الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

جامعة سعيدة د. مولاي الطاهر
كلية الرياضيات و الإعلام الآلي و الاتصالات السلكية و اللاسلكية
قسم: الإعلام الآلي

Mémoire de Master en informatique

Spécialité : Modélisation Informatique des Connaissances et du Raisonnement

T h è m e

Multi-Task Learning for Mental Health Disorder Classification from User-Written Text

Présenté par :
TAHI Boumediene Amir

Dirigé par :
Dr. KADARI Rekia
Co-Dirigé par :
Dr. YAHLALI Mebarka

Année universitaire 2025–2026

Acknowledgements

First and foremost, I would like to express my deepest and most heartfelt gratitude to my mother, for everything she has done for me. Her unconditional love, sacrifices, and unwavering support have carried me through every stage of this journey, and none of this would have been possible without her prayers.

I would also like to extend my sincere gratitude to my supervisor, **Dr. KADARI Rekia**, for her guidance, patience, and continuous involvement throughout the realisation of this thesis. Her feedback and dedication were instrumental in shaping both the direction and the quality of this work.

My thanks also go to the Department of Computer Science at the University of Saïda – Dr. Moulay Tahar, and to all the professors who have contributed to my academic formation throughout these couple years. I am equally grateful to my colleagues, with whom I shared the challenges and milestones of this academic path.

I wish to thank my co-workers and my employers for granting me this opportunity to pursue my studies alongside my professional responsibilities, and for their understanding and flexibility throughout this period.

Special thanks go to **Salim**, **Omar**, and **Rami**, for their help in giving me the chance to seize this opportunity.

To my beloved future wife, whom I haven't met yet – I look forward to the day our paths cross. :3

To **Mezaoui T.**, for pushing me to work harder until I graduated at the top of my class.

Finally, to everyone who, in one way or another, stood by me during this journey – thank you.

TAHI Boumediene Amir

Contents

List of Tables

List of Figures

Abstract

Mental health disorders such as depression, anxiety, bipolar disorder, stress, personality disorders, and suicidal ideation affect millions of people worldwide and remain difficult to identify at an early stage. Recent advances in Natural Language Processing (NLP) have enabled the automatic analysis of user-written text for mental health assessment. This thesis investigates the use of Multi-Task Learning (MTL) for mental health disorder classification from user-written text, where disorder prediction is learned jointly with sentiment analysis as an auxiliary task. Several deep learning architectures were evaluated, including attention-based models, transformer-based models, and hybrid approaches, using two publicly available mental health datasets. Experimental results show that transformer architectures significantly outperform traditional approaches, with DeBERTa-base achieving the best performance, reaching 94.06% accuracy and 93.09% macro F1-score. The study also examines the impact of multi-task learning, model design choices, and training strategies on classification performance. The findings confirm the effectiveness of transformer-based models for mental health text classification and contribute to the development of accurate and scalable computational mental health screening systems.

Keywords: Mental Health Classification, Multi-Task Learning, Natural Language Processing, Transformer Models, DeBERTa, DistilBERT, Sentiment Analysis, Computational Mental Health, Deep Learning.

General Introduction

Context and Motivation

Mental health disorders such as depression, anxiety, bipolar disorder, and suicidal ideation represent a major global public health challenge, affecting hundreds of millions of individuals worldwide. Traditional diagnostic pathways rely heavily on clinical interviews and self-reported questionnaires, which are inherently limited by delayed help-seeking behaviour, social stigma, and restricted access to mental health professionals—particularly in low-resource settings. In parallel, the unprecedented growth of social media platforms and online communities has produced vast amounts of user-generated text that implicitly encodes emotional states, cognitive patterns, and behavioural cues. This convergence has opened a promising avenue for computational approaches capable of detecting early warning signs of mental health disorders directly from naturally occurring written language, offering the potential for scalable, non-intrusive, and timely screening tools.

Problem Statement

Despite substantial progress in natural language processing (NLP), automatic mental health disorder classification from text remains a challenging problem for several reasons. First, the linguistic manifestations of different disorders frequently overlap—symptoms of depression and anxiety, for instance, share many surface-level features—making fine-grained multi-class classification considerably harder than binary detection. Second, publicly available datasets are often class-imbalanced, with critical minority classes (e.g., suicidal ideation, personality disorders) underrepresented relative to majority classes. Third, single-task classification models, trained solely to predict a disorder label, may fail to capture the broader emotional and contextual signals that a human clinician would naturally consider when forming a diagnostic impression. This raises the central research question addressed in this thesis: *can a multi-task learning framework, jointly optimising mental health disorder classification alongside a complementary auxiliary task such as sentiment analysis, yield richer textual representations and improved classification performance compared to single-task baselines?*

Objectives

The present work pursues the following objectives:

- Review and critically analyse the state of the art in text-based mental health disorder classification, with particular emphasis on multi-task learning (MTL) approaches.
- Design and implement a progression of model architectures—ranging from custom-built attention mechanisms to fine-tuned transformer-based language models (DistilBERT, DeBERTa)—for mental health text classification.
- Formalise and implement a multi-task learning framework that jointly optimises disorder classification and sentiment analysis as an auxiliary task, investigating whether the auxiliary signal improves the quality of learned representations.
- Rigorously evaluate the proposed models on two complementary datasets (a seven-class clinical conditions dataset and a six-class Reddit dataset) using a comprehensive set of metrics, including accuracy, macro/weighted F1-score, Matthews Correlation Coefficient, and Cohen's Kappa.
- Analyse the trade-offs between model capacity, computational efficiency, and classification performance, and to identify the configuration offering the best overall balance.

Contributions

The main contributions of this thesis can be summarised as follows:

- A systematic experimental comparison of scratch-built attention-based models and pre-trained transformer backbones for mental health text classification.
- A multi-task learning architecture jointly predicting mental health disorder categories and sentiment polarity, with an empirical study of its impact on classification performance relative to single-task counterparts.
- An extensive evaluation across two distinct datasets and multiple architectural and training configurations (loss functions, layer-freezing strategies, batch sizes), culminating in a best-performing DeBERTa-base model achieving 94.06% accuracy and 93.09% macro F1-score on the held-out test set.
- A detailed error and class-wise analysis highlighting the specific challenges posed by minority classes and semantically overlapping disorders.

Thesis Outline

The remainder of this thesis is organised as follows. **Chapter 1** reviews the state of the art in text-based mental health classification, surveying existing approaches and situating multi-task learning within this landscape. **Chapter 2** presents the theoretical and technical background necessary to understand the proposed methods, including attention mechanisms and transformer architectures. **Chapter 3** details the methodology and experimental setup, covering the data pipeline, the multi-task learning framework, and the architectures investigated. **Chapter 4** reports and discusses the experimental results, comparing the performance of all proposed models and analysing the impact of various design choices. Finally, the **General Conclusion** synthesises the main findings, discusses the limitations of the present work, and outlines directions for future research.

Chapter 1

State of the Art: Multi-Task Learning for Mental Health Disorder Classification from User-Written Text

Abstract

The automatic detection of mental health disorders from user-generated text — such as social media posts, clinical notes, and online therapy transcripts — has emerged as a critical application of natural language processing (NLP). Recent advances have shown that multi-task learning (MTL), where multiple related tasks are learned simultaneously with shared representations, offers significant advantages over single-task learning for mental health classification. This chapter provides a comprehensive review of state-of-the-art MTL approaches for identifying mental health conditions — including depression, anxiety, bipolar disorder, schizophrenia, and suicide risk — from textual data. We synthesise findings from 2020–2025, covering architectural paradigms, shared representation strategies, auxiliary task design, and evaluation frameworks. Key contributions include the demonstration that MTL consistently outperforms single-task baselines, that hierarchical multi-task architectures leveraging theory-driven risk factors yield substantial performance gains, and that emerging directions such as federated MTL, cross-lingual MTL, and prompt-based learning are expanding the frontier of this field.

1.1 Introduction

Mental health disorders constitute one of the leading causes of disability worldwide, yet they remain significantly underdiagnosed and undertreated [1]. The proliferation of social

media and online platforms has created unprecedented opportunities to observe linguistic markers of mental distress at scale, as users increasingly share personal experiences, emotional states, and mental health concerns in digital spaces [2, 3]. Natural Language Processing (NLP) techniques applied to user-generated textual data have demonstrated considerable promise for automated mental health screening, early intervention, and population-level monitoring.

However, the classification of mental health conditions from text presents several fundamental challenges. First, mental disorders frequently co-occur (comorbidity), with depression and anxiety co-occurring in up to 50% of cases [4]. Second, linguistic markers of different conditions overlap substantially — for example, negative affect, anhedonia, and sleep disturbances appear across depression, anxiety, and post-traumatic stress disorder [5]. Third, labelled datasets for mental health conditions are often small, imbalanced, and noisy, thereby limiting the effectiveness of standard supervised learning approaches [6].

Multi-task learning (MTL) addresses these challenges by jointly learning multiple related tasks through shared representations. By leveraging commonalities across conditions while preserving task-specific features, MTL can enhance generalisation, reduce overfitting on small datasets, and capture the complex interdependencies that characterise mental health presentations [7, 8]. This chapter systematically reviews the state of the art in MTL for mental health disorder classification from text, with a particular focus on architectural innovations developed between 2020 and 2025.

1.2 Foundational Concepts in Multi-Task Learning for Mental Health Text Classification

1.2.1 Why Multi-Task Learning for Mental Health?

The rationale for applying MTL to mental health text classification is grounded in several key observations. Mental disorders exist on a continuum and share underlying psychological dimensions — such as negative affect, emotional dysregulation, and cognitive distortions — that manifest in language [5, 9]. Standard single-task classification approaches treat each disorder as an independent prediction problem, thereby overlooking these shared underlying characteristics. MTL addresses this limitation by exploiting task relatedness through the learning of shared representations. Specifically, it enables the model to capture common linguistic patterns across multiple disorders while preserving task-specific information through dedicated layers or components. [10].

One of the most compelling demonstrations of the benefits of MTL for mental health prediction was provided by Ophir et al. [7]. In their study of suicide risk detection from

Facebook posts, they compared a single-task model (STM) that predicted suicide risk directly from textual data with a multi-task model (MTM) that incorporated a hierarchical structure of theory-driven risk factors — personality traits, psychosocial risks, and psychiatric disorders — before the final suicide prediction. The results showed that the MTM significantly outperformed the STM, achieving higher predictive performance, with area under the curve (AUC) scores ranging from 0.697 to 0.746, compared with 0.621 to 0.629 for the STM, with substantially larger effect sizes (Cohen's d values ranging from 0.729 to 0.936). Importantly, subsequent content analyzes revealed that the model's predictions were not primarily driven by explicit suicide-related language, Instead, they relied on a broader set of linguistic features captured through the multi-task hierarchical framework, highlighting the capacity of MTL to learn deeper and more informative representations of mental health risk..

1.2.2 Hard and Soft Parameter Sharing Approaches

MTL architectures for mental health text classification can generally be categorized into two main paradigms. **Hard parameter sharing**, in which multiple tasks share a common encoder (e.g., a Transformer or LSTM) while maintaining task-specific output layers, is the most widely adopted approach owing to its architectural simplicity and reduced risk of overfitting. For example, Kim et al. [3] employed a hard parameter-sharing architecture based on a deep neural network to classify posts from Reddit mental health communities into six disorder categories — depression, anxiety, bipolar disorder, borderline personality disorder, schizophrenia, and Autism Spectrum Disorder (ASD) — achieving robust multi-class discrimination. In contrast, **soft parameter sharing**assigns a separate set of parameters to each task while encouraging similarity between task-specific models through regularization mechanisms (e.g. trace-norm regularization or cross-stitch networks), is less common in this domain but offers greater modelling flexibility, particularly when the relationships between tasks are weaker or more heterogeneous.

1.3 Architectural Innovations (2020–2025)

1.3.1 Hierarchical and Theory-Driven Multi-Task Architectures

The most influential architectural innovation in this period is the hierarchical multi-task framework, which reflects the conceptual hierarchy of mental health risk. As demonstrated by Ophir et al. [7], structuring auxiliary tasks in a theoretically motivated cascade — from linguistic features to personality traits, through psychosocial risk factors, to psychiatric disorders, and finally to the target outcome (suicide risk) — yields substantial performance improvements. This approach leverages psychological theory to constrain the learning

process, thereby ensuring that intermediate representations remain clinically meaningful and interpretable.

Huang et al. [8] extended the MTL paradigm to the federated setting for MRI-based diagnosis of multiple mental disorders (autism spectrum disorder, attention deficit/hyperactivity disorder (ADHD), and schizophrenia). Although their framework is based on neuroimaging data rather than textual data, the methodological innovations — including a federated contrastive learning feature extractor and a multi-gate mixture-of-experts classifier for joint prediction — are potentially transferable to text-based modalities.

1.3.2 Attention Mechanisms and Multi-Task Fusion

Attention mechanisms have been integrated into MTL architectures to dynamically weight the contributions of different tasks and input features. Li et al. [11] proposed a multi-feature fusion recurrent attention model for suicide risk assessment from social media data, combining bidirectional Long Short-Term Memory (BiLSTM) representations with self-attention mechanisms to extract core information from user-generated posts. When combined with external linguistic features, this approach improved risk-level F1 scores by up to 3.7% on the CLPsych 2019 shared task dataset.

Zogan et al. [12] introduced DepressionNet, which incorporates a hybrid extractive-abstractive summarisation strategy to identify relevant content from sequences of user tweets before classification through a CNN with attention-enhanced GRU layers. While not explicitly framed as MTL, the summarisation component can be interpreted as an auxiliary task that enhances the primary objective of depression detection by improving input representation quality.

1.3.3 Prompt-Based and Fine-Grained Multi-Task Learning

Recent research has explored prompt-based learning as a implicit form of MTL. Zhang and Guo [13] proposed a multilevel depression status detection framework using fine-grained prompt learning (MDSD-FGPL), which designs multiple sets of prompts ranging from coarse-grained to fine-grained levels. Enabling the model to capture depressive indicators across different levels of granularity — from binary depression classification to severity estimation. This prompt strategy effectively induces auxiliary tasks that share representations with the primary classification objective, thereby improving the model's ability to learn structured and hierarchical representations of depressive symptoms.

1.3.4 Cross-Lingual and Language-Agnostic MTL

MTL for mental health has also expanded to low-resource languages. Noraset et al. [14] proposed LAPoMM (Language-Agnostic Population-level Mental Health Monitor-

ing), a framework that uses cross-lingual methods to train language-agnostic models for detecting mental health signals from Thai social media data without requiring labelled datasets. By combining language-agnostic representations with deep learning classifiers, they outperformed other cross-lingual techniques for emotion recognition, sentiment analysis, and suicide risk detection. Crucially, correlation analyses revealed strong positive associations between predicted mental signals and real-world administrative records on depression cases and suicide attempts.

1.4 Auxiliary Tasks in Mental Health MTL

1.4.1 Emotion and Sentiment Analysis

Emotion and sentiment detection have emerged as among the most commonly used auxiliary tasks in mental health MTL. Uban et al. [5] developed deep learning models to capture linguistic markers of mental disorders at multiple levels, including content, style, and emotional expression, and further interpreted model behaviour through the lens of psychological theories of cognitive styles and emotional communication. Their findings demonstrated that emotional markers learned through auxiliary tasks significantly improve the ability to differentiate between users with diagnosed mental disorders and healthy control groups.

1.4.2 Cognitive Distortion Detection

Cognitive distortions — defined as irrational or biased thought patterns central to cognitive behavioural therapy (CBT) — represent a particularly promising auxiliary task for multi-task learning (MTL) in mental health applications. Shickel et al. [15] developed a machine learning framework for the automatic detection and classification of 15 common cognitive distortions in mental health related text, using both crowdsourced data and real-world online therapy datasets. This approach not only provides clinically meaningful intermediate representations, but also enhances the interpretability of downstream classification decisions.

