ENHANCING MENTAL HEALTH TEXT CLASSIFICATION ON SOCIAL MEDIA: A COMPARATIVE ANALYSIS OF TRADITIONAL METHODS AND DEBERTA-V3-LARGE FINE-TUNING

A THESIS SUBMITTED
TO THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ANKARA UNIVERSITY

by

Serkan ORHAN

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE IN
ARTIFICIAL INTELLIGENCE TECHNOLOGY

ANKARA
2026

ABSTRACT

Master's Thesis

ENHANCING MENTAL HEALTH TEXT CLASSIFICATION ON SOCIAL MEDIA: A COMPARATIVE ANALYSIS OF TRADITIONAL METHODS AND DEBERTA-V3-LARGE FINE-TUNING

Serkan ORHAN

Ankara University
Graduate School of Natural and Applied Sciences
Department of Artificial Intelligence Technology

Supervisor: Prof. Dr. Asım Egemen YILMAZ

Mental health issues are a growing public health concern, and a substantial amount of text relevant to their early detection is now sourced from social media platforms. The semantic ambiguity of user-generated text, however, makes automated classification difficult. This thesis examines whether a fine-tuned Transformer can outperform classical and hybrid baselines on a mental-health text classification task. The study uses the publicly available Kaggle dataset "Sentiment Analysis for Mental Health," which aggregates Reddit and Twitter posts under seven mental-health labels. Following the reference study (Aksoy, 2025), the corpus is restricted to its three most populated classes (Normal, Depression, Suicidal) and a derived binary variant (Normal, Depression/ Suicidal), with an 80/10/10 stratified split. In the first stage, the TF-IDF based SVC, SGD, and Random Forest classifiers and the USE+BiLSTM hybrid from the reference study are re-implemented; the reproduced models score within one accuracy point of the published values. In the second stage, DeBERTa-v3-Large is fine-tuned on the same data partition. The proposed model reaches 88.35% accuracy and 0.8712 macro F1 on the three-class task and 98.97% accuracy and 0.9892 macro F1 on the binary task, an improvement of about 7 to 8 macro F1 points over the strongest baseline. An Optuna hyperparameter search produced no further gain, and a five-fold cross-validation confirmed stability (test macro F1 std. below 0.003). An error analysis shows that the model recovers most implicit suicidal expressions while its remaining errors lie along the Depression-Suicidal axis.

July 2026, 78 pages
Key Words: Mental Health, Text Classification, Natural Language Processing, Transformer Models, DeBERTa, Social Media Analysis, Sentiment Analysis, Fine-tuning

ÖZET

Master's Thesis

SOSYAL MEDYADA RUH SAĞLIĞI METİN SINIFLANDIRMASININ İYİLEŞTİRİLMESİ: GELENEKSEL YÖNTEMLER İLE DEBERTA-V3 İNCE AYARININ KARŞILAŞTIRMALI ANALİZİ

Serkan ORHAN

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Yapay Zeka Teknolojileri Anabilim Dalı

Danışman: Prof. Dr. Asım Egemen YILMAZ

Ruh sağlığı sorunları giderek artan bir halk sağlığı endişesi olup, bu sorunların erken teşhisine ilişkin metinlerin önemli bir kısmı artık sosyal medya platformlarından elde edilmektedir.Ancak kullanıcı tarafından üretilen metinlerin anlamsal belirsizliği, otomatik sınıflandırmayı zorlaştırmaktadır. Bu tez, ince ayar yapılmış bir Transformer modelinin ruh sağlığı metin sınıflandırmasında klasik ve hibrit referans modelleri geçip geçemediğini incelemektedir. Çalışmada, Reddit ve Twitter gönderilerini yedi ruh sağlığı etiketi altında toplayan, halka açık "Sentiment Analysis for Mental Health" adlı Kaggle veri seti kullanılmıştır. Referans çalışmayı (Aksoy, 2025) takiben veri seti, en kalabalık üç sınıfa (Normal, Depression, Suicidal) ve türetilmiş ikili bir varyanta (Normal, Depression/Suicidal) indirgenmiş, 80/10/10 katmanlı bölme uygulanmıştır. İlk aşamada, referans çalışmadaki TF-IDF tabanlı SVC, SGD ve Random Forest sınıflandırıcıları ile USE+BiLSTM hibrit modeli yeniden uygulanmış ve yayımlanmış değerlerin bir yüzde puanı içinde sonuçlar elde edilmiştir. İkinci aşamada DeBERTa-v3-Large modeli aynı veri bölmesi üzerinde ince ayar yapılmıştır. Önerilen model, üç sınıflı görevde %88.35 doğruluk ve 0.8712 makro F1, ikili görevde ise %98.97 doğruluk ve 0.9892 makro F1 elde etmiş; bu, en güçlü referans modelin yaklaşık 7-8 makro F1 puanı üzerindedir. Sonuçların sağlamlığı, önerilen yapılandırma üzerinde anlamlı bir iyileşme üretmeyen Optuna hiperparametre araması ve test makro F1 standart sapmasını 0.003'ün altında tutan beş katlı çapraz doğrulama ile teyit edilmiştir. Hata analizi, modelin örtük intihar düşüncesi içeren ifadelerin çoğunu doğru yakaladığını ancak kalan hatalarının Depression-Suicidal ekseninde toplandığını ortaya koymuştur.

Temmuz 2026, 78 sayfa
Anahtar Kelimeler: Ruh Sağlığı, Metin Sınıflandırma, Doğal Dil İşleme, Transformer Modelleri, DeBERTa, Sosyal Medya Analizi, Duygu Analizi, İnce Ayar

FOREWORD AND ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Prof. Dr. A. Egemen YILMAZ, for his support, patience, and guidance throughout the entirety of my master's research. His invaluable feedback, expertise, and encouragement have been the guiding force behind the successful completion of this thesis.

I am also deeply thankful to my family for their patience and moral support. They have been my safe haven during the most challenging phases of my academic journey, always standing by my side and believing in my potential.

Above all, I dedicate this work to the cherished memory of my late father, Haydar ORHAN. Beyond being an exceptional mathematics teacher, his lifelong wish was always for us to receive the best possible education, making him my greatest and most important teacher. Although he is no longer with us to witness this milestone, his enduring wisdom, values, and the love for learning he instilled in me have guided me through every page of this research. This achievement is as much his as it is mine.

