential to, the main narrative.

**C.1 Length and Dialect Distributions**

Table C.1 summarizes example statistics for post length and dialect score buckets.

Table C.1: Distribution of posts by length bucket with mean dialect scores

|  |  |  |
| --- | --- | --- |
| Bucket | Proportion of Posts | Mean Dialect Score |
| Short (<30 tokens) | 0.41 | 0.32 |
| Medium (30–80 tokens) | 0.44 | 0.47 |
| Long (>80 tokens) | 0.15 | 0.51 |

**C.2 Extended Severity-Level Results**

Table C.2 illustrates an extended view of severity-stratified performance for the BERT baseline (M2), complementing the summary in Chapter 5. These figures reinforce the trend that performance degrades, and bias amplification increases, as linguistic severity rises.

Table C.2: Extended severity-level performance for M2 (example values)

|  |  |  |  |
| --- | --- | --- | --- |
| Severity | Accuracy | Macro-F1 | BA |
| Mild | 0.88 | 0.84 | 0.09 |
| Moderate | 0.86 | 0.83 | 0.12 |
| Severe | 0.80 | 0.77 | 0.14 |

**APPENDIX D: CODE IMPLEMENTATION EXAMPLES**

This appendix documents representative code snippets used for data collection, preprocessing, and model preparation. The excerpts are abbreviated for clarity. Full implementations are available in the Colab notebooks referenced in Appendix A.

**D.1 Data Cleaning Function**

import re

def clean\_text(text):

"""

Clean raw Reddit post text by removing URLs, normalizing whitespace,

and converting to lowercase.

Args:

text (str): Raw text from Reddit post

Returns:

str: Cleaned text

"""

# Remove URLs

text = re.sub(r"http\S+", "", text)

# Remove user mentions

text = re.sub(r"@\w+", "", text)

# Normalize whitespace

text = re.sub(r"\s+", " ", text)

# Lowercase and strip

text = text.lower().strip()

return text

**D.2 Counterfactual Gender Swap for Data Augmentation**

GENDER\_SWAP = {

"he": "she",

"she": "he",

"him": "her",

"her": "him",

"his": "her",

"hers": "his",

"boyfriend": "girlfriend",

"girlfriend": "boyfriend",

"husband": "wife",

"wife": "husband"

}

def apply\_gender\_swap(text):

"""

Apply counterfactual gender swapping to text for data augmentation.

Args:

text (str): Input text

Returns:

str: Text with gender terms swapped

"""

def repl(match):

token = match.group(0)

lower = token.lower()

if lower in GENDER\_SWAP:

swapped = GENDER\_SWAP[lower]

# Preserve capitalization

if token[0].isupper():

swapped = swapped.capitalize()

return swapped

return token

pattern = r"\b(" + "|".join(GENDER\_SWAP.keys()) + r")\b"

return re.sub(pattern, repl, text, flags=re.IGNORECASE)

**D.3 Training Loop Skeleton (Illustrative)**

def train\_model(model, train\_loader, val\_loader, num\_epochs, optimizer, loss\_fn):

"""

Basic training loop with validation.

Args:

model: PyTorch model

train\_loader: DataLoader for training data

val\_loader: DataLoader for validation data

num\_epochs: Number of training epochs

optimizer: Optimizer instance

loss\_fn: Loss function

Returns:

Trained model

"""

for epoch in range(num\_epochs):

# Training phase

model.train()

train\_loss = 0

for batch in train\_loader:

optimizer.zero\_grad()

inputs, labels = batch

outputs = model(inputs)

loss = loss\_fn(outputs, labels)

loss.backward()

optimizer.step()

train\_loss += loss.item()

# Validation phase

model.eval()

val\_loss = 0

with torch.no\_grad():

for batch in val\_loader:

inputs, labels = batch

outputs = model(inputs)

loss = loss\_fn(outputs, labels)

val\_loss += loss.item()

print(f"Epoch {epoch+1}/{num\_epochs} - "

f"Train Loss: {train\_loss/len(train\_loader):.4f}, "

f"Val Loss: {val\_loss/len(val\_loader):.4f}")

return model

These snippets reflect the main implementation style used in the experimental pipeline, while full scripts are retained in the separate code repository and Colab notebooks.

**APPENDIX E: FAIRNESS METRICS FORMALIZATION**

This appendix formalizes the fairness metrics used in the thesis and provides worked examples to clarify interpretation.

**E.1 Notation**

Let:

Y ∈ {0,1} denote the ground-truth label (1 = negative class)

Ŷ ∈ {0,1} denote the model prediction (1 = negative class)

A ∈ {female, male} denote the (inferred) gender group

**E.2 Demographic Parity Difference**

Demographic Parity Difference (DPD) for the negative class is defined as:

DPD = P(Ŷ = 1 | A = female) − P(Ŷ = 1 | A = male)

A positive value indicates that the model predicts negative sentiment more often for the female group, conditional on no other information. Perfect demographic parity requires DPD = 0.