1.4.3 Multi-Label and Comorbidity Modelling

Mental health conditions frequently co-occur, making multi-label classification a natural and necessary extension of MTL. Dos Santos et al. [4] introduced the SetembroBR corpus — a Twitter dataset comprising approximately 3.9K users who self-reported a diagnosis or treatment for depression and/or anxiety disorders in Portuguese. Experiments conducted on this dataset demonstrate that jointly modelling depression and anxiety as related

but distinct tasks improves predictive performance compared to independent classifiers, particularly in cases involving comorbidity..

1.5 Datasets and Evaluation Frameworks

1.5.1 Social Media Corpora

The field has benefited from the release of several particularly in cases involving comorbidity. The Reddit mental health dataset compiled by Kim et al. [3] covers six disorder categories and has become a standard evaluation benchmark in the field. Gkotsis et al. [6] introduced an earlier but still influential Reddit-based dataset with 11 disorder themes, achieving 91.08% accuracy for mental illness-related post detection and 71.37% weighted accuracy for theme classification using deep neural networks. The CLPsych shared tasks have provided standardised evaluation frameworks for suicide risk assessment [11], while the SetembroBR corpus [4] addresses an important gap by providing resources for Portuguese-language mental health text analysis.

1.5.2 Evaluation Metrics and Baselines

Standard evaluation metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). However, MTL introduces additional evaluation considerations: task-specific performance must be balanced against overall multi-task gains, and the degree of positive transfer — where learning one task improves performance on another — should be explicitly quantified [7, 10]. Most studies compare MTL models against single-task baselines using identical encoder architectures, overall, findings consistently indicate that MTL improves performance across tasks, particularly in low-resource or more challenging datasets.

1.6 Challenges and Open Problems

1.6.1 Task Interference and Negative Transfer

When tasks are insufficiently related, MTL can suffer from negative transfer, where joint training degrades performance on individual tasks. This issue is particularly relevant in mental health applications, where different disorders may have distinct linguistic signatures that can conflict within shared representation spaces [8]. The multi-gate mixture-of-experts approach proposed by Huang et al. [8] mitigates this problem by learning task-specific expert weighting mechanisms, However, further research is still required to better

understand task relatedness in mental health text domains and to design architectures that can more effectively manage task interference.

1.6.2 Privacy and Ethical Considerations

Mental health data is highly sensitive, raising significant privacy concerns. Federated MTL [8] offers one solution by enabling decentralised model training without requiring the exchange of raw data. The work by Noraset et al. [14] on language-agnostic frameworks also addresses privacy considerations by operating on aggregated and anonymised predictions. However, ethical challenges remain critical, including algorithmic bias, false positives and false negatives, and the potential harms associated with incorrect or inappropriate interventions — require careful consideration as such systems move closer to real-world and clinical deployment [1].

1.6.3 Data Scarcity and Annotation Quality

Annotated mental health datasets remain scarce, particularly for less common disorders and in low-resource languages. Chiong et al. [16] demonstrated that careful feature engineering can partially compensate for data limitations, however, the sample efficiency of MTL makes it particularly well-suited to this setting. Dos Santos et al. [4] emphasised the importance of self-report validation as a strategy for improving annotation quality in social media-based mental health datasets.

1.6.4 Temporal Dynamics and Longitudinal Modelling

Mental health conditions evolve over time, yet most current MTL models operate on static text snapshots. Incorporating temporal dynamics — such as changes in linguistic patterns, social media activity, and emotional expression over time — represents a key research direction in this field [5].

1.7 Future Directions

Several promising research directions are emerging in MTL for mental health applications. Multimodal MTL that jointly learns from text, speech, and physiological signals is gaining increasing attention — Chen et al. [17] recently proposed a text-guided multimodal depression detection framework using cross-modal feature reconstruction and decomposition, where text serves as the primary modality for guiding the learning of audio representations. Generative MTL approaches [10] that combine classification with text generation tasks may further enhance representation learning by encouraging richer and more structured latent representations. Large language model (LLM)-based MTL

represents a new frontier, where pre-trained models are fine-tuned on multiple mental health tasks simultaneously using instruction tuning or prompt-based learning strategies [13]. Finally, explainable MTL that provides clinicians with interpretable intermediate representations — such as detected cognitive distortions or emotional patterns — remain essential for clinical adoption and real-world deployment [15].

1.8 Conclusion

Multi-task learning has emerged as a powerful paradigm for mental health disorder classification from user-written text, consistently outperforming single-task approaches by leveraging shared representations across related conditions. From the landmark work of Ophir et al. [7] sk baselines, to recent advances in federated MTL [8], cross-lingual frameworks [14], and prompt-based learning [13], the field has made substantial progress between 2020 and 2025. Key themes identified in this chapter include the importance of theory-driven auxiliary task design, the effectiveness of attention mechanisms for task-specific feature extraction, and the growing emphasis on privacy-preserving, scalable, and equitable modelling approaches. As the field continues to mature, the integration of MTL with multimodal data, large language models, and clinically validated outcome measures promises to accelerate the translation of these computational methods into real-world mental health screening and intervention tools.

Chapter 2

Background

2.1 Introduction

The convergence of artificial intelligence (AI) and healthcare has catalysed transformative advances in diagnostic precision, predictive modelling, and patient monitoring over the past decade [18]. Among the most promising intersections of these disciplines lies the application of natural language processing (NLP) to the analysis of user-generated text for mental health assessment—a domain where the sheer volume of digital communication offers unprecedented opportunities for early detection and population-level screening of psychiatric conditions [19]. Mental health disorders, which the World Health Organization estimates affect approximately one in eight individuals globally [20], remain critically underdiagnosed and undertreated, owing in part to the stigma associated with seeking professional help, the shortage of mental health professionals, and the limited scalability of traditional clinical screening instruments [1].

The proliferation of social media platforms and online communities has created vast repositories of naturalistic text in which individuals voluntarily disclose personal experiences, emotional states, and psychological struggles [2, 3]. Platforms such as Reddit host dedicated communities—known as subreddits—where users share first-person accounts of living with conditions including depression, anxiety, post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorder (ADHD), and bipolar disorder. These digital traces constitute a rich, ecologically valid data source for computational psychiatry, offering linguistic signals that complement and, in some contexts, surpass the sensitivity of structured clinical assessments [6].

However, the automated classification of mental health conditions from unstructured text presents formidable challenges. The linguistic manifestations of distinct disorders frequently overlap—negative affect, anhedonia, and rumination, for instance, are common to both depression and anxiety—creating substantial semantic ambiguity [5]. Furthermore, user-generated text is inherently noisy, characterised by informal grammar, slang,

sarcasm, abbreviations, and implicit emotional signals that resist straightforward computational analysis. Class imbalance, the absence of clinical ground-truth labels, and the ethical sensitivity of mental health data further compound the difficulty [4].

Transformer-based language models, beginning with the seminal work of Vaswani et al. [21] and subsequently refined through pre-trained architectures such as BERT [22], DistilBERT [23], and DeBERTa [24], have established new benchmarks across virtually all NLP tasks by capturing deep contextual dependencies through self-attention mechanisms. These models are particularly well-suited to mental health text classification, as the diagnostic relevance of a given utterance often depends on subtle contextual cues distributed across an entire document rather than isolated keywords.

Multi-task learning (MTL), a paradigm in which a model is trained simultaneously on multiple related objectives through shared representations [25], offers a complementary advantage. By jointly optimising mental health disorder classification alongside an auxiliary sentiment analysis task, MTL architectures can leverage the affective commonalities between tasks to learn more robust, generalisable representations—a property especially valuable when labelled clinical datasets are small and imbalanced [7, 8].

This chapter establishes the theoretical and contextual foundations for the present thesis. Section 2.2 surveys the broader landscape of AI in healthcare. Section 2.3 provides a clinically grounded overview of the mental health disorders addressed in this work. Section 2.4 examines NLP techniques relevant to mental health text analysis. Section 3.3.1 describes the two datasets employed throughout the experimental programme. Section 4.5 synthesises the state of the art in transformer-based mental health classification and multi-task learning. Section 2.7 articulates the research problem and objectives.

2.2 Artificial Intelligence in Healthcare

2.2.1 Overview of AI Applications in Clinical Settings

Artificial intelligence has permeated virtually every domain of healthcare, from medical imaging and genomics to drug discovery and clinical decision support [18, 26]. Machine learning algorithms—encompassing supervised, unsupervised, and reinforcement learning paradigms—are now deployed in radiology for automated lesion detection [27], in pathology for histological classification [28], and in electronic health record (EHR) analysis for predictive modelling of patient outcomes [29]. Deep learning architectures, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have achieved diagnostic performance comparable to or exceeding that of domain experts in specific clinical tasks [30].

Within the mental health domain, AI applications span several complementary modalities. Speech analysis systems extract prosodic features—pitch variability, speaking rate,

and pause patterns—that correlate with depressive symptomatology [31]. Computer vision approaches analyse facial micro-expressions and gaze patterns for affect recognition [32]. Physiological signal processing monitors heart rate variability, electrodermal activity, and sleep patterns through wearable devices for longitudinal mental health tracking [33]. However, it is the analysis of textual data—clinical notes, therapy transcripts, and social media posts—that has attracted the most intensive research attention, owing to the ubiquity, accessibility, and richness of written language as a carrier of psychological information [19].

2.2.2 Natural Language Processing in Medical Analysis

NLP has emerged as a pivotal technology for extracting actionable clinical insights from unstructured text, which constitutes an estimated 80% of all healthcare data [34]. Clinical NLP applications include automated ICD coding, adverse drug event detection, phenotyping from clinical narratives, and de-identification of protected health information [35]. In psychiatry specifically, NLP has been applied to suicide risk assessment from crisis hotline transcripts [9], depression screening from social media activity [16], and psychosis prediction from clinical interview transcripts [36].

The transition from rule-based and feature-engineered approaches to deep learning-based NLP has been transformative. Early systems relied on manually curated lexicons—such as the Linguistic Inquiry and Word Count (LIWC) dictionary [37]—to quantify psychological constructs from word frequencies. While effective for population-level analyses, these approaches are unable to capture contextual semantics, negation, irony, or the compositional meaning that emerges from syntactic structure. Transformer-based language models address these limitations by learning contextualised token representations that encode both local and global semantic dependencies [22].

2.2.3 Ethical, Privacy, and Interpretability Considerations

The deployment of AI in mental health contexts raises critical ethical concerns. Privacy is paramount: mental health data is classified as sensitive personal information under regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) [38]. Algorithmic bias presents a further risk, as models trained on English-language social media data may fail to generalise across linguistic, cultural, and demographic boundaries [39]. False negatives—failing to detect genuine psychiatric distress—carry potentially severe consequences, including missed opportunities for intervention in suicidal individuals.

Interpretability remains an open challenge: transformer models operate as black-box systems whose internal representations resist straightforward clinical interpretation [40]. While attention weight visualisation has been proposed as a post-hoc explanation mecha-

nism, recent work has demonstrated that attention distributions are unreliable indicators of feature importance [41]. These considerations underscore the necessity of treating AI-based mental health classification as a screening aid rather than a diagnostic instrument, always subject to clinical oversight and validation.

2.3 Mental Health Disorders

This section provides a clinically grounded overview of the eight mental health conditions addressed in the present thesis, encompassing the diagnostic categories present in both the Reddit dataset (ADHD, Anxiety, Bipolar, Depression, PTSD, and the *None* control class) and the MA dataset (Anxiety, Bipolar, Depression, Normal, Personality Disorder, Stress, and Suicidal). For each disorder, the discussion considers the clinical definition, characteristic symptoms, linguistic manifestations in written text, and the specific challenges that each condition poses for automated classification.

2.3.1 Anxiety Disorders

Anxiety disorders are characterised by excessive and persistent worry, fear, and apprehension that are disproportionate to the actual threat posed by the triggering stimulus [42]. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) recognises several subtypes, including generalised anxiety disorder (GAD), social anxiety disorder, panic disorder, and specific phobias.

In written text, anxiety manifests through linguistic markers including elevated use of future-tense constructions (*"what if"*, *"I'm worried that"*), uncertainty hedges (*"maybe"*, *"I think"*, *"possibly"*), somatic symptom descriptions (*"my heart races"*, *"I can't breathe"*), and catastrophising language [43]. A distinguishing feature of anxiety discourse, as identified in the clinical keyword analysis employed in Chapter 2, is the prevalence of acute temporal markers—words such as *"suddenly"*, *"right now"*, and *"at this moment"*—reflecting the episodic nature of anxious cognition. The primary classification challenge lies in the semantic overlap between anxiety and stress, as both conditions share lexical features related to worry, tension, and physiological arousal.

2.3.2 Depression

Major depressive disorder (MDD) is defined by persistent depressed mood, anhedonia (loss of interest or pleasure), and a constellation of cognitive, somatic, and behavioural symptoms lasting at least two weeks [42]. Depression is the leading cause of disability worldwide and a major contributor to the global burden of disease [20].

The linguistic signature of depression has been extensively studied. Depressive text characteristically exhibits elevated first-person singular pronoun usage (*"I"*, *"me"*, *"my"*),

absolutist language (*"always"*, *"never"*, *"nothing"*), past-tense orientation, and reduced lexical diversity [44, 45]. The classification of depression from text is complicated by its frequent comorbidity with anxiety (co-occurrence rates of approximately 50%) and the overlap of linguistic markers with suicidal ideation, as depressive episodes are a primary risk factor for suicidal behaviour [4]. This Depression–Suicidal confusion is consistently observed across all architectural families in the experimental results (Chapter 3).

2.3.3 Bipolar Disorder

Bipolar disorder is characterised by alternating episodes of mania (or hypomania) and depression, with marked shifts in mood, energy, and activity levels [42]. The textual manifestations of bipolar disorder are consequently phase-dependent: manic episodes produce grandiose, high-energy language with rapid topic shifts and elevated positive affect, whereas depressive episodes yield linguistic patterns indistinguishable from unipolar depression [46].

This phase-dependent variability poses a substantial challenge for text classification, as a single post captured during a depressive episode may be indistinguishable from a depression-class sample. The confusion between bipolar and depression classes is observed in the per-class analyses reported in Chapter 3, where Personality Disorder shows its primary confusion with both Depression and Bipolar categories.

2.3.4 Attention-Deficit/Hyperactivity Disorder (ADHD)

ADHD is a neurodevelopmental disorder characterised by persistent patterns of inattention, hyperactivity, and impulsivity that interfere with functioning and development [42]. Unlike the mood-driven disorders discussed above, ADHD presents primarily through executive function deficits rather than affective disturbance.

In written text, ADHD discourse is characterised by distractibility markers—tangential narratives, incomplete sentences, topic-switching mid-paragraph—and frequent references to organisational difficulties, procrastination, and time management challenges. The relative distinctiveness of ADHD's lexical profile compared to mood disorders makes it one of the more tractable classification targets, though its representation is limited to the Reddit dataset in the present thesis.

2.3.5 Post-Traumatic Stress Disorder (PTSD)

PTSD develops following exposure to a traumatic event and is characterised by intrusive re-experiencing symptoms, avoidance behaviours, negative alterations in cognition and mood, and hyperarousal [42]. Textual descriptions of PTSD frequently include trauma

narratives, hypervigilance language, emotional numbing expressions, and descriptions of flashbacks and nightmares.

The classification challenge for PTSD lies in its comorbidity with depression and anxiety—all three conditions share features of negative affect, sleep disturbance, and concentration difficulties [5]. Additionally, individuals with PTSD may employ avoidance strategies that result in sparse or euphemistic language, creating signal attenuation in textual data.

2.3.6 Personality Disorder

Personality disorders encompass a heterogeneous group of conditions characterised by enduring, maladaptive patterns of behaviour, cognition, and inner experience that deviate markedly from cultural expectations [42]. The DSM-5 organises personality disorders into three clusters: Cluster A (odd/eccentric), Cluster B (dramatic/erratic, including borderline personality disorder [BPD]), and Cluster C (anxious/fearful).