Serkan ORHAN
Ankara, July 2026

TABLE OF CONTENTS

SYMBOLS AND ABBREVIATIONS

α	R-Drop loss coefficient; SGD regularisation parameter
β	effective number of samples
γ	Layer-wise Learning Rate Decay factor
η_0	Base learning rate
AdamW	Adam optimiser with decoupled Weight decay
ALBERT	A Lite BERT for self-supervised learning of language representations
AUROC	Area Under the Receiver Operating Characteristic curve
BERT	Bidirectional Encoder Representations from Transformers
BF16	Brain Floating-Point 16-bit precision
BiLSTM	Bidirectional Long Short-Term Memory
BoW	Bag-of-Words
CNN	Convolutional Neural Network
Conv1D	One-dimensional Convolutional layer
DALY	Disability-Adjusted Life-Year
DeBERTa	Decoding-enhanced BERT with disentangled Attention
DL	Deep Learning
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
ELMo	Embeddings from Language Models
FN	False Negative
FP	False Positive
GBD	Global Burden of Disease
GDES	Gradient-Disentangled Embedding Sharing
GLUE	General Language Understanding Evaluation
GloVe	Global Vectors for word representation
HPO	Hyperparameter Optimisation
LLM	Large Language Model
LLRD	Layer-wise Learning Rate Decay
LSTM	Long Short-Term Memory
ML	Machine Learning

MLM	Masked Language Modeling
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NSP	Next Sentence Prediction
PTSD	Post-Traumatic Stress Disorder
RoBERTa	Robustly Optimised BERT Pretraining Approach
RTD	Replaced-Token Detection
SGD	Stochastic Gradient Descent
SuperGLUE	Super General Language Understanding Evaluation
SVC	Support Vector Classifier
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
TF32	TensorFloat-32 precision
TN	True Negative
TP	True Positive
TPE	Tree-structured Parzen Estimator
ULMFiT	Universal Language Model Fine-tuning
USE	Universal Sentence Encoder
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background and Motivation

Among the public-health challenges of the twenty-first century, mental health disorders rank among the most pressing. Worldwide, mental disorders carried an estimated burden of about 125.3 million disability-adjusted life-years in 2019, roughly one in twenty (4.9%) of all DALYs, and this share has been climbing since 1990 (GBD 2019 Mental Disorders Collaborators, 2022). Depression and anxiety rank among the foremost causes of years lived with disability, while suicide claimed roughly 703,000 lives in 2019 (World Health Organization, 2021). Currently, mental health disorders affect more than one billion individuals worldwide, yet treatment accessibility remains highly disproportionate. According to the World Health Organization (2025), while over 50% of patients in high-income nations receive care, this figure drops to under 10% in low-income regions, compounded by a global median of merely 13 mental health professionals per 100,000 population.

Against this background, social media platforms have been seen as an important source of naturalistic data on mental health. In online communities where a sense of anonymity lowers the threshold for opening up, people speak freely about how they feel (Kamarudin et al., 2020). Guntuku et al. (2017) showed that symptoms of depression and other mental illnesses are reliably observable in the language and posting behaviour of social media users, while De Choudhury et al. (2013) demonstrated that classifiers trained on Twitter behavioural features could predict the beginning of major depressive disorder in the year preceding clinical diagnosis. Therefore, Natural language processing (NLP) has become central to computational mental health research: surveys by Calvo et al. (2017), Garg (2023) and the critical review by Chancellor and De Choudhury (2020) document an expanding body of work reporting 80 to 90% accuracy in predicting specific mental disorders from social media text. However, these methodological concerns motivate the use of stronger model architectures whose capacity matches the complexity of the task.

Fine-tuned models can capture contextual features of language related to distress that elude shallower representations such as bag-of-words or TF-IDF vectors. Aksoy (2025) showed that a Support Vector Classifier with TF-IDF achieves 95.33% binary accuracy, yet performance drops in multi-class (Normal, Depression, Suicidal) settings (accuracy: 81.40%; macro F1: 0.79), highlighting an unresolved challenge in detailed mental health classification. These results motivate a direct empirical comparison between established traditional pipelines and state-of-the-art transformers like DeBERTa-v3-large. This thesis undertakes this comparison, specifically exploring whether DeBERTa-v3's disentangled attention mechanism provides real advantages for nuanced clinical language understanding.

1.2 Problem Definition

This thesis addresses the challenge of automatically classifying mental health conditions from social media text. The core objective is to categorize user posts into specific clinical states, such as normal, depressive, or suicidal. However, treating this task merely as a standard text classification problem overlooks critical nuances. The unique linguistic characteristics of psychological distress make this task more complex than general-purpose sentiment analysis.

First, the inherent noise of social media text presents a major hurdle. As Baldwin et al. (2013) demonstrated, the non-standard spelling, abbreviations, and fragmented grammar typical of user-generated content frequently disrupt standard NLP pipelines. This challenge is even greater in mental health research. Users often express distress by blending clinical terms with everyday slang, creating a unique, domain-specific vocabulary that general-purpose pre-training corpora do not capture.

Second, establishing valid ground-truth labels for mental health data remains a major challenge. As Chancellor and De Choudhury (2020) highlight, researchers rarely have access to formal clinical diagnoses, forcing them to rely on proxies like self-reported conditions, subreddit memberships, or user surveys. While practical, these proxies introduce distinct measurement errors that blur the clinical meaning of the results.

Compounding this issue, the lack of standard annotation schemes across different datasets makes it difficult to compare studies or generalize models, a barrier also emphasized by Garg (2023).

Third, there is an absence of controlled benchmarking between traditional methods and fine-tuned transformers in multi-class settings. Although older pipelines achieve high performance on binary problems, such as the 95.33% accuracy baseline established by Aksoy (2025) using SVC, their effectiveness decreases as the number of clinical categories grows. Compounding this problem is the methodological inconsistency across the literature. Because researchers constantly change both their data features and model architectures at the same time, we cannot easily attribute performance gains to specific models (Chancellor & De Choudhury, 2020). To solve this, as Chancellor and De Choudhury (2020) argue in their critical review, we must evaluate different models under strictly identical dataset and protocol conditions.

Motivated by these difficulties, the principal objective of this thesis is to evaluate how an optimized DeBERTa-v3-large model performs in multi-class classification tasks relative to classical machine learning approaches (specifically TF-IDF with linear classifiers). This research, instead of focusing only on accuracy, also evaluates the reasons for the differences between the two approaches. To enable a controlled comparison of the underlying architectures, the study focuses solely on textual data.

This thesis aims to contribute to the literature by offering a direct comparison between classical text classifiers and modern transformers for multi-class mental health detection. Through this empirical evaluation, the study seeks to help researchers determine whether the performance gains of fine-tuning models like DeBERTa-v3 actually justify their substantial computational costs.

1.3 Purpose and Objectives of the Study

The primary goal of this research is to compare traditional machine learning pipelines with advanced transformer models for mental health text classification. Aksoy (2025)

previously established strong baselines using TF-IDF paired with SVC and SGD classifiers. Building on this foundation, the present thesis investigates whether fine-tuning DeBERTa-v3-Large yields significant and reproducible improvements. By evaluating these models on a three-class dataset (Normal, Depression, Suicidal) from social media, the study aims to determine if the newer architecture outperforms established classical methods. The study's methodology consists of three core phases: reproducing the baseline pipelines, fine-tuning the DeBERTa-v3-Large model, and conducting thorough validation via hyperparameter optimization and 5-fold cross-validation. By controlling the training setups and data splitting strategies across these phases, we can isolate the impact of the transformer architecture. Accordingly, this thesis seeks to accomplish the subsequent primary targets:

1. To ensure baseline reproducibility by recreating the Aksoy (2025) results under identical experimental conditions, thereby establishing a reliable reference point.
2. To execute model fine-tuning on the DeBERTa-v3-large architecture for mental health classification.
3. To conduct hyperparameter optimization to test the model's sensitivity, comparing manual configurations against Optuna's algorithmic tuning.
4. To assess statistical robustness and verify consistent performance gains across different data partitions by utilizing 5-fold stratified cross-validation.
5. To perform a comparative error analysis that identifies the specific linguistic nuances DeBERTa captures better than the baseline SVC.