**E.3 Equalized Odds**

Let:

TPR<sub>g</sub> = P(Ŷ = 1 | Y = 1, A = g) (True Positive Rate for group g)

FPR<sub>g</sub> = P(Ŷ = 1 | Y = 0, A = g) (False Positive Rate for group g)

for group g ∈ {female, male}.

The Equalized Odds gaps are:

EO-TPR = TPR<sub>female</sub> − TPR<sub>male</sub>

EO-FPR = FPR<sub>female</sub> − FPR<sub>male</sub>

Values close to zero indicate more similar error behavior between groups. Perfect equalized odds requires both gaps equal zero.

**E.4 Bias Amplification**

Bias Amplification (BA) measures how much the predictive distribution diverges from the true label distribution across groups:

BA = [P(Ŷ = 1 | A = female) − P(Y = 1 | A = female)] − [P(Ŷ = 1 | A = male) − P(Y = 1 | A = male)]

A positive BA indicates that existing disparities in negative labels are being amplified by the model. BA = 0 indicates the model preserves the baseline disparity without amplification.

**E.5 Illustrative Example**

Suppose that on the test set:

P(Y = 1 | A = female) = 0.70

P(Y = 1 | A = male) = 0.68

P(Ŷ = 1 | A = female) = 0.78

P(Ŷ = 1 | A = male) = 0.70

Then:

DPD = 0.78 − 0.70 = 0.08

BA = (0.78 − 0.70) − (0.70 − 0.68) = 0.08 − 0.02 = 0.06

These values suggest a modest amplification of gender disparity in negative predictions relative to the underlying label distribution. The model predicts negative sentiment 8 percentage points more often for females, and this represents a 6 percentage point amplification beyond the baseline difference in true labels.

**APPENDIX F: ETHICAL CONSIDERATIONS AND DATA HANDLING**

This appendix supplements the ethical discussion in the main chapters by summarizing key steps taken to mitigate risks associated with working on mental-health-related user-generated content.

**F.1 Anonymization and Paraphrasing**

All example posts shown in this thesis are anonymized and paraphrased. Direct usernames, links, and identifiable details are removed or obfuscated. The aim is to preserve linguistic characteristics relevant to modeling while protecting the privacy of the original authors.

Where example posts are presented, they represent composite or synthetic examples that reflect typical patterns in the data rather than verbatim quotes from individual users.

**F.2 Compliance with Platform Policies**

Data was collected only from publicly accessible Reddit communities and used strictly for non-commercial academic research. Care was taken to respect Reddit's terms of service and community guidelines, including:

* Not attempting to deanonymize or contact individual users
* Restricting data collection to public posts
* Using the data exclusively for research purposes
* Not redistributing raw data without appropriate permissions

**F.3 Storage and Access Control**

Dataset files were stored on password-protected devices and in private cloud storage associated with the researcher's institutional account. Specific measures included:

* Encryption of local storage containing raw data
* Access-controlled cloud storage with multi-factor authentication
* Temporary artifacts (such as intermediate caches) in cloud environments (e.g., Google Colab) were cleared after experiments
* No raw dataset files are publicly distributed without appropriate licensing and ethical review

**F.4 Informed Consent and Public Data**

While Reddit posts are publicly available, users may not anticipate their content being used in research. This study follows the principle that:

* Public online content can be used for research with appropriate anonymization
* Individual posts are never attributed to specific users
* Sensitive content is handled with particular care
* The aggregated nature of the analysis minimizes individual privacy risks

**F.5 Scope and Limitations**

The analyses in this thesis are not intended to be used for:

* Clinical decision-making or diagnosis
* Automated moderation of real users without human oversight
* Direct intervention in mental health communities
* Profiling or classification of individual users

Instead, they provide an empirical case study of fairness issues in sentiment modeling on mental-health text. Any future deployment of similar systems in real-world settings would require:

* Additional ethical review by an institutional review board (IRB)
* Domain expert involvement (mental health professionals)
* More stringent validation with diverse populations
* Transparent communication with affected communities
* Ongoing monitoring for unintended consequences

**F.6 Potential Benefits and Risks**

Potential Benefits:

* Improved understanding of bias in mental health NLP systems
* Development of fairer sentiment analysis tools
* Contribution to responsible AI in healthcare contexts

Potential Risks:

* Misuse of techniques for surveillance or profiling
* Stigmatization through inaccurate predictions
* Privacy breaches if anonymization is insufficient
* Reinforcement of harmful stereotypes if findings are misinterpreted

This research prioritizes transparency about both the benefits and limitations of the work to enable informed evaluation by the research community and stakeholders.