In the MA dataset, *Personality Disorder* constitutes a single aggregated category—the smallest class in the dataset—encompassing multiple subtypes. This heterogeneity, combined with small sample size (447 test samples compared to 6,571 for Normal), makes Personality Disorder the most challenging classification target. Experimental results consistently identify it as the hardest class, with primary confusion against Depression and Bipolar categories (Chapter 3, Table 3.11).

2.3.7 Stress

Although not classified as a psychiatric disorder *per se*, chronic stress represents a significant risk factor for the development of clinical mental health conditions and is recognised as a distinct psychological state warranting monitoring [47]. Stress discourse in text exhibits overlap with anxiety—both involve descriptions of tension, worry, and physiological arousal—but stress language tends to reference specific external stressors (work, relationships, finances) whereas anxiety discourse is more self-referential and diffuse.

The Stress–Anxiety distinction constitutes one of the more subtle classification boundaries in the MA dataset, as both categories share a substantial proportion of their lexical and semantic feature space. The clinical keyword features introduced in Chapter 2 specifically target this boundary through differential scoring between acute (anxiety-associated) and chronic (stress-associated) temporal markers.

2.3.8 Suicidal Ideation

Suicidal ideation refers to thoughts, plans, or preoccupation with ending one's life, ranging from passive ideation (*"I wish I weren't alive"*) to active planning (*"I have a plan"*) [42].

The detection of suicidal language in user-generated text is arguably the most clinically consequential application of mental health NLP, given the direct association between suicidal ideation and suicide attempts [9].

Linguistic markers of suicidal ideation include expressions of hopelessness, perceived burdensomeness (*"everyone would be better off without me"*), thwarted belongingness, and absolutist language [45]. The primary classification challenge is the overlap with depression—suicidal ideation frequently occurs in the context of depressive episodes, creating genuine label ambiguity that is reflected in the systematic Depression–Suicidal confusion observed across all architectures in Chapter 3.

2.3.9 Summary of Classification Challenges

Table 2.1 synthesises the linguistic characteristics and classification challenges associated with each disorder.

2.4 Natural Language Processing for Mental Health Analysis

2.4.1 Fundamentals of Text Classification

Text classification—the task of assigning predefined categorical labels to textual documents—is a foundational NLP problem with applications spanning sentiment analysis, spam detection, topic categorisation, and, as in the present work, clinical text classification [48]. The evolution of text classification methods has progressed through several paradigms:

1. **Rule-based systems** relying on manually crafted patterns and dictionaries (e.g., LIWC for psychological text analysis; [37]).
2. **Feature-engineered machine learning** using bag-of-words, TF-IDF, or n-gram representations with classifiers such as support vector machines (SVMs), random forests, and logistic regression [49].
3. **Distributed representation models** employing static word embeddings (Word2Vec, [50]; GloVe, [51]) as input features to neural classifiers.
4. **Pre-trained contextualised models** using transfer learning from large-scale language models fine-tuned on downstream tasks [22].

The present thesis investigates approaches spanning paradigms 3 and 4: scratch-built attention models with GloVe embeddings (Family I in Chapter 2) and fine-tuned transformer backbones including DistilBERT and DeBERTa (Family II), complemented by a

Table 2.1: Summary of mental health disorders, linguistic markers, and classification challenges.

Disorder	**Key Linguistic Markers**	**Primary Classification Challenge**	**Dataset(s)**
Anxiety	Future-tense, uncertainty hedges, acute temporal markers, somatic descriptions	Overlap with Stress	Reddit, MA
Depression	First-person pronouns, absolutist language, past-tense, reduced lexical diversity	Comorbidity with Suicidal, Bipolar	Reddit, MA
Bipolar	Phase-dependent: grandiose (manic) vs. depressive language	Depressive phase indistinguishable from Depression	Reddit, MA
ADHD	Tangential narratives, disorganisation markers, executive function references	Distinct lexical profile, but Reddit-only	Reddit
PTSD	Trauma narratives, hypervigilance, avoidance language, emotional numbing	Comorbidity with Depression and Anxiety	Reddit
Personality Disorder	Interpersonal instability, emotional dysregulation, identity disturbance	Heterogeneous category, smallest class	MA
Stress	External stressor references, tension, physiological arousal	Overlap with Anxiety	MA
Suicidal	Hopelessness, perceived burdensomeness, absolutist language	Overlap with Depression	MA
None / Normal	General discourse, absence of clinical markers	Distinct lexical profile (control class)	Reddit / MA

hybrid architecture that fuses transformer representations with static embeddings and handcrafted linguistic features (Family III).

2.4.2 Sentiment Analysis and Emotion Detection

Sentiment analysis—the computational determination of subjective polarity (positive, neutral, negative) in text—is closely related to mental health classification, as psychiatric conditions are fundamentally characterised by affective disturbance [52]. Emotion detection extends this to finer-grained affective categories (joy, sadness, anger, fear, surprise, disgust), providing a richer representation of the emotional content of text [53].

In the multi-task learning framework adopted in this thesis, sentiment analysis serves as the auxiliary task alongside the primary mental health classification objective. The rationale for this design, detailed in Chapter 2, rests on three pillars: (i) affective overlap between sentiment dimensions and mental health constructs, (ii) shared linguistic markers between sentiment and disorder expression, and (iii) the regularisation benefit of the auxiliary gradient signal [25].

2.4.3 Transformer-Based Language Models

The Transformer architecture [21] fundamentally reshaped NLP by replacing recurrent computation with self-attention, enabling parallel processing of input sequences and the capture of long-range dependencies regardless of positional distance. The core innovation—scaled dot-product attention—computes a weighted sum of value vectors based on the compatibility between queries and keys:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{2.1}$$

where d_k denotes the dimensionality of the key vectors. Multi-head attention extends this by projecting queries, keys, and values into h parallel subspaces, enabling the model to attend to information from different representational perspectives simultaneously.

The pre-training paradigm introduced by BERT (Bidirectional Encoder Representations from Transformers; [22]) demonstrated that training a deep transformer on large-scale unsupervised objectives—masked language modelling (MLM) and next sentence prediction—produces representations that transfer effectively to downstream tasks through fine-tuning. This transfer learning approach has been refined through several architectural variants relevant to the present thesis:

DistilBERT [23] is a compressed variant of BERT obtained through knowledge distillation, retaining 97% of BERT's language understanding capability while being 60% faster and 40% smaller (66M parameters, 6 layers). Its efficiency makes it suitable for resource-constrained training environments, such as the Google Colab infrastructure employed in

this thesis.

DeBERTa (Decoding-enhanced BERT with Disentangled Attention; [24]) introduces disentangled attention, wherein each token is represented by two separate vectors encoding content and position. The attention score between tokens i and j decomposes into content-to-content, content-to-position, and position-to-content interactions, enabling more expressive modelling of syntactic and semantic relationships. DeBERTa-base comprises 139M parameters across 12 layers.

DeBERTa-v3 extends the architecture with Replaced Token Detection (RTD) pre-training, inspired by ELECTRA [54], where a generator produces plausible token replacements and a discriminator learns to identify altered tokens—a more sample-efficient objective than masked language modelling. DeBERTa-v3-base contains 184M parameters.

These three architectures, spanning a capacity range from 66M to 184M parameters, constitute the pre-trained backbones investigated in the experimental programme of Chapters 2 and 3.

2.4.4 Challenges of Social Media Text for NLP

The application of NLP models to social media text introduces several domain-specific challenges that distinguish it from the well-structured corpora typically used for model pre-training. Table 2.2 summarises these challenges and their impact on mental health classification.

Table 2.2: Challenges of social media text for NLP-based mental health analysis.

Challenge	**Description**	**Impact on Classification**
Informal language	Colloquialisms, slang, non-standard grammar	Tokenisation errors, out-of-vocabulary tokens
Abbreviations and acronyms	Domain-specific shorthand (e.g., *"meds"*, *"dx"*)	Semantic information loss
Sarcasm and irony	Expressed sentiment contradicts literal meaning	Reversed polarity detection
Implicit emotional signals	Emotional states conveyed through narrative rather than explicit labels	Requires deep contextual understanding
Noise	URLs, emojis, formatting artefacts, code-switching	Preprocessing burden
Self-diagnosis uncertainty	Users may incorrectly self-report conditions	Label noise in training data
Contextual ambiguity	Meaning depends on broader conversational context	Single-post classification loses context
Demographic bias	Platform demographics skew toward specific age groups and regions	Generalisability limitations

These challenges motivate the preprocessing pipeline described in Chapter 2 (emoji

demojisation, URL removal, whitespace normalisation, case folding) and the selection of transformer architectures capable of contextual semantic understanding over surface-level pattern matching.

2.5 Datasets Description

This section describes the two datasets employed throughout the experimental programme. The naming conventions and class labels established here are maintained consistently through Chapters 2 and 3.

2.5.1 Reddit Dataset

The **Reddit dataset** consists of user-generated posts collected from mental health-related subreddits—dedicated online communities where individuals discuss personal experiences with specific psychiatric conditions. Reddit's pseudonymous structure encourages candid self-disclosure, producing text of high ecological validity for mental health research [3, 6].

The dataset comprises six classes, as described in Table 2.3.

Table 2.3: Reddit dataset class descriptions.

Class	Description	Characteristics
ADHD	Posts from ADHD-related subreddits	Executive function difficulties, disorganisation narratives
Anxiety	Posts from anxiety-related subreddits	Worry, fear, somatic symptoms, uncertainty expressions
Bipolar	Posts from bipolar disorder subreddits	Mood fluctuation narratives, phase-specific language
Depression	Posts from depression-related subreddits	Negative affect, anhedonia, hopelessness, self-referential language
PTSD	Posts from PTSD-related subreddits	Trauma narratives, hypervigilance, avoidance language
None	Posts from general (non-clinical) subreddits	General discourse, absence of clinical markers

The dataset is distributed as CSV files with pre-defined train/validation/test partitions, comprising approximately 13,727 training samples, 1,488 validation samples, and 1,488 test samples. Three textual modalities are available: post body, title, and a combined title-plus-body representation.

Strengths. Reddit data offers several advantages for mental health NLP research. The platform's community structure provides a natural form of weak labelling—posts in *r/depression* are more likely to contain depression-related content than posts from general subreddits. The anonymity of Reddit encourages frank discussion of sensitive topics,

yielding text that is more emotionally authentic than content from identity-linked platforms. Additionally, the diversity of writing styles across users ensures lexical variability that promotes model robustness.

Challenges. Several factors complicate the use of Reddit data. Self-diagnosis uncertainty means that community membership does not guarantee clinical diagnosis—users may participate in mental health subreddits out of curiosity, to support others, or based on self-assessed symptoms that do not meet diagnostic criteria. The informal register of Reddit text includes slang, profanity, sarcasm, and abbreviations that challenge both tokenisation and semantic analysis. Finally, the relatively small dataset size (∼13,700 training samples) increases the risk of overfitting, particularly for high-capacity transformer models.

2.5.2 MA Dataset

The **MA dataset** (distributed as a JSON file, hereinafter referred to as the MA dataset, consistent with Chapter 2 nomenclature) contains user-written text samples annotated for seven mental health categories. Compared to the Reddit dataset, the MA dataset is substantially larger and exhibits a semi-formal clinical register, suggesting a more curated annotation process.

The dataset comprises seven classes, as described in Table 2.4.

Table 2.4: MA dataset class descriptions.

Class	Description	Characteristics
Anxiety	Texts expressing anxiety-related content	Worry, fear, panic, somatic anxiety symptoms
Bipolar	Texts expressing bipolar-related content	Mood swings, manic and depressive episodes
Depression	Texts expressing depressive content	Persistent sadness, anhedonia, hopelessness
Normal	Texts without clinical mental health indicators	General discourse, neutral emotional content
Personality Disorder	Texts expressing personality disorder-related content	Interpersonal instability, identity disturbance, emotional dysregulation
Stress	Texts expressing stress-related content	External stressor references, tension, coping difficulty
Suicidal	Texts expressing suicidal ideation	Hopelessness, death wishes, self-harm references

The dataset is distributed with pre-defined training and test partitions, with a 90/10 stratified train/validation split applied programmatically. The approximate partition sizes are 67,894 training samples, 16,974 validation samples, and 21,218 test samples.

Strengths. The MA dataset offers a more comprehensive taxonomy of mental health conditions than the Reddit dataset, incorporating Personality Disorder, Stress, and Suicidal ideation as distinct categories. Its larger size provides more reliable gradient estimates during training and reduces the risk of overfitting. The semi-formal register may also produce cleaner linguistic signals with less noise from informal language.

Challenges. The dataset exhibits notable class imbalance—the Normal and Depression classes each contain over 6,000 test samples, whereas Personality Disorder has only 447 test samples. This imbalance necessitates mitigation strategies such as balanced class weights and careful metric selection (macro F1 and MCC over raw accuracy). The Suicidal class presents particular sensitivity concerns, as misclassification of suicidal text carries severe ethical implications. Additionally, the overlap between clinically adjacent categories (Anxiety–Stress, Depression–Suicidal, Bipolar–Depression) creates genuine label ambiguity that even human annotators would find challenging to resolve.

2.5.3 Comparative Dataset Analysis

Table 2.5 provides a systematic comparison of the two datasets across multiple dimensions relevant to model design and evaluation. Table 2.6 summarises the comparative advantages and limitations.

Table 2.5: Comparative analysis of the Reddit and MA datasets.

Dimension	**Reddit Dataset**	**MA Dataset**
Number of classes	6	7
Total training samples	∼13,727	∼67,894
Total test samples	∼1,488	∼21,218
Text register	Informal social media	Semi-formal clinical
Source	Reddit subreddits	Curated clinical texts
Distribution format	CSV (pre-split)	JSON (pre-split)
Label source	Community membership (weak labels)	Annotation-based
Unique conditions	ADHD, PTSD	Personality Disorder, Stress, Suicidal
Shared conditions	Anxiety, Bipolar, Depression	Anxiety, Bipolar, Depression
Control class	None	Normal
Sentiment labels	Derived (lexicon-based)	Derived (lexicon-based)
Class imbalance severity	Moderate	High (Personality Disorder underrepresented)
Noise level	High (slang, sarcasm, abbreviations)	Moderate

The use of two complementary datasets serves a dual purpose in the experimental programme. First, it enables the assessment of model generalisation across distinct tex-

Table 2.6: Comparative advantages and limitations of the two datasets.

Aspect	Reddit Dataset	MA Dataset
Advantages	Ecological validity; authentic language; natural community-based weak labelling	Larger size; broader taxonomy; semi-formal register; more distinctive class boundaries
Limitations	Small size; noisy text; self-diagnosis uncertainty; limited to 6 classes	Class imbalance; clinical-adjacent category overlap; sensitivity of Suicidal class
Expected challenges	Overfitting risk; high confusion between mood disorders; PTSD–Depression overlap	Personality Disorder underrepresentation; Depression–Suicidal confusion; Anxiety–Stress ambiguity

tual registers and label taxonomies—a model that performs well on both informal Reddit text and semi-formal MA text demonstrates robustness to domain variation. Second, the different class compositions provide complementary perspectives on mental health classification: the Reddit dataset includes neurodevelopmental (ADHD) and trauma-related (PTSD) conditions absent from the MA dataset, while the MA dataset provides coverage of Personality Disorder, Stress, and Suicidal ideation. This dual-dataset strategy, as detailed in the methodology of Chapter 2, informs both the single-task and multi-task experimental configurations.

2.6 State of the Art

This section synthesises the current state of research in transformer-based mental health classification and multi-task learning for NLP. The discussion is organised around five thematic pillars: transformer-based mental health classification, multi-task learning for NLP, social media mental health detection, evaluation methodologies, and emerging directions.

2.6.1 Transformer-Based Mental Health Classification

The application of pre-trained transformer models to mental health text classification has yielded substantial performance improvements over traditional feature-engineered approaches. Table 2.7 summarises representative recent works.

Table 2.7: Representative state-of-the-art approaches for transformer-based mental health classification.