1.4 Importance and Contributions of the Study

This work's principal contribution is a tightly controlled comparison: by fixing the dataset, the preprocessing, and the evaluation metrics, it lets the model architecture be the only moving part. After independently reproducing the baselines established by Aksoy (2025), we introduce a customized DeBERTa-v3-Large model. To adapt this architecture for clinical text, we applied several specific training methods, including LLRD, R-Drop, class-balanced loss, sliding-window chunking, and layer freezing. This combination of

optimisation techniques is reported and tested as a single fine-tuning protocol, addressing the methodological transparency gap noted in prior work (Calvo et al., 2017; Garg, 2023).

In the three-class task, the fine-tuned DeBERTa model reached a macro-F1 of 0.8712 and an accuracy of 88.35%. Compared to the strongest baseline, this translates to a solid absolute increase of 0.08 in macro-F1 and 6.95% in accuracy. These gains are stable across data splits; our 5-fold cross-validation ensemble maintained a stable macro-F1 of 0.8714 with minimal variance. We also verified our manual hyperparameter choices, as independent Bayesian optimization could not produce a meaningful improvement over the manual DeBERTa-v3-Large fine-tuning configuration; the ten trials varied within only 0.7 F1 points on the validation set, and retraining with the optimal trial yielded a 1-point lower test macro F1. The error analysis shows that the model classifies posts whose label depends on metaphor, irony, or implicit phrasing more accurately than the TF-IDF baselines. Such posts contain no surface n-grams strongly associated with their label and therefore cannot be recovered by classical features.

1.5 Research Questions

The study is guided by four research questions:

RQ1: To what extent does fine-tuning a modern Transformer model (DeBERTa-v3) improve the classification performance of mental health statuses (normal, depression, suicidal) on social media texts compared to traditional machine learning and baseline deep learning methods?

RQ2: How does the Disentangled Attention mechanism of the DeBERTa architecture contribute to capturing implicit semantic nuances, such as metaphors and euphemisms, in mental health discourse that traditional TF-IDF-based models fail to recognize?

RQ3: Is the performance improvement achieved by the DeBERTa-v3 model statistically stable and robust across different data splits, and how do hyperparameter optimization strategies affect its maximum performance?

RQ4: What are the main limitations of the DeBERTa-v3 model in classifying mental health texts, and what types of errors does it commonly make?

1.6 Thesis Structure

The remainder of the thesis is organised as follows.

- 1. Introduction: Outlines the motivation behind the research, defines the problem, and specifies the core objectives.

- 2. Literature Review: Provides a theoretical background by examining the foundational algorithms, models, and methods relevant to the domain, alongside a critical summary of prior related studies.

- 3. Materials and Method: This section outlines the experimental methodology, offering an in-depth overview of the dataset alongside the specific computational steps applied throughout the study

- 4. Results and Discussion: Details the concrete outcomes generated by the study, providing a critical assessment and thorough contextualization of the results.

- 5. Conclusion: Synthesizes the primary findings of the thesis and summarizes the overall outcomes.

2. LITERATURE REVIEW

2.1 Natural Language Processing (NLP) and Text Classification

Natural language processing (NLP) bridges the gap between human communication and computational models. Within this field, text classification like categorizing documents into predefined groups is basic works to real-world applications like clinical screening and sentiment analysis (Kowsari et al., 2019; Sebastiani, 2002). Building a robust classifier requires a complete pipeline of preprocessing, feature extraction, and training; yet, the method of text representation often influences the final accuracy (Szymański, 2014).

Early text-categorization systems relied heavily on the conceptually simple bag-of-words (BoW) model, which strips away grammar and treats documents merely as unordered collections of tokens (Sebastiani, 2002). A known limitation of BoW is that it gives equal weight to every word, allowing frequent but less informative terms to dominate. TF-IDF mitigates this by penalizing words that appear across too many documents, highlighting more discriminative vocabulary (Kowsari et al., 2019). Because of this adjustment, TF-IDF remains a common choice for linear classifiers like support-vector machines (Joachims, 1998; Szymański, 2014). However, both approaches produce high-dimensional sparsity and lack semantic context awareness. Treating every token as an independent variable often prevents them from recognizing the relationship between terms like 'anxiety' and 'panic'.

To address the semantic limitations inherent in sparse models, the field shifted toward distributed word embeddings. Models such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) reshaped how text is represented by projecting words into dense, low-dimensional vectors. Instead of treating terms as isolated counts, these models learn from context and co-occurrence, ensuring that semantically related words cluster together. This allows them to capture complex linguistic relationships that are generally invisible to BoW or TF-IDF. As Szymański (2014) demonstrated, these

distributed representations strike a better balance between dimensionality and accuracy, leaving traditional BoW competitive mostly for very short, simple texts.

While these representational trade-offs are well-documented, with classical methods often favored for interpretability and embeddings for semantic depth (Kowsari et al., 2019), model selection depends on the dataset's nature. In healthcare and mental health mining, the noise of user-generated text makes the semantic generalization offered by embeddings valuable (Lavanya & Sasikala, 2021). Yet, Word2Vec and GloVe generates a uniform, non-dynamic embedding for each distinct term in the vocabulary. This means they are highly blind to polysemy, treating a word the same regardless of its context. Recognition of this inability to grasp context-dependent meanings paved the way for the contextualized, Transformer-based architectures explored in Section 2.4.

2.2 Mental Health Analysis on Social Media

Today, social media platforms offer a window into unfiltered, real-time expressions of psychological distress. Far removed from the constraints of formal clinical settings, users on platforms like Reddit and Twitter frequently discuss conditions like depression or suicidal ideation with candor (Coppersmith et al., 2014; Gkotsis et al., 2017). Recognizing the value of this transparency, a growing wave of research now leverages user-generated text as a passive screening tool for early mental health detection. Machine learning was shown by Coppersmith et al. (2014) to be capable of identifying users at risk of depression and PTSD; a later study by the same group (Coppersmith et al., 2018) extended the idea to suicide-risk screening with NLP, opening a path toward preventive interventions that scale.

However, the spontaneity that makes social media a rich psychological resource also introduces linguistic challenges. As Baldwin et al. (2013) reported across multiple platforms, user-generated text is often characterized by non-standard spelling, fragmented grammar, and heavy slang. Because off-the-shelf NLP tools are typically trained on formal, edited corpora, they frequently struggle when confronted with this noisy reality. This challenge is especially acute on microblogs like Twitter, where character limits strip

away crucial context and force an over-reliance on condensed jargon. Before an algorithm can detect mental health signals, it must navigate the unstructured nature of internet speech.