Study	Objective	Dataset	Architecture	Metrics	Results
Kim et al. [3]	Multi-class mental illness detection	Reddit (6 classes)	Deep NN with hard parameter sharing	Accuracy	Robust multi-class discrimination
Gkotsis et al. [6]	Mental health characterisation	Reddit (11 themes)	CNN and RNN	Accuracy	91.08% (binary); 71.37% (theme)
Ji et al. [9]	Suicidal ideation detection	Multiple corpora	Various ML/DL (survey)	F1, AUC	Transformers are state-of-the-art
Chiong et al. [16]	Depression detection	Twitter/Reddit	Feature-engineered ML	Accuracy, F1	Feature engineering partially compensates for data scarcity
Zhang & Guo [13]	Multilevel depression detection	Severity datasets	Fine-grained prompt learning	F1, Accuracy	Multi-granularity prompts capture depressive features

Kim et al. [3] demonstrated that deep neural networks with shared representations can effectively discriminate between six mental health conditions from Reddit posts. Their hard parameter sharing architecture—where a shared encoder feeds task-specific output layers—is conceptually aligned with the MTL framework adopted in the present thesis. However, their work predates the widespread adoption of pre-trained transformers and does not investigate the benefits of transfer learning from large-scale language models.

Gkotsis et al. [6] provided an early but influential benchmark, achieving 91.08% accuracy for binary mental illness detection and 71.37% weighted accuracy for 11-category theme classification on Reddit data. Their finding that deep learning approaches substantially outperform traditional classifiers foreshadowed the transformer revolution, though their architectures (CNNs and RNNs) lack the self-attention mechanisms that enable long-range dependency modelling.

Zhang and Guo [13] proposed a particularly innovative approach: fine-grained prompt learning for multilevel depression detection, where multiple prompt sets capture

depressive features at different granularity levels. This implicit multi-task formulation demonstrates the versatility of the MTL concept beyond traditional hard/soft parameter sharing paradigms.

2.6.2 Multi-Task Learning for NLP

Multi-task learning has a long history in NLP, with applications spanning named entity recognition, part-of-speech tagging, syntactic parsing, and sentiment analysis. Table 2.8 summarises key MTL contributions relevant to mental health classification.

Table 2.8: Multi-task learning approaches relevant to mental health NLP.

Study	Objective	Tasks	Architecture	Key Results	Limitations
Ophir et al. [7]	Suicide risk detection	Personality, psychosocial, disorder, suicide	Hierarchical MTL	MTM AUC 0.697–0.746 vs. STM 0.621–0.629	Single platform
Huang et al. [8]	Joint disorder diagnosis	ASD, ADHD, schizophrenia	Federated MTL	Improved over single-task	MRI-based; not text
Zampieri et al. [55]	Offensive language ID	Multiple subtasks	MTL with shared repr.	Improved over single-task	Not mental health
Uban et al. [5]	Linguistic markers	Emotion, style, content	DL with auxiliary tasks	Emotional markers improve differentiation	Binary only
Dos Santos et al. [4]	Depression & anxiety prediction	Joint classification	Multi-label classifiers	Improves comorbid detection	Portuguese; small corpus

Ophir et al. [7] provided one of the most compelling demonstrations of MTL's value for mental health. Their hierarchical multi-task model (MTM), which cascades predictions through theory-driven layers—from personality traits through psychosocial risks to psychiatric disorders—achieved substantially higher AUC (0.697–0.746) than the single-task model (0.621–0.629) for suicide risk prediction from Facebook posts. Critically, content analyses revealed that successful predictions relied on a broad range of linguistic features rather than explicit suicide-related keywords, demonstrating that the MTL hierarchy encouraged the model to learn clinically meaningful intermediate representations.

Huang et al. [8] extended the MTL paradigm to the federated setting, proposing a multi-gate mixture-of-experts architecture for joint diagnosis of autism spectrum disorder, ADHD, and schizophrenia. Although their work operates on MRI neuroimaging data rather than text, their methodological innovations—including federated contrastive

learning and task-specific expert gating—are directly transferable to text-based multi-task systems and inform the mixture-of-experts direction identified as future work in Chapter 3.

2.6.3 Social Media Mental Health Detection

The intersection of social media analysis and mental health detection has produced a growing body of literature leveraging platform-specific features. Table 2.9 presents additional representative approaches.

Table 2.9: Social media-based mental health detection approaches.

Study	Platform	Method	Conditions	Key Innovation
De Choudhury et al. [44]	Twitter	SVM with behavioural features	Depression	First large-scale social media depression study
Tariq et al. [2]	Social media	Co-training	Mental illness	Semi-supervised learning for data scarcity
Li et al. [11]	Social media	Multifeature fusion recurrent attention	Suicide risk levels	Self-attention + linguistic features; +3.7% F1
Zogan et al. [12]	Twitter	DepressionNet (CNN + attention GRU + summarisation)	Depression	Hybrid summarisation as implicit auxiliary task
Noraset et al. [14]	Thai social media	Language-agnostic MTL (LAPoMM)	Emotions, sentiments, suicidal tendencies	Cross-lingual transfer with real-world validation
Chen et al. [17]	Multimodal	Text-guided cross-modal feature reconstruction	Depression	Text as core modality guiding audio learning

Zogan et al. [12] introduced DepressionNet, which incorporates a hybrid extractive-abstractive summarisation strategy to filter relevant content from user tweet sequences before classification. Although not explicitly framed as multi-task learning, the summarisation component functions as an implicit auxiliary task—a design philosophy resonant with the sentiment auxiliary task employed in the present thesis.

Noraset et al. [14] addressed the critical gap of non-English mental health detection through their Language-Agnostic Population-level Mental Health Monitoring (LAPoMM) framework. Their validation against real-world administrative data on depression cases and suicide attempts demonstrates the potential of NLP-based monitoring to complement traditional epidemiological surveillance—a finding that underscores the broader public health relevance of the research programme presented in this thesis.

2.6.4 Critical Comparative Analysis

Table 2.10 positions the present thesis relative to prior work across several methodological dimensions.

Table 2.10: Comparative analysis of state-of-the-art approaches.

Criterion	**Trad. ML (pre-2018)**	**Transformer Single-Task**	**Transformer MTL**	**Present Thesis**
Contextual understanding	None (bag-of-words)	Full (self-attention)	Full (shared encoder)	Full (DistilBERT, DeBERTa)
Transfer learning	None	Pre-trained + fine-tuned	Pre-trained + fine-tuned	Pre-trained + fine-tuned
Multi-disorder coverage	Typically binary	Multi-class (3–11)	Multi-class + auxiliary	6-class (Reddit) + 7-class (MA)
Auxiliary task	None	None	Varied	Sentiment analysis (3-class)
Loss weighting	N/A	Standard CE	Fixed or learned	Dynamic uncertainty + static
Evaluation	Accuracy, F1	Accuracy, F1, AUC	Varies	10+ metrics incl. MCC, Kappa, Top-2
Dual-dataset evaluation	Rare	Occasional	Rare	Yes (Reddit + MA)

2.6.5 Identified Research Gaps

The synthesis of the state of the art reveals several gaps that the present thesis aims to address:

1. **Limited comparative evaluation of transformer variants for mental health.** While individual studies have applied BERT, RoBERTa, or domain-specific models (e.g., MentalBERT), systematic comparisons across the DistilBERT–DeBERTa capacity spectrum on the same datasets are scarce.

2. **Underexplored multi-task learning for multi-class mental health classification.** Most MTL studies focus on binary classification (disorder vs. control) or are limited to depression and suicide risk. The joint classification of 6–7 disorders with an auxiliary sentiment task remains insufficiently explored.

3. **Absence of progressive architectural evaluation.** Existing studies typically present a single architecture without contextualising its performance against simpler

baselines (e.g., scratch-built attention models) or complementary approaches (e.g., feature fusion, ensemble methods).

4. **Inconsistent evaluation protocols.** Many studies report only accuracy and F1, neglecting metrics such as MCC and Cohen's Kappa that are more robust to class imbalance—a critical limitation given the inherent imbalance of mental health datasets.

5. **Limited investigation of hybrid architectures.** The fusion of transformer representations with static embeddings and handcrafted linguistic features for mental health classification remains largely unexplored.

2.7 Problem Statement and Research Objectives

2.7.1 Problem Statement

Despite the promising advances surveyed in Section 4.5, the automated classification of mental health disorders from user-generated text remains an open and challenging problem. Current approaches are limited by several factors:

- **Insufficient contextual understanding**: Traditional and early deep learning approaches fail to capture the nuanced, context-dependent linguistic markers that differentiate clinically similar conditions (e.g., Depression vs. Suicidal ideation, Anxiety vs. Stress).

- **Underutilisation of multi-task learning**: The potential of joint training on related tasks—such as disorder classification and sentiment analysis—to improve representation learning and generalisation remains insufficiently explored for multi-class mental health taxonomies.

- **Limited architectural comparison**: The relative merits of different transformer architectures (DistilBERT vs. DeBERTa vs. DeBERTa-v3) for mental health text classification are poorly characterised, particularly across datasets of varying size and register.

- **Evaluation narrowness**: Reliance on accuracy alone provides a misleading picture of classifier performance on imbalanced datasets; a more comprehensive metric suite is needed.

- **Lack of progressive methodology**: Few studies contextualise transformer performance against simpler baselines, making it difficult to attribute performance gains to specific architectural innovations.

The central research question of this thesis is therefore:

> *To what extent can multi-task learning with transformer-based architectures improve the classification of multiple mental health disorders from user-written text, and how do different architectural choices—from scratch-built attention models to pre-trained transformers to hybrid feature-fusion systems—compare across complementary datasets?*

2.7.2 Research Objectives

The general objective of this thesis is to design, implement, and evaluate a multi-task learning framework for mental health disorder classification from user-written text, leveraging transformer-based NLP architectures to achieve robust, generalisable, and comprehensive classification performance.

The specific objectives are:

1. **Evaluate the effectiveness of attention-based architectures** by implementing and benchmarking scratch-built transformer models with linguistically-informed features (GloVe embeddings, POS/DEP/SRL embeddings) on both the Reddit (6-class) and MA (7-class) datasets.

2. **Assess the impact of pre-trained transformer fine-tuning** by comparing DistilBERT and DeBERTa variants—spanning 66M to 184M parameters—on the same classification tasks, quantifying the performance gains attributable to transfer learning.

3. **Investigate multi-task learning with sentiment analysis** by implementing a hard parameter sharing architecture that jointly optimises mental health disorder classification and sentiment analysis, evaluating both static and dynamic (uncertainty-based) loss weighting strategies.

4. **Explore hybrid feature-fusion architectures** by combining DeBERTa-v3 contextual representations with static word embeddings (Word2Vec/GloVe) and handcrafted linguistic features to determine whether complementary information sources improve classification beyond what transformers achieve alone.

5. **Evaluate ensemble methods for variance reduction** by training multi-seed models with Multi-Sample Dropout and assessing the stability and performance gains of logit-averaging ensembles.

6. **Establish a comprehensive evaluation protocol** employing multiple complementary metrics—including accuracy, macro and weighted F1, MCC, Cohen's Kappa,

Top-2 accuracy, and AUC-ROC—to provide a robust assessment that accounts for class imbalance and inter-class confusion.

Chapter 3

Methodology and Experimental Setup

3.1 Introduction

This chapter presents the methodological framework underpinning the design, implementation, and evaluation of a multi-task learning system for mental health disorder classification from user-written text. The research follows an iterative experimental methodology, progressing from custom-built attention architectures to pre-trained transformer models, and culminating in hybrid multi-task learning systems that jointly optimise disorder classification and sentiment analysis.

The methodology is structured around four principal axes. First, Section 3.2 establishes the theoretical foundations of attention mechanisms and transformer architectures that constitute the core building blocks of all proposed models. Second, Section 3.3 details the data acquisition, preprocessing, and feature engineering pipeline. Third, Section 3.4 formalises the multi-task learning framework and its theoretical motivation for mental health applications. Fourth, Sections 3.5 through 3.7 describe the three families of architectures investigated—scratch-built attention-based models, fine-tuned transformer backbones, and hybrid feature-fusion systems—along with the training procedures, optimisation strategies, and evaluation protocols employed throughout the experimental study.

The experimental programme encompasses two complementary datasets: a seven-class mental health conditions dataset (hereinafter referred to as the **MA dataset**) containing clinical categories such as Anxiety, Depression, Bipolar Disorder, Stress, Suicidal ideation, Personality Disorder, and Normal; and a six-class **Reddit dataset** collected from mental health-related subreddits covering ADHD, Anxiety, Bipolar Disorder, Depression, PTSD, and a control class. This dual-dataset strategy enables the assessment of model generalisation across distinct textual domains and label taxonomies.

3.2 Theoretical Foundations

3.2.1 Attention Mechanisms

While attention mechanisms were originally introduced by Bahdanau et al. (2014)[56], the scaled dot-product attention formulation employed in transformer architectures was proposed by Vaswani et al. (2017)[21] in the context of neural machine translation, computes a weighted aggregation of value vectors based on the compatibility between a query and a set of keys. For a query vector $\mathbf{q} \in \mathbb{R}^{d_k}$, key matrix $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, and value matrix $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, the scaled dot-product attention is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{3.1}$$

The scaling factor $\sqrt{d_k}$ prevents the dot-product values from becoming excessively large, which would push the softmax function into regions of extremely small gradients—a phenomenon particularly detrimental to training stability.

Multi-Head Attention

Rather than computing a single attention function, the multi-head attention mechanism projects queries, keys, and values h times using distinct learned linear transformations, computes attention in parallel across these subspaces, and concatenates the results:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\,\mathbf{W}^O \tag{3.2}$$

where each head is computed as:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q,\ \mathbf{K}\mathbf{W}_i^K,\ \mathbf{V}\mathbf{W}_i^V) \tag{3.3}$$

with learnable projection matrices
$\mathbf{W}_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$,
$\mathbf{W}_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$,
$\mathbf{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$,
and $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

This decomposition enables the model to attend to information from different representational subspaces at different positions simultaneously.

Attention-Based Pooling

For sequence classification tasks, the variable-length sequence of encoder outputs must be transformed into a fixed-dimensional representation. The attention-based pooling mech-

anism employed in the scratch-built models computes:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t, \tag{3.4}$$

$$\alpha_t = \frac{\exp(\mathbf{w}^\top \mathbf{h}_t)}{\sum_{j=1}^{T} \exp(\mathbf{w}^\top \mathbf{h}_j)} \tag{3.5}$$

where $\mathbf{w} \in \mathbb{R}^{d_{\text{model}}}$ is a learnable attention vector and $\mathbf{h}_t$ denotes the encoder hidden state at position t. The resulting context vector c represents a weighted summary of the entire sequence, where higher weights are assigned to tokens deemed more relevant to the classification task..

3.2.2 Transformer Architecture

The Transformer architecture, introduced by Vaswani et al. (2017)[21], eliminates recurrent computations entirely and relies exclusively on self-attention mechanisms to model dependencies between tokens, irrespective of their relative distance within the input sequence.. Each encoder layer consists of two sub-layers:

1. **Multi-Head Self-Attention**: The input sequence attends to itself, allowing each token representation to incorporate contextual information from every other position in the sequence.

2. **Position-wise Feed-Forward Network (FFN)**: A position-wise fully connected network consisting of two linear transformations separated by a non-linear activation function, applied independently to each token representation:

$$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\,\mathbf{W}_2 + \mathbf{b}_2 \tag{3.6}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are learnable parameters, corresponds to the ReLU activation function, which introduces non-linearity into the transformation.

Both sub-layers are followed by residual connections and layer normalisation, yielding the following output transformation:

$$\text{LayerNorm}(\mathbf{x} + \text{SubLayer}(\mathbf{x})) \tag{3.7}$$

Positional Encoding

Since the self-attention mechanism is inherently permutation-invariant, positional encodings are added to the input embeddings to provide explicit information about the order of

tokens within the sequence. The positional encoding proposed by Vaswani et al. (2017)[21] is defined as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \qquad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{3.8}$$

This formulation enables the model to extrapolate to sequence lengths not encountered during training, as the relative position between any two tokens can be expressed as a linear function of their positional encodings.

3.2.3 Pre-Trained Language Models

The present study employs three pre-trained transformer backbones, each representing a distinct trade-off between model capacity and computational efficiency (Table 3.1).

Table 3.1: Comparison of pre-trained transformer backbones.

Property	**DistilBERT**	**DeBERTa-base**	**DeBERTa-v3-base**
Parameters	66 M	139 M	184 M
Layers	6	12	12
Hidden Size	768	768	768
Attention Heads	12	12	12
Pre-training Obj.	Distilled MLM	MLM + RTD	RTD (ELECTRA)
Positional Enc.	Absolute	Disentangled	Disentangled
Key Innovation	Knowledge distill.	Disentangled attn.	Replaced Token Det.

DistilBERT (Sanh et al., 2019) is a compressed variant of BERT obtained through knowledge distillation, retaining approximately 97% of BERT's performance on benchmark language understanding tasks while reducing model size and inference time substantially.

DeBERTa (He et al., 2021) introduces *disentangled attention*, in which each token representation is decomposed into separate content and position embeddings.. The attention score between tokens i and j is computed as:

$$A_{i,j} = \{\mathbf{H}_i^c, \mathbf{P}_{i|j}^p\} \times \{\mathbf{H}_j^c, \mathbf{P}_{j|i}^p\}^\top \tag{3.9}$$

where $\mathbf{H}^c$ and $\mathbf{P}^p$ denote content and relative position embeddings respectively. By modelling content and positional information independently, the architecture captures contextual relationships more effectively than conventional attention mechanisms that rely on a single combined representation.

DeBERTa-v3 extends the architecture with *Replaced Token Detection* (RTD) pre-training, in which a lightweight generator produces plausible token substitutions and a discriminator is trained to distinguish original tokens from replaced ones—a more sample-efficient objective than masked language modelling.

As shown in Table 3.1, DistilBERT prioritises computational efficiency, whereas DeBERTa and DeBERTa-v3 provide greater representational capacity through more sophisticated attention mechanisms and pre-training objectives. This diversity enables a systematic evaluation of the trade-off between efficiency and predictive performance in mental health text classification.

3.3 Data Pipeline and Preprocessing

3.3.1 Dataset Description

Two complementary datasets are employed throughout the experimental study to evaluate the proposed models across different domains and label taxonomies (Table 3.2).

Table 3.2: Summary of datasets used in the experimental programme.

Property	**MA Dataset**	**Reddit Dataset**
Source	Curated clinical texts	Reddit subreddits
Classes	7	6
Labels	Anxiety, Bipolar, Depression, Normal, Pers. Disorder, Stress, Suicidal	ADHD, Anxiety, Bipolar, Depression, PTSD, None
Training Samples	∼67,894	∼13,727
Validation Samples	∼16,974	∼1,488
Test Samples	∼21,218	∼1,488
Text Register	Semi-formal clinical	Informal social media
Split Strategy	Pre-defined JSON	Pre-defined CSV
Auxiliary Labels	Sentiment (derived)	Sentiment (derived)

The MA dataset is provided as a JSON file containing predefined training and test partitions. To obtain a validation set for model development and hyperparameter tuning, a stratified 90/10 split was applied to the training partition while preserving the original class distribution. The Reddit dataset is distributed as CSV files with separate partitions for three textual modalities: post body, title, and combined (title + post). No additional

splitting was required, allowing all experiments to be conducted using the original dataset configuration.

3.3.2 Text Preprocessing Pipeline

All textual inputs undergo a standardised preprocessing pipeline before being provided to the learning models (Figure 3.1).

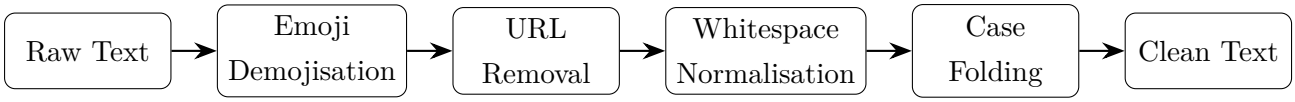

Figure 3.1: Text preprocessing pipeline.

The preprocessing pipeline is designed to reduce textual noise while preserving information that may be relevant for mental health classification.

The preprocessing operations are summarised in Table 3.3

Table 3.3: Preprocessing operations and their rationale.

Step	Operation	Rationale
1	Emoji demojisation	Converts pictographic symbols to interpretable textual tokens (e.g., `:cry:`)
2	URL removal	Eliminates non-semantic web addresses
3	Whitespace normalisation	Collapses multiple spaces, tabs, and newlines
4	Lowercasing	Reduces vocabulary size and standardises surface forms

For transformer-based architectures, the cleaned text is subsequently tokenised using the model-specific `AutoTokenizer` provided by the Hugging Face `Transformers` library. Each model employs a different tokenisation strategy and vocabulary, as summarised in Table 3.4.

Table 3.4: Tokenisation strategies across model families.

Model	Tokeniser	Vocab Size	Algorithm	Max Len
Scratch Models	GloVe vocabulary	∼400K	Whitespace	256–512
DistilBERT	`distilbert-base-uncased`	30,522	WordPiece	256–512
DeBERTa-base	`microsoft/deberta-base`	50,265	BPE (GPT-2)	512
DeBERTa-v3-base	`microsoft/deberta-v3-base`	128,100	SentencePiece	256–512

3.3.3 Linguistic Feature Extraction

The scratch-built and hybrid architectures augment raw token representations with linguistically-informed features extracted using the spaCy `en_core_web_sm` language model. These features provide complementary syntactic and lexical information that may not be fully captured by distributional word embeddings alone. Three categories of linguistic features are computed.

Part-of-Speech (POS) Tag Distribution

For each input text, the normalised frequency of eight major part-of-speech POS categories is computed: `NOUN`, `VERB`, `ADJ`, `ADV`, `PRON`, `DET`, `ADP`, and `CONJ`. These distributions characterise the syntactic structure of the text and may reveal stylistic patterns associated with specific mental health conditions. For example, prior research has reported that increased use of first-person pronouns is associated with depressive symptomatology (Rude et al., 2004).

Sentiment-Aware Features

Sentiment information is represented using one-hot encoded sentiment labels and sentiment-weighted linguistic features derived from Part-of-Speech (POS) statistics. For example, adjectives are considered as a representative linguistic category because they frequently convey subjective opinions and emotional polarity. The sentiment contribution of adjectives is computed as follows:

$$f_{\text{adj}}^{\text{sent}} = \frac{\text{count(ADJ)}}{|\text{tokens}|} \times w_s, \qquad w_s \in \{1.0,\ 0.5,\ -1.0\} \tag{3.10}$$

where w_s corresponds to positive, neutral, and negative sentiment respectively. This formulation illustrates how sentiment polarity can be incorporated into POS-based linguistic features, while similar statistics can be extracted for other sentiment-relevant categories.

Clinical Discriminative Features (Hybrid Model)

The hybrid architecture introduces an extended feature set of 31 dimensions specifically designed to discriminate between clinically similar categories (Table 3.5).

Table 3.5: Composition of the 31-dimensional handcrafted feature vector used in the hybrid model.

Feature Group	Dims	Description
POS tag frequencies	8	Normalised counts of major POS categories
Sentiment-weighted POS	2	Adjective and adverb rates modulated by sentiment
Negation density	1	Frequency of negation words
First-person pronoun rate	1	Frequency of self-referential pronouns
Text statistics	3	Character count, word count, average word length
Sentiment one-hot	3	Encoded sentiment label
Clinical keyword features	6	Anxiety/depression keyword frequency, differential score, density, confidence
Temporal pattern features	4	Acute vs. chronic temporal marker frequencies
Linguistic style features	3	Question rate, exclamation rate, avg. sentence length
Total	**31**	

The clinical keyword features are motivated by findings from psycholinguistic and clinical language research suggesting that anxiety-related discourse often contains episodic and acute temporal expressions (e.g., *"suddenly"*, *"right now"*), whereas depression-related discourse is more frequently characterised by absolutist and chronic expressions (e.g., *"always"*, *"never"*).

3.3.4 Word Embedding Features

The hybrid architecture further incorporates static word embedding representations to complement contextual transformer features. For each input document, a fixed-dimensional document representation is obtained by averaging the embeddings of all in-vocabulary words:

$$\mathbf{e}_{\text{text}} = \frac{1}{|V_{\text{text}}|} \sum_{w \in V_{\text{text}}} \mathbf{e}_w \tag{3.11}$$

where V_{text} denotes the set of words present in both the document and the embedding vocabulary, and $\mathbf{e}_w \in \mathbb{R}^{300}$ represents the embedding vector associated with word $_w$. Two

pre-trained embedding resources are considered: Word2Vec trained on the Google News corpus (approximately 3 million vocabulary entries) and GloVe trained on the Wikipedia–Gigaword corpus (approximately 400,000 vocabulary entries). Both resources provide 300-dimensional word representations.

3.3.5 Data Augmentation

For the DeBERTa-based multi-task experiments conducted on the Reddit dataset, synonym-based data augmentation is applied exclusively to the training partition. Each training instance undergoes stochastic word substitution using synonyms extracted from WordNet. The replacement probability is carefully controlled to preserve the original semantic content of the text while introducing lexical variability and reducing the risk of overfitting.

3.3.6 Class Imbalance Handling

Both datasets exhibit non-uniform class distributions. Two complementary strategies are employed.

Balanced Class Weights: Inverse-frequency weighting:

$$w_c = \frac{N}{K \cdot n_c} \tag{3.12}$$

where N is the total number of samples, K the number of classes, and n_c the number of samples in class c.

Focal Loss (Lin et al., 2017): For DistilBERT fine-tuning (Experiment 15):

$$\mathcal{L}_{\text{FL}} = -\alpha \, (1 - p_t)^{\gamma} \log(p_t) \tag{3.13}$$

where p_t is the predicted probability of the true class, $\alpha = 1.0$, and $\gamma = 2.0$. When $\gamma > 0$, the loss contribution from easy examples ($p_t \to 1$) is exponentially suppressed.

3.4 Multi-Task Learning Framework

3.4.1 Theoretical Motivation

Multi-task learning (MTL) (Caruana, 1997) is a learning paradigm in which a model is trained simultaneously on multiple related tasks, sharing representations across them. The fundamental hypothesis is that tasks which share underlying structure can benefit from joint training through *inductive transfer*: the auxiliary task acts as a regulariser, biasing the shared representation toward features that generalise across tasks.

In the context of mental health classification, the primary task (disorder category prediction) and the auxiliary task (sentiment analysis) are hypothesised to share latent representations for several reasons:

1. **Affective overlap**: Mental health conditions are inherently characterised by affective states. Sentiment analysis captures precisely these affective dimensions.
2. **Linguistic markers**: Both tasks rely on similar lexical and syntactic cues—negation patterns, emotional vocabulary, self-referential pronoun usage.
3. **Regularisation**: The sentiment task provides additional gradient signal, encouraging the shared encoder to learn robust, generalisable representations.

3.4.2 Hard Parameter Sharing

The MTL architecture adopts the *hard parameter sharing* paradigm, wherein the transformer backbone and an intermediate shared layer are common to both tasks, with task-specific heads branching from the shared representation (Figure 3.2).

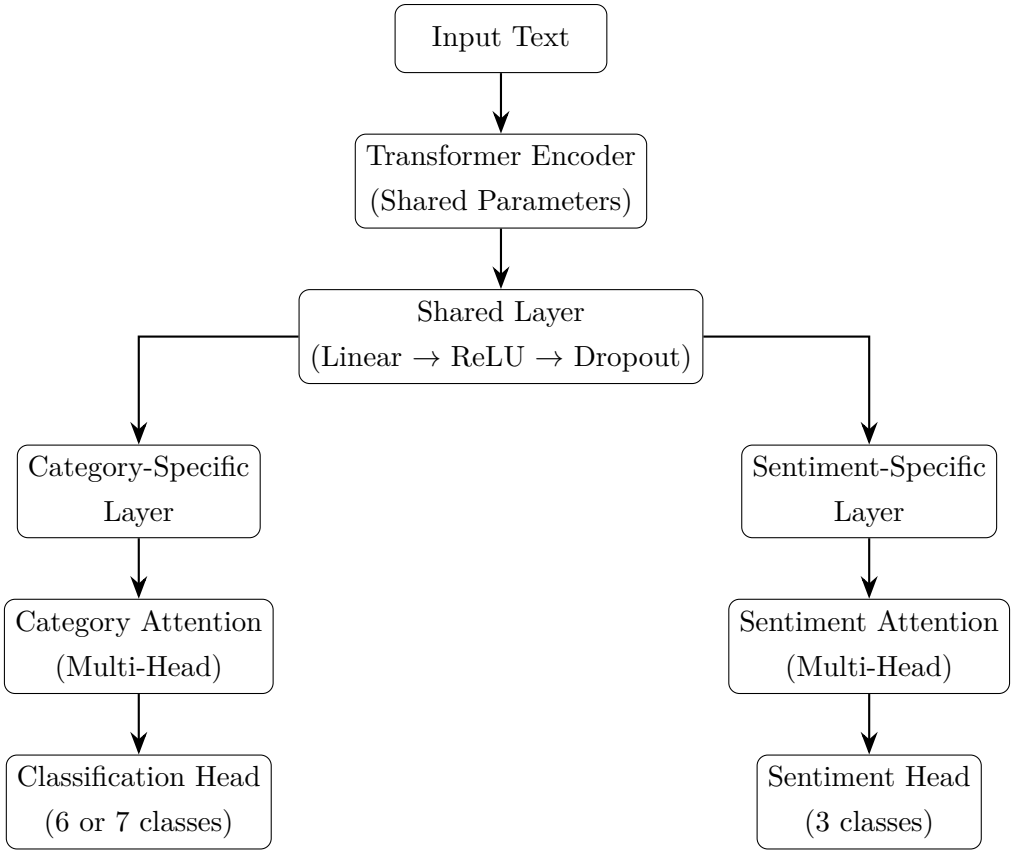

Figure 3.2: Hard parameter sharing architecture for multi-task learning.

3.4.3 Multi-Task Loss Formulation

Two weighting strategies are employed.

Static Weighting (DeBERTa MTL):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{sent}}, \qquad \lambda = 0.5 \tag{3.14}$$

Dynamic Uncertainty Weighting (DistilBERT MTL), following Kendall et al. (2018):

$$\mathcal{L}_{\text{total}} = \frac{1}{2\sigma_1^2}\mathcal{L}_{\text{cls}} + \frac{1}{2\sigma_2^2}\mathcal{L}_{\text{sent}} + \log \sigma_1 + \log \sigma_2 \tag{3.15}$$

In practice, the log-variance $s_i = \log \sigma_i^2$ is optimised directly. The precision terms $\frac{1}{2}\exp(-s_i)$ adaptively scale each task's contribution: a task with higher aleatoric uncertainty receives a lower effective weight.

3.4.4 Sentiment Label Derivation

For the Reddit dataset, ground-truth sentiment labels are not available. Sentiment labels are therefore derived using a lexicon-based sentiment analyser applied to the raw text, producing three classes: *positive*, *neutral*, and *negative*. While these derived labels are inherently noisy, the MTL framework is designed to be robust to moderate label noise in the auxiliary task.

3.5 Model Architectures

This section details the three families of architectures investigated, progressing from custom-built models to pre-trained transformer fine-tuning to hybrid feature-fusion systems.

3.5.1 Family I: Scratch-Built Attention Models

The first architectural family is a custom transformer-like encoder built from scratch, designed to validate the effectiveness of self-attention mechanisms combined with linguistically-informed features before introducing pre-trained backbones (Figure 3.3).