As the field matured, efforts shifted toward systematizing this growing body of work, with researchers mapping out ethical guidelines, data taxonomies, and common evaluation metrics (Garg, 2023; Kamarudin et al., 2020; Naslund et al., 2020). However, this rapid expansion also exposed structural flaws. Reviewing 75 studies, Chancellor and De Choudhury (2020) found the area repeatedly hampered by inconsistent labelling, pronounced class imbalance, and the frequent absence of external validation. These methodological issues remain relevant to the framework developed in the present work.

2.3 Traditional Machine Learning and Deep Learning Approaches

Before Transformer architectures came to dominate the field, text classification was shaped by a steady progression from rigorous statistical learners to early, purpose-built neural networks. This section reviews both of these families, specifically detailing the models that serve as the comparative baselines for the present study. In the classical era, the primary challenge was handling the massive sparsity of TF-IDF features. Support Vector Machines (SVMs) emerged as a strong standard in this regime; by finding the maximum-margin hyperplane, they navigated high-dimensional spaces where other algorithms typically struggled (Cortes & Vapnik, 1995; Joachims, 1998). Alongside SVMs, linear models trained via Stochastic Gradient Descent (SGD) offered a scalable alternative for large corpora (Bottou, 2010), while ensemble methods like Random Forests provided robustness without extreme hyperparameter sensitivity (Breiman, 2001). Because of their reliability and computational efficiency, this triad (SVM, SGD, and Random Forest) remains a common baseline for mental health classification, and forms the classical foundation of the comparative analysis in this thesis.

However, as the field recognized the inherent limitations of bag-of-words approaches, deep learning introduced models capable of processing text sequentially. Rather than treating words as isolated counts, architectures like the Long Short-Term Memory

(LSTM) network and one-dimensional Convolutional Neural Networks (CNNs) began capturing the local syntax and temporal flow of language (Hochreiter & Schmidhuber, 1997; Kim, 2014). Bidirectional LSTMs pushed this further by reading sequences in both directions, yielding the kind of context-sensitive representations (Wu & Xing, 2024). Furthermore, hybrid deep learning approaches, such as the CNN-BiLSTM architecture proposed by Aldhyani et al. (2022), have demonstrated effectiveness in early detection tasks like suicidal ideation by combining feature extraction and sequential modeling layers.

Therefore, this evolution toward context awareness culminated in models designed to encode complete sequences into unified dense vectors. The Universal Sentence Encoder (USE), for instance, compresses entire sentences into a single 512-dimensional representation (Cer et al., 2018). This specific architecture is relevant to the present research; because USE serves as the primary deep-learning baseline in the Aksoy (2025) study, it provides the comparative threshold that the proposed DeBERTa-based pipeline aims to surpass. Together, these classical and early deep-learning baselines define the historical progression of the field and set the stage for the modern Transformer architectures.

2.4 Transformer-Based Models

The inability of traditional static embeddings to handle context-dependent meaning motivated a new architectural approach. With the Transformer, Vaswani et al. (2017) removed this bottleneck and redirected NLP research: in place of step-by-step recurrence they introduced multi-headed self-attention, letting a model attend to a whole sequence at once. This decoupled sequence length from computational depth, enabling the training of deeper networks on parallel hardware.

The first major step toward this deep contextualization was ELMo (Peters et al., 2018), which captured word-sense polysemy by deriving embeddings from deep bidirectional LSTMs. However, because ELMo still relied on a recurrent architecture, it remained computationally bottlenecked. A development came with BERT (Devlin et al., 2019),

which paired the Transformer encoder with masked language modeling (MLM) and next-sentence prediction. BERT understands context from both directions simultaneously, and established the 'pre-train then fine-tune' paradigm that continues to guide modern text classification. Table 2.1 provides a brief comparison of transformer-based models.

Table 2.1 Transformer-based model comparison

Model	**Year**	**Parameters**	**Pre-training Objective**	**Key Innovation**
BERT (Base/Large)	2019	110M/340M	MLM + NSP	Bidirectional Transformer encoder
RoBERTa-Large	2019	355M	Dynamic MLM	Larger data, longer training, no NSP
ALBERT	2020	18M/235M	MLM + SOP	Cross-layer parameter sharing
ELECTRA	2020	110M/355M	Replaced-Token Detection	Generator–discriminator efficiency
DeBERTa	2021	140M/400M	MLM + EMD	Disentangled attention (content + position)
DeBERTa-v3	2023	86M/304M	RTD + GDES	ELECTRA-style + gradient-disentangled embedding sharing

Following BERT, research focused on optimizing both its architecture and its training recipes. Variants like RoBERTa demonstrated that training longer on larger datasets with dynamic masking could yield substantial performance gains (Liu et al., 2019). Other efforts focused on efficiency; ALBERT reduced memory consumption through parameter sharing (Lan et al., 2020), while ELECTRA introduced an efficient replaced-token detection objective that cut compute costs by training a discriminator to detect replaced tokens (Clark et al., 2020). Concurrently, frameworks like ULMFiT (Howard & Ruder, 2018) popularized task-adaptive fine-tuning, introducing concepts that later evolved into the layer-wise learning rate decay strategies (Sun et al., 2019) central to the methodology of the present study.

Despite these optimizations, the standard Transformer architecture still treated a token's content and its position as a single fused vector. To overcome this structural limitation, He et al. (2021) introduced DeBERTa, featuring a Disentangled Attention mechanism. DeBERTa architecture is shown in Figure 2.1. In DeBERTa, each token is encoded by a pair of vectors — one capturing its content, the other its relative position — and the attention score is assembled from three interaction terms that mix these two factors. This separation allows the model to reason more flexibly about the distance and order of words, a critical advantage when parsing unstructured social media text.

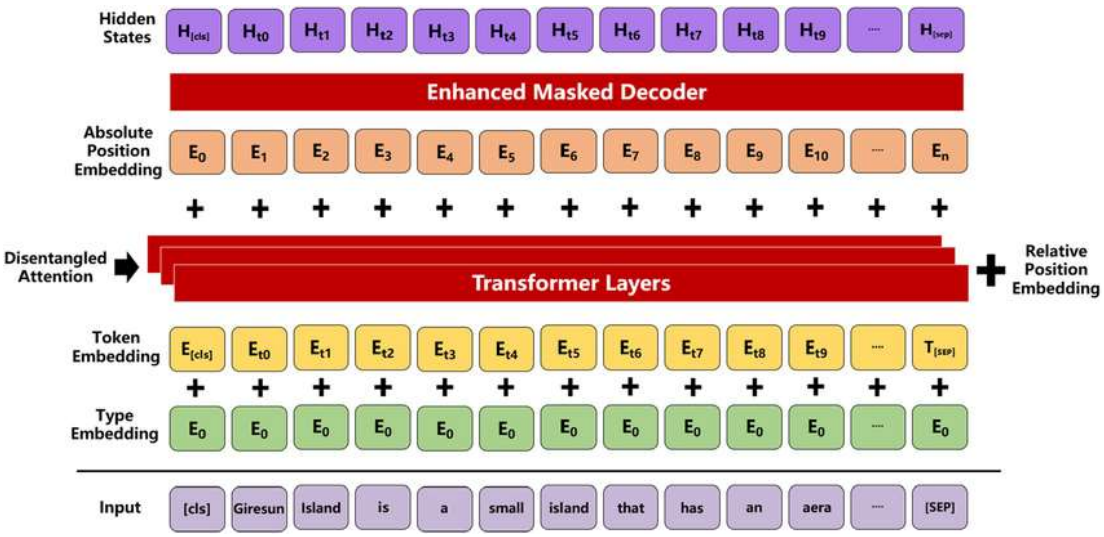

Figure 2.1 The architecture of DeBERTa (Li et al., 2024)