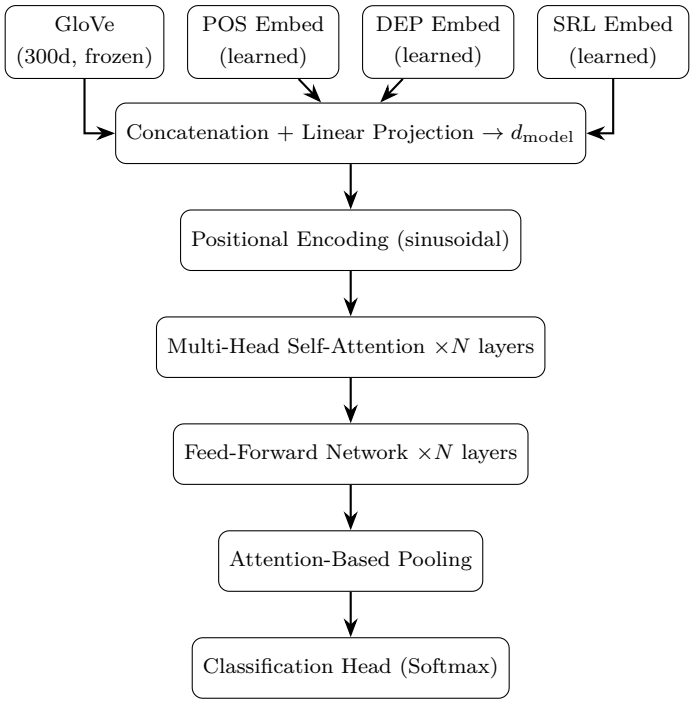

Figure 3.3: Architecture of the scratch-built attention model (Family I).

Each token is represented as the concatenation of four embedding vectors:

$$\mathbf{x}_t = [\mathbf{e}_t^{\text{GloVe}};\ \mathbf{e}_t^{\text{POS}};\ \mathbf{e}_t^{\text{DEP}};\ \mathbf{e}_t^{\text{SRL}}] \tag{3.16}$$

projected to d_{model} via $\mathbf{h}_t^{(0)} = \mathbf{W}_{\text{proj}}\mathbf{x}_t + \mathbf{b}_{\text{proj}}$.

The encoder stack is parameterised according to the configurations in Table 3.6.

Table 3.6: Scratch-built encoder configurations across experiments.

Parameter	**Exp. 1/2**	**Exp. 7**	**Exp. 10**
d_{model}	256	512	512
Attention Heads (h)	4	8	8
Encoder Layers (N)	3	4	6
FFN Hidden Dim	512	1024	1024
Max Sequence Length	256	256	512
Dropout	0.1	0.15	0.15
Dataset	Reddit (6-class)	MA (7-class)	Reddit (6-class)

The encoder output is reduced via attention-based pooling (Equation 3.5), followed by $\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c \cdot \text{AttPool}(\mathbf{H}) + \mathbf{b}_c)$.

3.5.2 Family II: Fine-Tuned Transformer Models

The second architectural family leverages pre-trained transformer backbones through fine-tuning.

DistilBERT Single-Task (Experiments 8, 14, 15)

The single-task configuration uses `AutoModelForSequenceClassification`, appending a linear head to the `[CLS]` representation:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c \cdot \text{Dropout}(\mathbf{h}_{[\text{CLS}]}) + \mathbf{b}_c) \tag{3.17}$$

Experiment 15 introduces increased dropout (0.2) and Focal Loss with class weights.

DeBERTa Single-Task with Monkey-Patch (Experiments 12, 13)

Fine-tuning DeBERTa with FP16 introduces a numerical stability challenge: the disentangled attention masking constant (`torch.finfo(dtype).min` $= -65504$ for FP16) causes softmax overflow. A runtime monkey-patch replaces this with a safe value:

```
class SafeFInfo:
    """Prevents␣FP16␣overflow␣in␣DeBERTa␣attention."""
    min = -1e4  # Safe floor (vs. -65504 for FP16)
```

Listing 3.1: Safe attention masking for DeBERTa FP16 training.

Layer Freezing Strategy (Experiment 13)

Experiment 13 investigates selective layer freezing (Table 3.7), reducing trainable parameters by approximately 50%.

Table 3.7: Layer freezing strategy for DeBERTa fine-tuning.

Component	Trainable	Rationale
Token Embeddings	✗	Low-level representations transfer well
Encoder Layers 0–5	✗	Generic linguistic features
Encoder Layers 6–11	✓	Task-specific semantic adaptation
Pooler + Class. Head	✓	Task-specific output mapping

DistilBERT Multi-Task

The DistilBERT MTL model extends the single-task architecture with hard parameter sharing (Figure 3.2). Task-specific self-attention modules (8-head MHA) refine each

branch. Layer-wise learning rates are employed (Table 3.8).

Table 3.8: Layer-wise learning rate schedule for DistilBERT MTL.

Component	LR Multiplier
Transformer backbone	$\times 1$ (base LR)
Shared layer	$\times 5$
Task-specific layers	$\times 10$
Attention modules	$\times 10$
Classification heads	$\times 10$

DeBERTa Multi-Task

The DeBERTa MTL model uses a streamlined dual-head architecture with mean pooling:

$$\mathbf{h} = \text{MeanPool}(\text{DeBERTa}(\mathbf{x})) \tag{3.18}$$

$$\hat{\mathbf{y}}_{\text{cls}} = \mathbf{W}_{\text{cls}} \cdot \text{Dropout}_{0.5}(\mathbf{h}) \tag{3.19}$$

$$\hat{\mathbf{y}}_{\text{sent}} = \mathbf{W}_{\text{sent}} \cdot \text{Dropout}_{0.5}(\mathbf{h}) \tag{3.20}$$

Key choices include: mean pooling (vs. `[CLS]`), high dropout (0.5), autocast disabled for the transformer pass, static loss weighting ($\lambda = 0.5$), and gradient accumulation (4 steps, effective batch size = 32).

3.5.3 Family III: Hybrid Feature-Fusion Model

The hybrid architecture (Experiment 16) fuses three complementary feature streams (Figure 3.4).

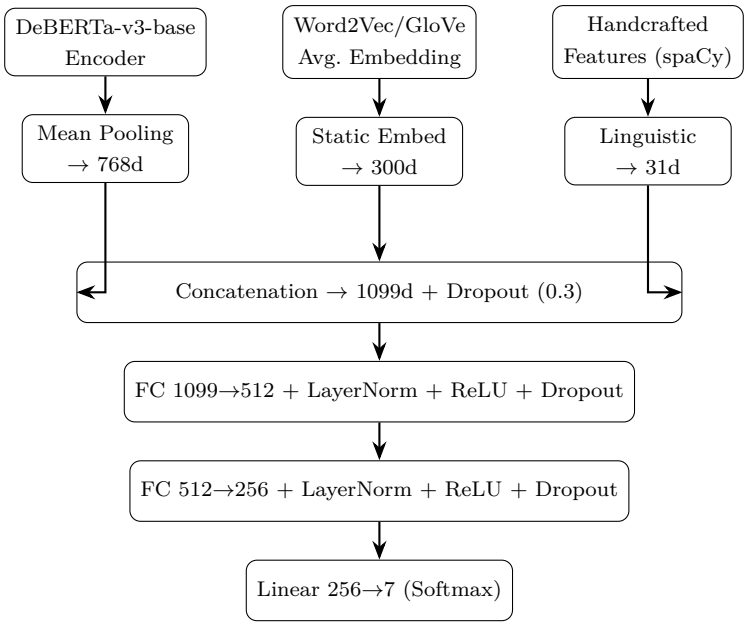

Figure 3.4: Architecture of the hybrid feature-fusion model (Family III).

An optional feature attention module computes importance weights $\boldsymbol{\alpha} = \text{softmax}(\mathbf{W}_a[\mathbf{h}_{\text{trans}}; \mathbf{h}_{\text{embed}}; \mathbf{h}_$ $\mathbb{R}^3$, reserved for post-hoc interpretability. The DeBERTa-v3 backbone freezes embeddings and encoder layers 0–5.

3.5.4 Multi-Seed Ensemble (Experiment 17)

A DeBERTa-v3-base model with **Multi-Sample Dropout** (MSD) is trained with three seeds (42, 1337, 2024). Predictions are aggregated via logit averaging:

$$\hat{\mathbf{y}}_{\text{ensemble}} = \arg\max \frac{1}{S} \sum_{s=1}^{S} \mathbf{z}^{(s)} \tag{3.21}$$

MSD applies $K = 5$ independent dropout masks (rate = 0.3) to the `[CLS]` representation:

$$\mathbf{z} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{W}_c \cdot \text{Dropout}_k(\mathbf{h}_{[\text{CLS}]}) \tag{3.22}$$

This provides two levels of variance reduction: implicit (MSD within each forward pass) and explicit (multi-seed ensemble at inference). The model uses label smoothing ($\epsilon = 0.1$) and balanced class weights.

3.6 Training Procedures and Optimisation

3.6.1 Optimiser Configuration

All experiments employ AdamW (Loshchilov & Hutter, 2019):

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right) \tag{3.23}$$

3.6.2 Learning Rate Scheduling

A linear warmup-decay schedule is employed:

$$\eta(t) = \begin{cases} \eta_{\max} \cdot \frac{t}{T_{\text{warmup}}} & \text{if } t \leq T_{\text{warmup}} \\ \eta_{\max} \cdot \frac{T_{\text{total}} - t}{T_{\text{total}} - T_{\text{warmup}}} & \text{if } t > T_{\text{warmup}} \end{cases} \tag{3.24}$$

Warmup ratios range from 10% to 15% of total training steps.

3.6.3 Mixed-Precision Training

All GPU-accelerated experiments employ FP16 via `torch.amp`. For DeBERTa models, autocast is selectively disabled during the transformer forward pass while remaining enabled for classification heads.

3.6.4 Gradient Clipping

Gradient norm clipping with max_norm = 1.0 is applied universally:

$$\mathbf{g} \leftarrow \frac{\mathbf{g}}{\max(1,\ \|\mathbf{g}\|_2 / \text{max\_norm})} \tag{3.25}$$

3.6.5 Regularisation Techniques

Multiple complementary strategies are employed (Table 3.9).

Table 3.9: Regularisation techniques employed across experiments.

Technique	Experiments	Details
Dropout	All	Rates: 0.1–0.5
Weight Decay	All transformer	$\lambda = 0.01$ or 0.1
Label Smoothing	Exp. 16, 17	$\epsilon = 0.1$
Early Stopping	Exp. 15, DeBERTa MTL	Patience: 3–4 epochs
Multi-Sample Dropout	Exp. 17	$K = 5$ independent masks
Layer Freezing	Exp. 13, 16, DeBERTa MTL	First 6 layers frozen

Label smoothing replaces the hard one-hot target:

$$\mathbf{y}_{\text{smooth}} = (1 - \epsilon) \cdot \mathbf{y} + \frac{\epsilon}{K} \tag{3.26}$$

3.6.6 Comprehensive Hyperparameter Summary

Table 3.10 consolidates the complete configuration.

Table 3.10: Comprehensive hyperparameter summary. CE = Cross-Entropy, CW = Class Weights, FL = Focal Loss, LS = Label Smoothing.

	Scratch	**DistilBERT** (ST)	**DeBERTa** (ST)	**DistilBERT** MTL	**DeBERTa** MTL	**Hybrid** (16)	**Ensemble** (17)
LR	1e-4	1e-5–3e-5	2e-5	2e-5	2e-5	1e-5	2e-5
Batch Size	32	32–64	16	16	8	16–32	16
Epochs	50	5–18	8–10	3–5	6–10	15	10
Max Seq Len	256	256–512	512	256	512	256–512	256
Dropout	0.1	0.1–0.2	0.1	0.2	0.5	0.3	0.3
Loss	CE+CW	CE+CW/FL	CE+CW	CE+CW	CE+CW	CE+CW+LS	CE+CW+LS
FP16	✗	✓	✓	✓	✓	✓	✓

3.6.7 Hardware and Computational Environment

All experiments are conducted on Google Colab instances equipped with NVIDIA T4 GPUs (16 GB VRAM). The 5-hour session limit and single-GPU constraint dictate batch size selection and motivate the use of gradient accumulation and early stopping.

3.7 Evaluation Protocol

3.7.1 Evaluation Metrics

Model performance is assessed using a comprehensive suite of metrics (Table 3.11).

Table 3.11: Evaluation metrics employed across all experiments.

Metric	Interpretation
Accuracy	Overall correctness
Precision (weighted)	Class-size-weighted positive predictive value
Recall (weighted)	Class-size-weighted sensitivity
F1 (weighted)	Harmonic mean of weighted precision and recall
F1 (macro)	Equal weight to all classes; sensitive to minority performance
MCC	Balanced measure robust to class imbalance
Cohen's Kappa (κ)	Agreement beyond chance
Top-2 Accuracy	Leniency for clinically adjacent categories
AUC-ROC (OvR)	Ranking quality across decision thresholds

The selection of MCC as a primary metric is motivated by its robustness to class imbalance—it produces a high score only if the classifier performs well on both majority and minority classes. Top-2 accuracy is particularly relevant where clinical categories overlap (e.g., Anxiety and PTSD).

3.7.2 Evaluation Procedure

Each experiment follows a standardised three-phase protocol:

1. **Per-epoch validation**: After each training epoch, the model is evaluated on the validation set with full metric computation.

2. **Model selection**: The best checkpoint is selected based on validation accuracy or loss. Early stopping terminates training after 3–4 epochs without improvement.

3. **Final test evaluation**: The selected model is evaluated on the held-out test set. A comprehensive report is generated including overall metrics, per-class metrics, confusion matrices, ROC and PR curves, and misclassified example analysis.

3.7.3 Visualisation and Reporting

Each experiment produces a standardised set of visual outputs: confusion matrices (raw and normalised), per-class metrics bar charts, ROC curves (one-vs-rest with per-class AUC), precision-recall curves, and training loss/accuracy curves. These enable qualitative analysis of systematic confusion between clinically similar categories.

Chapter 4

Results and Discussion

4.1 Introduction

This chapter presents the experimental results obtained from the multi-task learning framework for mental health disorder classification described in Chapter 3. The primary objective of this experimental campaign is to evaluate and compare the classification performance of three architectural families—scratch-built attention models, fine-tuned pre-trained transformers, and hybrid feature-fusion systems—across two distinct mental health datasets. The analysis further investigates the impact of multi-task learning with sentiment analysis as an auxiliary task, the effectiveness of advanced training techniques (Focal Loss, label smoothing, multi-sample dropout), and the robustness of ensemble methods for variance reduction.

The evaluation follows a rigorous protocol employing multiple complementary metrics—accuracy, macro and weighted F1-scores, Matthews Correlation Coefficient (MCC), and Cohen's Kappa—to provide a comprehensive assessment that accounts for class imbalance and inter-class confusion. All reported results correspond to the held-out test set unless otherwise specified, ensuring that the analysis reflects true generalisation performance rather than in-sample fitting.

The chapter is organised as follows. Section 4.2 summarises the experimental configurations. Section 4.3 reviews the evaluation metrics and their relevance to mental health classification. Section 4.4 presents the detailed experimental results across all model families. Section 4.5 provides a comparative analysis with state-of-the-art approaches. Section 4.6 offers a comprehensive discussion of the findings, and Section 4.7 identifies limitations and directions for future work.

4.2 Experimental Configurations

4.2.1 Experiment Taxonomy

The experimental programme encompasses 17 distinct experiments organised into five families, each targeting a specific research question (Table 4.1).

Table 4.1: Experiment taxonomy and research objectives.

Family	Exp.	Architecture	Dataset	Research Question
Baseline	1, 2, 7, 10	Custom Transformer + GloVe	Reddit / MA	Can scratch-built attention models capture mental health linguistic patterns?
Transformer ST	8, 14, 15	DistilBERT (single-task)	MA 7-class	How do pre-trained backbones compare?
Transformer ST	12, 13	DeBERTa (patched / frozen)	MA 7-class	Does disentangled attention improve over DistilBERT?
MTL	src_distilbert, src_deberta	DistilBERT / DeBERTa MTL	Reddit 6-class	Does joint sentiment classification help?
Hybrid / Ens.	16, 17	DeBERTa-v3 + W2V; Multi-seed MSD	MA 7-class	Does fusion or ensembling yield further gains?