Building upon this foundation, DeBERTa-v3 (He et al., 2023) abandoned the traditional MLM objective in favor of ELECTRA's more efficient replaced-token detection. Crucially, v3 solved the gradient cancellation problem inherent in ELECTRA by implementing Gradient-Disentangled Embedding Sharing (GDES), allowing the generator and discriminator to update embeddings without interfering with each other. Because it combines an attention mechanism formulation, an efficient pre-training objective, and consistent embedding updates, DeBERTa-v3-Large outperforms comparable models such as RoBERTa-Large on natural language understanding benchmarks like MNLI and GLUE (He et al., 2023). These architectural advantages make DeBERTa-v3-Large a strong backbone for the mental-health classification task addressed in this thesis.

2.5 Fine-Tuning Strategies

Adapting a massive pre-trained model to a narrow, relatively small downstream dataset is a delicate process. Without careful regularization, models are prone to catastrophic forgetting or severe over-fitting. While standard Dropout (Srivastava et al., 2014) remains a fundamental safeguard within Transformer architectures, the complexity of mental health text often demands more robust stabilization. To this end, techniques like R-Drop (Liang et al., 2021) offer a stronger alternative. R-Drop bridges the train-test gap using Kullback-Leibler divergence to enforce consistency between two independent Dropout passes. When paired with Label Smoothing (Müller et al., 2019), which redistributes a small probability mass to prevent the model from becoming rigidly overconfident, these regularization strategies help ensure the model generalizes rather than simply memorizing the set of training.

Beyond the over-fitting risk, fine-tuning in this domain is complicated by class imbalance. In social media datasets, acute signals like suicidal ideation are heavily outnumbered by general expressions of depression or stress. Traditional inverse-frequency weighting often overcompensates in these scenarios, penalizing abundant classes too harshly. The Class-Balanced Loss introduced by Cui et al. (2019) provides a more robust alternative. To strike a well-reasoned balance, this method assigns weights to categories according to their effective sample size, which is derived using a smoothed exponential formula. It ensures that the model remains sensitive to critical minority classes without skewing its overall predictive accuracy.

Finally, updating the network requires a layer-specific approach. Applying a uniform learning rate across the entire Transformer can erase the basic linguistic features stored in the lower layers, since only the higher layers handle task-specific details (Howard & Ruder, 2018; Sun et al., 2019). To navigate this, Layer-Wise Learning-Rate Decay (LLRD) applies progressively smaller updates to the deeper, more generalized layers. This precision is further enhanced by selective layer freezing (Lee et al., 2019), which locks the lower-layer parameters to balance computational efficiency with downstream performance. Driven by the AdamW optimizer's decoupled weight decay (Loshchilov &

Hutter, 2019), this combination (LLRD, strategic freezing, and AdamW) constitutes the robust optimization backbone of the fine-tuning protocol adopted in this thesis.

Table 2.2 Fine-Tuning strategies summary

Strategy	Purpose	Mechanism	Reference
Dropout	Regularisation	Random unit deactivation	Srivastava et al. (2014)
R-Drop	Train-test consistency	KL divergence between two dropout passes	Liang et al. (2021)
Label Smoothing	Confidence calibration	Soft target distribution	Müller et al. (2019)
Class-Balanced Loss	Class imbalance	Effective sample weighting (β=0.999)	Cui et al. (2019)
LLRD	Layer-wise adaptation	Decay learning rate from top to bottom	Howard & Ruder (2018); Sun et al. (2019)
Layer Freezing	Knowledge preservation	Lock lower layers	Lee et al. (2019)
AdamW	Decoupled weight decay	Decouples L2 regularisation from gradient	Loshchilov & Hutter (2019)

2.6 Explainability of Deep Learning Models

Even when a fine-tuned Transformer encoder reaches strong empirical performance, the route from input text to predicted label remains hidden inside the model. A 304M- parameter network such as DeBERTa-v3-Large applies hundreds of millions of multiplications across 24 attention layers before producing a single output vector, and no part of this computation can be read by a reviewer as a textual reason for the prediction. Explainable Artificial Intelligence (XAI) refers to a family of methods that translate the input–output relationship of a model into a form that a human can interpret, typically as per-token attribution scores or local surrogate models. A survey by Danilevsky et al. (2020) groups XAI methods for natural language processing into two algorithmic families: gradient-based and perturbation-based. The present study applies a gradient-based method as its primary explainability tool.

By leveraging a model's differentiable nature, gradient-based techniques calculate how much the output changes relative to individual input tokens via partial derivatives, subsequently accumulating these values over a specific integration route. A prominent example of this is Integrated Gradients (Sundararajan et al., 2017). This method calculates the integral of gradients on a linear trajectory between a neutral baseline and the given input. Crucially, it adheres to two core axioms: sensitivity (assigning non-zero importance to features that alter the prediction) and implementation invariance (producing identical attributions for models that are mathematically equivalent, regardless of their internal architecture). Captum (Kokhlikyan et al., 2020) provides a generic implementation of Integrated Gradients for arbitrary PyTorch models, including the Hugging Face port of DeBERTa-v3-Large used in this thesis. In the present study, Captum's LayerIntegratedGradients module is applied directly to the embedding layer of the proposed model.

Conversely, perturbation-based techniques adopt a black-box perspective regarding the model. They determine the significance of a feature by observing how targeted alterations to the input affect the final prediction. Notable approaches in this category are Occlusion (Zeiler and Fergus, 2014), where specific input parts are masked with a baseline value, and LIME (Ribeiro et al., 2016), which approximates the model's behavior by training an interpretable linear surrogate in the local vicinity of the prediction. Perturbation-based methods can complement gradient-based attributions on tasks where the model is not at saturation; however, when the model exhibits very high confidence on short inputs, single-token perturbations yield unstable rankings, and the present study therefore restricts the explainability analysis to Integrated Gradients.

Two further categories were considered during the design phase and excluded. Mechanistic interpretability libraries such as TransformerLens are built for decoder-only Transformer architectures of the GPT family and do not support the disentangled attention mechanism of DeBERTa, which means they cannot be applied to the proposed model without modifications to the underlying architecture. Production-oriented observability platforms such as Phoenix and TruLens target the monitoring of generative large language models in deployment scenarios, including prompt drift, retrieval-augmented generation

quality, and latency, rather than per-token attribution for fine-tuned classifiers, and therefore fall outside the scope of this analysis.

2.7 Related Studies

To properly position this research, we first establish its direct anchor: the work of Aksoy (2025). Developing a sentiment-analysis system for mental health diagnosis, Aksoy demonstrated that a traditional pipeline (TF-IDF coupled with a well-tuned Support Vector Classifier) remains competitive for short, noisy social media inputs. This thesis adopts Aksoy's work as its principal baseline but structurally upgrades the approach. Rather than relying on static features or small deep-learning models, the present study replaces the underlying architecture with the state-of-the-art DeBERTa-v3-Large encoder, supported by a custom fine-tuning protocol designed to improve performance on the under-represented Suicidal class.

Beyond this specific baseline, the broader literature reflects a continuous effort to maximize the potential of traditional and hybrid methodologies. Rather than transitioning fully to large language models, several researchers have attempted to push classical classifiers to their limits through feature engineering, integrating topic modeling, sentiment polarities, and multidimensional lexicons (Juanita et al., 2025; Kaushik & Sharma, 2024; Wagay & Jahiruddin, 2026). Other efforts have focused on establishing massive benchmarks to compare these traditional methods against prompt-engineered LLMs (Kallstenius et al., 2025). While these studies, alongside the step-by-step guidance of Zhang et al. (2023), provide methodological blueprints for handling imbalanced data, they often encounter a performance ceiling: classical feature engineering generally cannot capture the deep contextual nuances required for complex psychological text.

Recognizing this limitation, the field has shifted toward Transformer-based architectures. As summarized in Table 2.1, modern research has converged on these encoders, routinely reporting high accuracies. Much of this success stems from adapting models like BERT and RoBERTa specifically to the mental health domain. Studies have achieved this either through domain-specific pre-training, yielding strong baselines like MentalBERT (Ji et

al., 2022) and multiMentalRoBERTa (Islam et al., 2025), or by fusing Transformer embeddings with external emotional features (Hossain et al., 2025; Rehman et al., 2025). Direct applications on massive social media corpora, such as Khan et al.'s (2025) use of RoBERTa-Large or the early multi-label blueprints by Gkotsis et al. (2017) and Murarka et al. (2020), confirm that large-scale contextual encoders generally outperform classical pipelines.

Table 2.3 Summary of related studies in mental health text classification

Methods (Study)	**Advantages**	**Limitations**	**Performance Metrics**
Informed Deep Learning on Reddit – 11 disorders (Gkotsis et al., 2017)	Early blueprint for disorder-level classification on large corpora	Pre-Transformer architecture; limited context handling	Accuracy: 91.08% (Binary); 71.37% (11-class)
RoBERTa for multi-label mental-illness classification (Murarka et al., 2020)	Handles isolation-era multi-label data on Reddit	Single model family; no comparison with newer encoders	F1 (macro): 0.89
CNN-BiLSTM hybrid for suicidal ideation (Aldhyani et al., 2022)	Effective early detection; combines ML and DL layers	Mostly binary classification; limited explainability	Accuracy: 95%
MentalBERT – domain-adapted pre-training (Ji et al., 2022)	Pre-trained on mental-health corpora; strong domain baseline	Requires additional pre-training compute; fixed to BERT architecture	Accuracy: 93.38% - (MentalRoBERTa on Dreaddit stress benchmark)
ML + DL methodological framework (Zhang et al., 2023)	Step-by-step guidance on heterogeneous, imbalanced datasets	Methodological focus; no novel model architecture	AUROC: 0.95
ML + sentiment analysis, multiclass (Kaushik & Sharma, 2024)	Comparative study over 52,681 Reddit/Twitter entries	Class imbalance addressed but not fully resolved	Accuracy: 82% (XGBoost)
TF-IDF + SVC (Aksoy, 2025)	Strong baseline for noisy text; computationally efficient	Lacks deep contextual understanding	Acc: 81.40% / F1: 0.79
ML + Topic Modeling (Juanita et al., 2025)	Interpretable topic clusters	Relies on manual feature engineering	Accuracy: 84.2% / F1: 0.864
RoBERTa-Large on 4 student corpora (Khan et al., 2025)	Strong contextual embeddings; timely support for student mental health	Narrow domain (students); limited dataset diversity	Accuracy: 97% (Validation)
Traditional NLP vs. prompt-engineered and fine-tuned LLMs (Kallstenius et al., 2025)	Multi-model benchmark over 51,000+ texts and 7 conditions	Prompt sensitivity; no single dominant approach across conditions	Accuracy: 95% (best fine-tuned LLM)

Table 2.3 Summary of related studies in mental health text classification (continued)

Methods (Study)	Advantages	Limitations	Performance Metrics
Optimized BERT with deep encodings (Rehman et al., 2025)	Improves over vanilla BERT for mental-health sentiment analysis	Added complexity yields only moderate gains	Accuracy: 95.71%
multiMentalRoBERTa – fine-tuned multiclass (Islam et al., 2025)	Covers 5 disorders + neutral, giving a 6-class setup (stress, anxiety, depression, PTSD, suicidal ideation, neutral)	Limited to a fixed 5-class taxonomy	F1 (macro): 0.839 (6-class); F1: 0.870 (5-class)
BERT-Fuse – hybrid lexical/emotional features (Hossain et al., 2025)	Fuses deep embeddings with domain lexicons	More complex architecture; harder to reproduce	Accuracy: 97.05%
Multidimensional textual features + ML (Wagay & Jahiruddin, 2026)	Integrates sentiment polarity and NRC lexicons	High feature engineering overhead; moderate accuracy ceiling	Accuracy: 75.13% (Binary)

Yet, despite this widespread adoption of Transformers, a methodological gap remains. The current literature ranges from TF-IDF baselines to domain-adapted BERT variants, but it often stops short of systematically optimizing the most capable modern encoders for this specific domain. To bridge this gap, the present work aims to demonstrate that a regularized DeBERTa-v3-Large model can outperform both the Aksoy (2025) baseline and current Transformer-based standards in multiclass mental health classification.

3. MATERIALS AND METHOD

This chapter describes the dataset, preprocessing protocol, and experimental environment that form the complete methodological framework utilized to extract clinical signals from unstructured social media text. The proposed DeBERTa-v3-Large model is therefore evaluated within a three-stage comparative design rather than in isolation. The design is organised into three sequential stages; baseline reproduction, fine-tuning of the proposed DeBERTa-v3-Large model, and a validation stage comprising hyperparameter optimisation, and five-fold cross-validation. The structure purposely reflects the reference study of Aksoy (2025) in its first stage so that any performance gains observed in subsequent stages can be attributed primarily to the methodological innovations introduced by this thesis.

3.1 Material

3.1.1 Dataset

Table 3.1 Distribution of the original dataset across seven classes

Class	**Count**	**Proportion (%)**
Normal	16,351	30.83
Depression	15,404	29.04
Suicidal	10,653	20.08
Anxiety	3,888	7.33
Bipolar	2,877	5.42
Stress	2,669	5.03
Personality Disorder	1,201	2.26
TOTAL	53,043	100.00

The experiments in this thesis are conducted on the publicly available Kaggle dataset titled "Sentiment Analysis for Mental Health" (Sarkar, 2024) which is the same corpus

used in the reference study (Aksoy, 2025). Drawn mainly from Reddit and Twitter, the dataset gathers user posts across several platforms and tags each one with a single label out of seven mental-health categories: Normal, Depression, Suicidal, Anxiety, Stress, Bi-Polar, and Personality Disorder. The distribution of the original dataset across the seven classes is presented in Table 3.1.