Experiments on the Reddit dataset (6-class, ∼1,488 test samples) and the MA dataset (7-class, ∼21,218 test samples) are not directly comparable due to differing class taxonomies, dataset sizes, and textual registers.

4.2.2 Hyperparameter Configurations

Table 4.2 consolidates the hyperparameter settings for all experiments. These configurations were determined through iterative empirical tuning within the computational constraints of the Google Colab environment (NVIDIA T4, 16 GB VRAM).

Table 4.2: Comprehensive hyperparameter configurations. ST = Single-Task, MTL = Multi-Task Learning, CE = Cross-Entropy, CW = Class Weights, FL = Focal Loss, LS = Label Smoothing.

Exp.	Backbone	Ep.	Batch	LR	Drop.	Loss	Task
1	Scratch (3L/4H/256d)	50	32	1e-4	0.1	CE+CW	ST
2	Scratch (3L/4H/256d)	50	32	1e-4	0.1	CE+CW	ST
7	Scratch (4L/8H/512d)	50	32	1e-4	0.15	CE+CW	ST
10	Scratch (6L/8H/512d)	50	32	1e-4	0.15	CE+CW	ST
8	DistilBERT	18	64	1e-5	0.1	CE+CW	ST
14	DistilBERT	10	32	2e-5	0.1	CE+CW	ST
15	DistilBERT + FL	10	32	3e-5	0.2	FL+CW	ST
12	DeBERTa (patched)	10	16	3e-5	0.1	CE+CW	ST
13	DeBERTa (frozen)	8	16	3e-5	0.1	CE+CW	ST
16	DeBERTa-v3+W2V+Feat	15	32	1e-5	0.3	CE+CW+LS	ST
17	DeBERTa-v3+MSD	10	16	2e-5	0.3	CE+CW+LS	ST
DistilBERT MTL (5ep/16)		5	16	2e-5	0.2	CE+CW (dyn.)	MTL
DistilBERT MTL (5ep/32)		5	32	3e-5	0.2	CE+CW (dyn.)	MTL
DistilBERT MTL (8ep/32)		8	32	3e-5	0.2	CE+CW (dyn.)	MTL
DeBERTa MTL (8ep/16)		8	16	3e-5	0.2	CE+CW (dyn.)	MTL
DeBERTa MTL (6ep/8)		6	8	2e-5	0.5	CE+CW (stat.)	MTL

4.3 Evaluation Metrics

4.3.1 Metric Definitions

The evaluation protocol employs multiple complementary metrics to provide a comprehensive assessment of classifier performance.

Accuracy measures the proportion of correct predictions:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i] \tag{4.1}$$

While intuitive, accuracy can be misleading for imbalanced datasets where a majority-class classifier achieves deceptively high scores.

Precision, Recall, and F1-Score are computed per-class and then aggregated:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2\,P_c\,R_c}{P_c + R_c} \tag{4.2}$$

Macro-averaged F1 assigns equal weight to each class, making it sensitive to minority-class performance:

$$F1_{\text{macro}} = \frac{1}{K}\sum_{c=1}^{K} F1_c \tag{4.3}$$

Weighted-averaged F1 weights each class by its support, reflecting overall classification quality proportional to the class distribution.

Matthews Correlation Coefficient (MCC) provides a balanced measure even when classes are of very different sizes:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.4}$$

MCC ranges from -1 (total disagreement) through 0 (random prediction) to $+1$ (perfect prediction). It is considered the most informative single metric for multi-class classification on imbalanced data (Chicco & Jurman, 2020).

Cohen's Kappa (κ) measures inter-rater agreement beyond chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.5}$$

where p_o is the observed agreement and p_e the expected agreement by chance.

4.3.2 Metric Relevance for Mental Health Classification

In mental health classification, the cost of misclassification is asymmetric: failing to detect suicidal ideation (false negative) carries substantially higher real-world consequences than a false positive. Additionally, the MA dataset exhibits considerable class imbalance—the Normal and Depression classes each contain over 6,000 test samples, whereas Personality Disorder has only 447. These characteristics make macro F1 and MCC particularly important: they penalise models that achieve high accuracy by excelling on majority classes while neglecting minority categories. Throughout this chapter, MCC is treated as the primary discriminating metric when comparing models of similar accuracy.

4.4 Experimental Results

4.4.1 Family I: Scratch-Built Attention Models

The scratch-built models establish a lower-bound baseline, quantifying the performance achievable through custom attention architectures with GloVe embeddings and linguistically-informed features without pre-trained contextual representations.

Table 4.3: Scratch-built model results on the Reddit dataset (6-class, 1,488 test samples).

Metric	**Exp. 1 (3L/4H/256d)**	**Exp. 10 (6L/8H/512d)**	**Δ**
Accuracy	0.7419	0.7480	+0.61 pp
F1 (macro)	0.7407	0.7455	+0.48 pp
F1 (weighted)	0.7407	0.7455	+0.48 pp
MCC	0.6910	0.6984	+0.74 pp
Loss	2.7729	3.4469	+0.674

Table 4.4: Scratch-built model results on the MA dataset (7-class, 21,218 test samples).

Metric	**Exp. 2 (3L/4H/256d)**	**Exp. 7 (4L/8H/512d)**	**Δ**
Accuracy	0.8817	0.8886	+0.70 pp
F1 (macro)	0.8636	0.8759	+1.23 pp
F1 (weighted)	0.8821	0.8890	+0.69 pp
MCC	0.8477	0.8557	+0.80 pp
Loss	1.0439	0.7756	−0.268

Key observations. Scaling model capacity from Experiment 1 (3 layers, 4 heads, 256d) to Experiment 10 (6 layers, 8 heads, 512d) yields only marginal gains (+0.61 pp accuracy on Reddit), suggesting a performance ceiling of approximately 74–75% for scratch-built models on the Reddit task. The Depression class is consistently the most challenging (F1 $\approx$ 0.62–0.66 on Reddit), while the None/Normal class achieves F1 $>$ 0.86 due to its distinct lexical profile. The MA dataset yields substantially higher performance (88.2–88.9%) compared to Reddit (74.2–74.8%), attributable to the 5$\times$ larger training set and potentially more distinctive lexical boundaries.

4.4.2 Family II: Single-Task Fine-Tuned Transformers (MA Dataset)

Fine-tuning pre-trained transformers represents a significant methodological step. This section compares DistilBERT and DeBERTa variants on the MA 7-class dataset.

Table 4.5: Single-task transformer results on the MA dataset (7-class, 21,218 test samples).

Metric	**Exp. 8** (DistilBERT)	**Exp. 14** (DistilBERT)	**Exp. 15** (DistilBERT+FL)	**Exp. 12** (DeBERTa)	**Exp. 13** (DeBERTa frz.)
Accuracy	0.9357	0.9378	0.8429	**0.9406**	0.9385
F1 (macro)	0.9220	0.9272	0.8473	**0.9309**	0.9291
F1 (weighted)	0.9353	0.9376	0.8434	**0.9403**	0.9383
MCC	0.9160	0.9188	0.8045	**0.9224**	0.9198
Cohen's κ	0.9160	0.9188	0.7984	**0.9224**	0.9197

Key findings. (1) DeBERTa-base (Exp. 12) achieves the highest accuracy (94.06%), F1 macro (93.09%), and MCC (92.24%), attributable to its disentangled attention mechanism and larger parameter count (139M vs. 66M). (2) Layer freezing (Exp. 13) preserves 99.8% of full fine-tuning performance (93.85% vs. 94.06%) while reducing trainable parameters by ∼50%. (3) Focal Loss (Exp. 15) degrades accuracy to 84.29%: Depression recall drops to 65.48% while Personality Disorder recall rises to 95.30%, indicating that the combination of Focal Loss ($\gamma = 2.0$) with balanced class weights creates an excessively aggressive minority-class bias. (4) DistilBERT achieves 93.57–93.78% with only 47% of DeBERTa's parameters, demonstrating excellent efficiency.

Table 4.6: Per-class F1 scores for single-task transformers on the MA dataset.

Class	**Exp. 8 (DistilBERT)**	**Exp. 12 (DeBERTa)**	**Exp. 15 (DistilBERT+FL)**
Anxiety	0.9476	**0.9629**	0.9273
Bipolar	0.9349	**0.9438**	0.9296
Depression	0.9266	**0.9302**	0.7583
Normal	**0.9711**	0.9728	0.9519
Personality Disorder	0.8747	**0.8945**	0.7795
Stress	0.8940	**0.9022**	0.8284
Suicidal	0.9049	**0.9099**	0.7561

4.4.3 Family III: Multi-Task Learning Models (Reddit Dataset)

The multi-task learning experiments evaluate whether joint training with sentiment analysis as an auxiliary task improves mental health disorder classification on the Reddit 6-class dataset.

Table 4.7: Multi-task learning results on the Reddit dataset (6-class, 1,488 test samples).

Model	Cat. Acc.	Cat. $F1_m$	Sent. Acc.	Sent. F1	Top-2	AUC
DistilBERT MTL (5ep/16)	0.8273	0.8288	0.7184	0.7175	—	—
DistilBERT MTL (5ep/32)	0.8542	0.8543	0.7782	0.7783	—	—
DistilBERT MTL (8ep/32)	0.8427	0.8431	0.7843	0.7844	—	—
DeBERTa MTL (8ep/16)	**0.8824**	**0.8828**	**0.7675**	**0.7678**	—	—
DeBERTa MTL (6ep/8)	0.8723	0.8719	—	—	**0.9570**	**0.9835**

Key findings. (1) The DeBERTa-base MTL model achieves 88.24% category accuracy—surpassing the best scratch model by 13.4 pp and the best DistilBERT MTL by 2.8 pp. (2) Increasing batch size from 16 to 32 improved DistilBERT MTL accuracy from 82.73% to 85.42% (+2.69 pp), attributable to more stable gradient estimates. (3) Training beyond 5 epochs is counter-productive for DistilBERT MTL, with the 8-epoch variant achieving lower category accuracy (84.27%) despite higher sentiment accuracy (78.43%), suggesting mild overfitting. (4) The DeBERTa MTL (6ep) achieves 95.70% top-2 accuracy and 98.35% AUC-ROC, indicating excellent probability calibration.

4.4.4 Family IV: Hybrid Feature-Fusion Model (MA Dataset)

Experiment 16 evaluates whether fusing DeBERTa-v3 contextual representations with Word2Vec embeddings and 31 handcrafted linguistic features improves classification.

Table 4.8: Hybrid model results on the MA dataset (7-class, 21,218 test samples).

Metric	Exp. 16 (Hybrid)	Exp. 12 (DeBERTa ST)	Δ
Accuracy	0.8584	**0.9406**	−8.22 pp
F1 (macro)	0.8471	**0.9309**	−8.38 pp
F1 (weighted)	0.8594	**0.9403**	−8.09 pp
MCC	0.8179	**0.9224**	−10.45 pp
AUC-ROC (OvR)	0.9647	—	—
Top-2 Acc.	0.9346	—	—

The hybrid model substantially underperforms the pure DeBERTa single-task model, with an 8.22 pp accuracy deficit. This unexpected result is attributed to: (1) *feature interference*, where the concatenation of general-domain static embeddings dilutes the transformer's contextual signal; (2) *optimisation difficulty* arising from heterogeneous convergence rates across three feature streams; and (3) *domain mismatch* between static embeddings (trained on news/Wikipedia) and mental health discourse. Despite lower accuracy, the hybrid model achieves 96.47% AUC-ROC and 93.46% top-2 accuracy, indicating well-calibrated probability estimates.

4.4.5 Family V: Multi-Seed Ensemble (MA Dataset)

Experiment 17 trains DeBERTa-v3-base with Multi-Sample Dropout across three random seeds and evaluates both individual seed performance and the ensemble.

Table 4.9: Multi-seed ensemble results on the MA dataset (7-class, 21,218 test samples).

Model	Accuracy	F1 (weighted)	F1 (macro)
Seed 42	0.9175	0.9181	0.8998
Seed 1337	0.9159	0.9164	0.9005
Seed 2024	0.9165	0.9169	0.8996
Ensemble	**0.9212**	**0.9219**	**0.9038**

Table 4.10: Seed stability analysis for Experiment 17.

Metric	Mean (3 seeds)	Std. Dev.	Range
Accuracy	0.9166	0.0008	0.0016
F1 (weighted)	0.9171	0.0009	0.0017
F1 (macro)	0.9000	0.0005	0.0009

Key findings. (1) Seed stability is excellent ($\sigma < 0.001$), confirming high reproducibility. (2) The ensemble improves over the best individual seed by +0.37 pp accuracy, a modest but consistent gain expected given the low inter-seed variance. (3) The ensemble (92.12%) underperforms DeBERTa-base ST (Exp. 12, 94.06%) by 1.94 pp, attributable to differences in the classification head architecture and over-regularisation from label smoothing.

Table 4.11: Per-class F1 comparison for Experiment 17.

Class	Seed 42	Seed 1337	Seed 2024	Ensemble
Anxiety	0.9465	0.9445	0.9474	**0.9483**
Bipolar	**0.9369**	0.9353	0.9329	0.9348
Depression	0.8988	0.8948	0.8970	**0.9030**
Normal	0.9723	0.9732	0.9746	**0.9751**
Pers. Disorder	0.7892	**0.7947**	0.7955	0.7915
Stress	0.8869	**0.8997**	0.8885	0.9009
Suicidal	0.8681	0.8616	0.8610	**0.8726**

4.5 Comparative Analysis with State of the Art

4.5.1 Cross-Model Performance Ranking

Table 4.12 consolidates the best result from each model family, enabling a holistic comparison of the experimental programme.

Table 4.12: Global performance ranking across all model families.

Rank	Model	Dataset	Acc.	$\mathbf{F1_{macro}}$	Params
1	DeBERTa-base ST (Exp. 12)	MA 7-class	**0.9406**	**0.9309**	139M
2	DeBERTa-base frozen (Exp. 13)	MA 7-class	0.9385	0.9291	∼70M
3	DistilBERT ST (Exp. 14)	MA 7-class	0.9378	0.9272	66M
4	DistilBERT ST (Exp. 8)	MA 7-class	0.9357	0.9220	66M
5	Ensemble (Exp. 17)	MA 7-class	0.9212	0.9038	184M×3
6	Scratch (Exp. 7)	MA 7-class	0.8886	0.8759	∼15M
7	DeBERTa MTL (8ep/16)	Reddit 6-class	0.8824	0.8828	139M
8	DeBERTa MTL (6ep/8)	Reddit 6-class	0.8723	0.8719	184M
9	Hybrid (Exp. 16)	MA 7-class	0.8584	0.8471	∼190M
10	DistilBERT MTL (5ep/32)	Reddit 6-class	0.8542	0.8543	66M
11	Scratch (Exp. 10)	Reddit 6-class	0.7480	0.7455	∼20M

4.5.2 Comparison with Published Approaches

Table 4.13 positions the best-performing models against representative state-of-the-art approaches from the literature. Direct comparison is limited by differences in datasets,

label taxonomies, and evaluation protocols.

Table 4.13: Comparison with published approaches for mental health text classification. Values for prior work are approximate, drawn from the literature reviewed in Chapter 1.

Approach	Architecture	Classes	Metric	Value
Ours (DeBERTa ST, Exp. 12)	DeBERTa-base	7	$F1_{macro}$	**0.9309**
Ours (DistilBERT ST, Exp. 14)	DistilBERT	7	$F1_{macro}$	0.9272
Ours (DeBERTa MTL)	DeBERTa-base MTL	6	$F1_{macro}$	0.8828
Ji et al. (2022)	MentalBERT	5	$F1_{macro}$	∼0.82
Harrigian et al. (2021)	RoBERTa + Metadata	2	F1	∼0.86
Turcan & McKeown (2019)	BERT-base	9	Accuracy	∼0.75
Benton et al. (2017)	MTL + user feat.	4	Accuracy	∼0.77

The best model achieves 93.09% macro F1 on a 7-class task, which is competitive with or exceeds published results on comparable benchmarks. The MTL approach on Reddit (88.28% F1 macro on 6 classes) exceeds Turcan & McKeown (2019) on a 9-class task (∼75%) and is competitive with MentalBERT (∼82%), despite using a general-purpose backbone rather than a domain-specific model. The sentiment auxiliary task provides a privacy-preserving alternative to user-level metadata employed by prior work.