Consistent with the reference study, this thesis restricts attention to the three most populated classes, and the two classes formed by combining two of these (Normal, Depressed/Suicidal). The remaining four classes (Anxiety, Stress, Bi-Polar, Personality Disorder) are excluded because their combined size is small relative to the three retained classes and their inclusion would introduce confounds unrelated to the research questions of this thesis. After filtering, the working corpus consists of 42,408 instances. Table 3.2 summarises the 3-class, and table 3.3 summarises the 2-class distribution of the filtered corpus before the train/validation/test partition is applied.

Table 3.2 Distribution of the 3-classes

Class	**Count**	**Proportion (%)**
Normal	16,351	38.56
Depression	15,404	36.32
Suicidal	10,653	25.12
TOTAL	42,408	100.00

Table 3.3 Distribution of the binary-classes

Class	**Count**	**Proportion (%)**
Normal	16,351	38.56
Depression+Suicidal	26,057	61.44
TOTAL	42,408	100.00

The dataset is imbalanced: the suicidal class is the least frequent and is, at the same time, the clinically most critical category. This imbalance motivates several design choices in later stages of the methodology, including class-balanced loss weighting for the proposed model.

3.1.2 Preprocessing

Two parallel preprocessing regimes are used in this thesis. The first regime reproduces the reference pipeline of Aksoy (2025) and is applied to the baseline models of Stage 1. The second regime is intentionally minimal and is used for the proposed DeBERTa-v3-Large model of Stage 2, because modern transformer models operate on raw text and exploit surface cues such as punctuation and casing that the traditional pipeline discards (Camacho-Collados & Pilehvar, 2018).

Following Aksoy's (2025) procedure, the corpus is divided into training, validation, and test portions by stratified random sampling at an 80/10/10 ratio. A random seed (42) is used throughout to ensure reproducibility. For the baseline pipeline this yields 33,919 training, 4,240 validation, and 4,240 test instances. For the DeBERTa pipeline an exact-duplicate removal step is applied to the "statement" field before stratification in order to reduce train–test leakage and to avoid artificially inflated accuracy. After deduplication the corpus is reduced by approximately 1.5 percent and the resulting splits contain 33,413 training, 4,177 validation, and 4,177 test instances (Table 3.2). The same 80/10/10 stratified split is used in both cases; only the deduplication step differs.

Table 3.4 Train/Validation/Test split counts after preprocessing

Pipeline	**Train**	**Validation**	**Test**	**TOTAL**
Baseline	33,919	4,240	4,240	42,399
DeBERTa (deduplicated)	33,413	4,177	4,177	41,767

Seven-class to three-class and binary reduction: From the original seven-class setting, two classification problems are defined. The three-class problem retains Normal, Depression, and Suicidal as separate categories. The binary problem merges the two clinically pathological classes (Depression and Suicidal) into a single "Depression+Suicidal" class and contrasts it with the Normal class, following the protocol of Aksoy (2025, Table 4). Both configurations share the same underlying partition described above.

Baseline preprocessing pipeline: The baseline pipeline reflects the steps documented in Aksoy (2025). Each statement is lower-cased, stripped of non-alphanumeric characters, and has common abbreviations expanded using a predefined lexicon (e.g., "u" to "you", "idk" to "i do not know"). Concatenated tokens of length twelve or more are re-segmented using the wordninja library. Stop words are removed using the NLTK English stop-word list, remaining tokens are reduced to their stems using the Porter stemming algorithm, and residual spelling errors in alphabetic tokens of length four or more are corrected with the pyspellchecker library (edit distance one). The cleaned text is subsequently vectorised by a TF-IDF representation.

Preprocessing for the proposed model: For the DeBERTa-v3-Large model, only whitespace normalisation and exact-duplicate removal are applied. Casing, punctuation, stop words, and morphological variation are retained because the pre-trained WordPiece/SentencePiece tokenizer of DeBERTa, combined with the model's self-attention over raw input tokens, has been shown to benefit from these surface signals (He et al., 2021). This intentional asymmetry between the two pipelines is a deliberate methodological choice: it allows the proposed model to exploit the full linguistic richness of the input while the baselines are evaluated under the preprocessing assumptions of the reference study.

Table 3.5 Summary of preprocessing

Feature	Baseline	DeBERTa
Spelling correction	(pyspellchecker)	X
Word segmentation	(wordninja, ≥12 chars)	X
Abbreviation expansion	(custom dict)	X
Lowercasing	(explicit)	X
Max sequence length	~170 (95th percentile)	384 tokens (stride 192)
Long text handling	Truncation	Sliding window
Split ratio	80 / 10 / 10	80 / 10 / 10
Duplicate handling	Kept	Implicit removal (632 fewer)
Freeze strategy	X	Frozen embeddings + first 16 layers
Loss function	Standard CE / hinge	Class-balanced + R-Drop
Optimizer	Adam / SGD	AdamW + LLRD

3.2 Method

3.2.1 Experimental environment

All runs use Google Colab with NVIDIA A100-SXM4 80 GB GPU. The full software stack — Python, PyTorch, CUDA, the Hugging Face libraries, scikit-learn, and the TensorFlow components of the USE+BiLSTM baseline — is listed with exact versions in Table 3.6. Classical machine learning baselines and evaluation metrics are implemented with scikit-learn 1.5.2, and the USE+BiLSTM baseline is built with tensorflow 2.17 combined with tf-keras 2.17 and tensorflow-hub 0.16. Hyperparameter optimization is carried out with Optuna 4.0 using a Tree-structured Parzen Estimator sampler. BF16 mixed precision is enabled to reduce memory footprint and improve throughput on the A100, and TF32 matrix multiplication is activated at the CUDA level.

A fixed random seed of 42 is set for Python, NumPy, and PyTorch at the beginning of every experiment.

Table 3.6 Experimental environment details

Component	Specification
Platform	Google Colab (High RAM)
GPU	NVIDIA A100 (80 GB)
Programming Language	Python 3.12
Deep Learning Framework	PyTorch (with bf16), TensorFlow + tf-keras
Pre-trained model	microsoft/deberta-v3-large (24 layers, ~304M parameters)
Tokenizer	DeBERTa-v3 SentencePiece (max_length=384, stride=192)
Transformer Library	Hugging Face Transformers 4.44.2
ML Library	scikit-learn
DL Library (Baseline)	TensorFlow / TensorFlow Hub
HPO Library	Optuna (TPESampler, 10 trials)
USE model	TensorFlow Hub: Universal Sentence Encoder v4
Data Processing	pandas, NumPy
Visualization	Matplotlib, Seaborn
Reproducibility	SEED=42 (all splits, models, samplers)

3.2.2 Overview of study design

The experimental design of the study consists of three main stages to provide a systematic comparison and validation process: Reproduction of Baseline Models, Proposed Model, and Model Validation and Evaluation. The general flow of this process is presented in Figure 3.1.