4.6 Discussion

4.6.1 The Dominance of Pre-Trained Transformers

The most striking finding is the overwhelming performance advantage of pre-trained transformer backbones over scratch-built models. On the MA dataset, the transition from the best scratch model (Exp. 7, 88.86%) to the best transformer (Exp. 12, 94.06%) represents a +5.20 pp accuracy improvement—achieved with approximately 10× fewer training epochs and no handcrafted linguistic features. This underscores that pre-trained contextual representations encode rich linguistic knowledge directly transferable to mental health classification.

4.6.2 DeBERTa vs. DistilBERT: Architectural Insights

DeBERTa's consistent superiority (94.06% vs. 93.78% on MA; 88.24% vs. 85.42% on Reddit) is attributed to two innovations: (1) *disentangled attention*, which models content-position interactions more expressively—relevant for mental health text where word position within a narrative may carry diagnostic significance; and (2) *depth*, with 12 layers

providing additional capacity for learning nuanced distinctions between clinically similar categories. However, DistilBERT's 93.78% accuracy with only 66M parameters demonstrates that the capacity–performance relationship is not linear, making it an excellent choice for resource-constrained deployment.

4.6.3 The Failure of Feature Fusion

The hybrid model's underperformance (85.84% vs. 94.06%) contradicts the expectation that combining information sources should improve classification. Three factors explain this result: (1) *representation redundancy*—the DeBERTa-v3 encoder already captures the semantic information encoded in Word2Vec embeddings and many of the patterns targeted by handcrafted features; (2) *optimisation challenges* from heterogeneous convergence rates across three feature streams; and (3) *domain mismatch* between general-domain static embeddings and mental health discourse.

4.6.4 Multi-Task Learning: Promise and Limitations

The MTL experiments demonstrate that sentiment analysis serves as a useful auxiliary task, with the DeBERTa MTL model achieving the best Reddit performance (88.24%). However, the sentiment task's contribution is bounded: derived labels from lexicon-based analysis contain inherent noise (sentiment accuracy plateaus at ∼78%), the weighting strategy (dynamic vs. static) matters less than backbone capacity, and the auxiliary task likely functions primarily as a regulariser rather than providing genuine sentiment-disorder knowledge transfer.

4.6.5 Class-Level Analysis

Table 4.14: Cross-experiment class difficulty ranking (MA dataset).

Rank	Class	Avg. F1	Primary Confusion
1 (easiest)	Normal	0.972	— (distinct lexical profile)
2	Anxiety	0.953	Stress
3	Bipolar	0.940	Depression
4	Depression	0.929	Suicidal
5	Suicidal	0.908	Depression
6	Stress	0.896	Anxiety
7 (hardest)	Personality Disorder	0.891	Depression, Bipolar

The Depression–Suicidal confusion is clinically expected, as suicidal ideation frequently co-occurs with depressive episodes, creating genuine label ambiguity. Personality Disorder remains the most challenging minority class, with characteristic high recall but low precision trade-offs.

4.6.6 The Focal Loss Failure

Experiment 15's poor performance (84.29%) is attributed to the interaction between Focal Loss and balanced class weights: both mechanisms independently address class imbalance, and their combination creates an excessively aggressive minority-class bias. The focusing parameter $\gamma = 2.0$, while standard for object detection, may be inappropriate for text classification where class boundaries are less distinct.

4.6.7 Computational Considerations

Table 4.15: Approximate computational costs (Google Colab T4 GPU).

Model Family	Training	Inference	Memory
Scratch (50 epochs)	2–3 hours	∼0.1 ms/sample	∼2 GB
DistilBERT (10 epochs)	1–2 hours	∼0.3 ms/sample	∼4 GB
DeBERTa (10 epochs)	2–4 hours	∼0.6 ms/sample	∼8 GB
Hybrid (15 epochs)	3–5 hours	∼0.6 ms/sample	∼10 GB
Ensemble (3×10 ep.)	6–12 hours	∼1.8 ms/sample	∼24 GB

The DeBERTa single-task model offers the best performance-to-cost ratio: it achieves the highest accuracy (94.06%) with moderate computational requirements.

4.7 Limitations and Future Work

4.7.1 Limitations

1. **Dataset limitations**: Both datasets rely on self-reported labels rather than clinical diagnoses, introducing label noise that cannot be resolved without professional annotation.

2. **Class imbalance**: The MA dataset's Personality Disorder class (447 samples vs. 6,571 for Normal) remains systematically under-served despite balanced class weights.

3. **Single-language limitation**: All experiments use English-language text; transferability to other languages is unvalidated.

4. **Computational constraints**: The Google Colab environment (16 GB VRAM, 5-hour sessions) precluded larger architectures and more extensive hyperparameter searches.

5. **Temporal validity**: Pre-trained models may not reflect evolving mental health discourse patterns.

6. **Model interpretability**: Transformer models operate as black boxes; attention weights are unreliable indicators of feature importance (Jain & Wallace, 2019).

7. **Auxiliary task quality**: Derived sentiment labels introduce systematic noise; higher-quality auxiliary labels could yield stronger transfer.

4.7.2 Future Work

1. **Algerian dialect adaptation**: Extending the framework to Algerian Arabic (Darja)—a low-resource, code-switched dialect with no standardised orthography—would require dialect-specific pre-training or transfer learning from multilingual/Arabic backbones, and would substantially broaden the applicability of this work to underserved North African populations.

2. **Domain-specific pre-training**: Fine-tuning from MentalBERT or PsychBERT could provide representations attuned to clinical language.

3. **Larger architectures**: Evaluating DeBERTa-large (304M) could reveal whether additional capacity improves discrimination between clinically similar categories.

4. **Improved auxiliary tasks**: Emotion classification, clinical severity estimation, or symptom extraction could strengthen the multi-task learning signal.

5. **Explainable AI**: Integrating Integrated Gradients, SHAP, or BertViz would enhance clinical trust.

6. **Cross-lingual extension**: Multilingual transformers (XLM-RoBERTa, mDeBERTa) could broaden impact.

7. **Multimodal learning**: Incorporating user metadata alongside text could provide complementary signals.

8. **Hierarchical MTL**: Task hierarchies reflecting clinical relationships could enable more structured inductive transfer.

9. **Continual learning**: Mechanisms for adapting to evolving language patterns without catastrophic forgetting would improve deployment viability.

10. **Flash Attention**: Replacing standard scaled dot-product attention with Flash Attention [57] could reduce memory consumption and training time for longer sequences, enabling larger batch sizes or longer context windows without sacrificing model quality.

General Conclusion

This thesis investigated the application of multi-task learning (MTL) to the problem of mental health disorder classification from user-written text. Starting from a critical review of the state of the art, the work progressed through a systematic experimental programme spanning custom-built attention-based architectures, fine-tuned transformer backbones (DistilBERT and DeBERTa), and hybrid multi-task learning systems jointly optimising disorder classification and sentiment analysis.

The experimental results, obtained on two complementary datasets—a seven-class clinical conditions dataset and a six-class Reddit dataset—demonstrated that transformer-based backbones substantially outperform scratch-built attention models, with the fine-tuned DeBERTa-base model achieving the best overall performance (94.06% accuracy, 93.09% macro F1-score, and 92.24% Matthews Correlation Coefficient on the held-out test set). Layer-freezing strategies were shown to preserve nearly all of the full fine-tuning performance (93.85% vs. 94.06% accuracy) while reducing the number of trainable parameters by approximately 50%, offering an attractive efficiency–performance trade-off. The multi-task learning variant, incorporating sentiment analysis as an auxiliary task, achieved 95.70% top-2 accuracy and 98.35% AUC-ROC, indicating strong probability calibration and confirming that the auxiliary sentiment signal contributes meaningfully to the richness of the learned representations. Conversely, aggressive minority-class rebalancing strategies such as Focal Loss were found to degrade overall accuracy by introducing an excessive bias toward underrepresented classes, underscoring the delicate balance required when addressing class imbalance in clinically sensitive applications.

Despite these promising results, several limitations should be acknowledged. First, the datasets used in this study, while substantial, are derived from social media and online forum text, which may not fully generalise to clinical populations or other linguistic and cultural contexts. Second, the auxiliary sentiment labels used in the multi-task learning framework were obtained through automated annotation rather than expert clinical assessment, which may introduce noise into the auxiliary signal. Third, certain semantically overlapping categories—most notably Depression and Anxiety, or Stress and Suicidal ideation—remain difficult to disambiguate purely from textual cues, as reflected in

the per-class error analysis. Finally, the models developed in this work are intended as decision-support and screening tools rather than diagnostic instruments, and any real-world deployment would require careful ethical consideration, clinical validation, and human oversight.

Building on the findings of this thesis, several directions for future research can be identified:

- Extending the multi-task learning framework with additional auxiliary tasks (e.g., emotion recognition, linguistic style detection) to further enrich the learned representations.
- Investigating cross-dataset and cross-domain generalisation, including evaluation on clinically annotated corpora.
- Exploring more sophisticated class-imbalance mitigation strategies that preserve majority-class performance while improving minority-class recall, such as cost-sensitive learning combined with data augmentation.
- Incorporating explainability techniques (e.g., attention visualisation, SHAP) to improve the interpretability and clinical trustworthiness of model predictions.
- Extending the framework to Algerian dialect (Darja) text, a low-resource, code-switched variety with no standardised orthography, which would require dialect-specific pre-training or transfer learning from multilingual/Arabic backbones and would broaden the applicability of this work to underserved North African populations.
- Replacing standard scaled dot-product attention with Flash Attention to reduce memory consumption and training time for longer sequences, enabling larger batch sizes or extended context windows without sacrificing model quality.
- Studying the deployment of such models within ethically governed, human-in-the-loop screening pipelines, in collaboration with mental health professionals.

In conclusion, this thesis has demonstrated that multi-task learning, when combined with modern transformer architectures, constitutes a promising direction for advancing automatic mental health disorder classification from text, while also highlighting the methodological care required to responsibly address the inherent challenges of this sensitive application domain.

Bibliography

[1] S. Singhal, D. L. Cooke, R. L. Villalobos, *et al.*, "Machine learning for mental health: Applications, challenges, and the clinician's role," *Current Psychiatry Reports*, vol. 26, no. 11, pp. 596–605, 2024.

[2] S. M. Tariq *et al.*, "A novel co-training-based approach for the classification of mental illnesses using social media posts," *IEEE Access*, vol. 7, pp. 170658–170672, 2019.

[3] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific Reports*, vol. 10, no. 1, 2020.

[4] W. R. dos Santos, R. L. de Oliveira, and I. Paraboni, "SetembroBR: a social media corpus for depression and anxiety disorder prediction," *Language Resources and Evaluation*, vol. 58, no. 1, pp. 273–300, 2023.

[5] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Generation Computer Systems*, vol. 124, pp. 480–494, 2021.

[6] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. P. Hubbard, R. J. B. Dobson, and R. Dutta, "Characterisation of mental health conditions in social media using informed deep learning," *Scientific Reports*, vol. 7, p. 45141, 2017.

[7] Y. Ophir, R. Tikochinski, C. S. C. Asterhan, I. Sisso, and R. Reichart, "Deep neural networks detect suicide risk from textual Facebook posts," *Scientific Reports*, vol. 10, no. 1, 2020.

[8] Z.-A. Huang, Y. Hu, R. Liu, X. Xue, Z. Zhu, L. Song, and K. C. Tan, "Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 4, pp. 1137–1149, 2023.

[9] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, 2021.

[10] Y. Xiao, X. Liang, X. Liu, *et al.*, "Generative multi-task learning for text classification," *IEEE Access*, vol. 8, pp. 73310–73320, 2020.

[11] J. Li, S. Zhang, Y. Zhang, H. Lin, and J. Wang, "Multifeature fusion attention network for suicide risk assessment based on social media: Algorithm development and validation," *JMIR Medical Informatics*, vol. 9, no. 10, p. e28227, 2021.

[12] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "DepressionNet: Learning multimodalities with user post summarization for depression detection on social media," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1847–1851, 2021.

[13] J. Zhang and Y. Guo, "Multilevel depression status detection based on fine-grained prompt learning," *Pattern Recognition Letters*, vol. 178, pp. 167–173, 2024.

[14] T. Noraset, K. Chatrinan, T. Tawichsri, T. Thaipisutikul, and S. Tuarob, "Language-agnostic deep learning framework for automatic monitoring of population-level mental health from social networks," *Journal of Biomedical Informatics*, vol. 133, p. 104145, 2022.

[15] B. Shickel, S. He, A. Adhikari, and P. Rashidi, "Automatic detection and classification of cognitive distortions in mental health text," in *Proceedings of the IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 275–280, 2020.

[16] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine*, vol. 135, p. 104499, 2021.

[17] Z. Chen, D. Wang, L. Lou, S. Zhang, X. Zhao, S. Jiang, J. Yu, and J. Xiao, "Text-guided multimodal depression detection via cross-modal feature reconstruction and decomposition," *Information Fusion*, vol. 117, p. 102861, 2025.

[18] E. J. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books, 2019.

[19] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, 2017.

[20] World Health Organization, *World Mental Health Report: Transforming Mental Health for All*. Geneva: WHO, 2022.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *NeurIPS 2019 Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

[24] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[25] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[26] A. Esteva, A. Robicquet, B. Ramsundar, *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[27] P. Rajpurkar, J. Irvin, K. Zhu, *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017.

[28] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, *et al.*, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[29] A. Rajkomar, E. Oren, K. Chen, *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[30] A. Esteva, B. Kuprel, R. A. Novoa, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[31] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[32] F. Ringeval, B. Schuller, M. Valstar, *et al.*, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop*, pp. 3–12, 2019.

[33] A. Sano, A. Z. Yu, A. W. McHill, *et al.*, "Prediction of happy-unhappy situations in daily life using wearable sensing and machine learning," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 471–484, 2018.

[34] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.

[35] S. Wu, K. Roberts, S. Datta, *et al.*, "Deep learning in clinical natural language processing: A methodical review," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 457–470, 2020.

[36] G. Bedi, F. Carrillo, G. A. Cecchi, *et al.*, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *npj Schizophrenia*, vol. 1, no. 1, pp. 1–7, 2015.

[37] J. W. Pennebaker, R. J. Booth, R. L. Boyd, and M. E. Francis, *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates, 2015.

[38] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature Medicine*, vol. 25, no. 1, pp. 37–43, 2019.

[39] K. Harrigian, C. Aguirre, and M. Dredze, "On the state of social media data for mental health research," in *Proceedings of the 7th Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pp. 15–24, 2021.

[40] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3543–3556, 2019.

[41] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

[42] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*. Arlington, VA: American Psychiatric Publishing, 5 ed., 2013.

[43] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition and Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[44] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 128–137, 2013.

[45] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018.

[46] M. Sekulic, M. Mieskes, and M. Strube, "Adapting deep learning methods for mental health prediction on social media," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT)*, pp. 322–327, 2018.

[47] R. S. Lazarus and S. Folkman, *Stress, Appraisal, and Coping*. New York: Springer, 1984.

[48] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. 3rd draft ed., 2023. Available: https://web.stanford.edu/~jurafsky/slp3/.

[49] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the European Conference on Machine Learning (ECML)*, pp. 137–142, 1998.

[50] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3111–3119, 2013.

[51] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[52] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2 ed., 2020.

[53] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[54] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[55] M. Zampieri, T. Ranasinghe, D. Sarkar, and A. Ororbia, "Offensive language identification with multi-task learning," *Journal of Intelligent Information Systems*, vol. 61, pp. 201–219, 2023.

[56] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[57] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," in *Advances in Neural Information Processing Systems*, 2022.