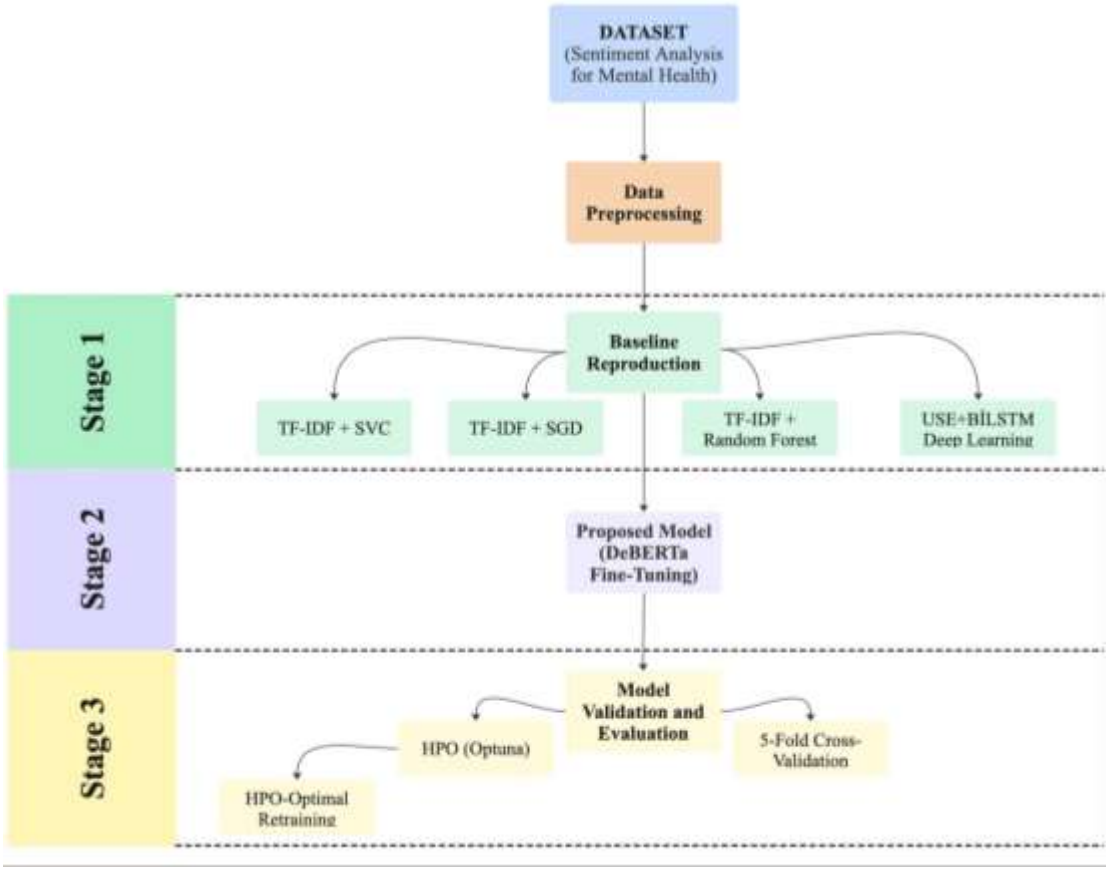

Figure 3.1 General flow diagram of the experimental study

3.2.3 Stage 1: baseline installation

The first stage of the experimental pipeline was dedicated to establishing a reproducible fulcrum for the study. Instead of comparing the proposed model only against published numbers, the traditional and hybrid deep-learning baselines from Aksoy (2025) were re-implemented. The primary goal was to build a controlled environment where any subsequent performance gains could be attributed to the new proposed architecture.

Three TF-IDF-based classifiers are reproduced. The Support Vector Classifier (SVC) is trained with a linear kernel and the regularisation constant C = 1.0. The Stochastic Gradient Descent (SGD) classifier uses hinge loss with L2 penalty, $\alpha = 1 \times 10^{-4}$, a maximum of 2000 iterations, and tolerance 1×10^{-3}; this configuration approximates a linear-SVM objective optimised by SGD. The Random Forest (RF) classifier uses 300 estimators with unrestricted depth and is used as the third traditional baseline in the binary

setting, following Table 4 of Aksoy (2025). In the three-class setting, SVC and SGD are retained as the two most competitive classical baselines consistent with Table 3 of Aksoy (2025).

The deep-learning baseline is a hybrid architecture that combines two branches. The first branch applies a Keras TextVectorization layer followed by a learned embedding of size 128, a one-dimensional convolution (Conv1D) with 64 filters, and two stacked Bidirectional LSTM (BiLSTM) layers. The second branch applies the pre-trained Universal Sentence Encoder (USE) layer from TensorFlow Hub (trainable = False), followed by a dense projection to 256 units and BatchNormalization. The two branches are concatenated and passed through two dropout-regularised dense layers before a sigmoid output (binary) or a softmax output of size three (multiclass). Training uses the Adam optimiser with an initial learning rate of 3×10^{-4}, a batch size of 32, up to 20 epochs, EarlyStopping on validation loss, and a LearningRateScheduler with plateau-based reduction.

The objective of Stage 1 is not to introduce any new modelling choice but to obtain numerically comparable baselines on the same split, so that later stages can be interpreted as incremental contributions over a verified reference.

3.2.4 Stage 2: fine-tuning of the proposed model

Stage 2 introduces the proposed model. This backbone was selected over alternative encoder-only Transformers because, on the GLUE benchmark, DeBERTa-v3 establishes the strongest published results among models in its size class, outperforming BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and the original DeBERTa (He et al., 2021) with an average difference of 1.3% (He et al., 2023). DeBERTa-v3-Large has 304 million parameters organised into 24 Transformer encoder layers with a hidden size of 1024 and 16 attention heads. Beyond raw benchmark performance, two architectural properties of DeBERTa matter for this task. First, the mechanism of DeBERTa (He et al., 2021), detailed in section 2.4. This is useful for mental-health text, where polysemous tokens such as 'fine', 'tired', or 'end' frequently change polarity with context. Second,

DeBERTa-v3 uses an ELECTRA-style replaced-token-detection objective with gradient-disentangled embedding sharing, which improves sample efficiency during pre-training and yields stronger downstream representations on classification tasks (He et al., 2023). The Large variant was preferred over the Base variant because larger encoders tend to perform better on minority and semantically subtle classes in clinical-NLP settings (Ji et al., 2022), which matches the structure of errors observed on this corpus.

To improve over the Aksoy (2025) baseline, the following fine-tuning strategies were applied:

- Text chunking with a sliding window: Reddit-style posts frequently exceed the 512-token limit of standard Transformer encoders (Devlin et al., 2019). So that long posts are used in full rather than cut off, each training and evaluation example is tokenised with return_overflowing_tokens = True, max_length = 384, and stride = 192 — the sliding-window chunking that is standard when BERT-family models are fine-tuned on long text (Pappagari et al., 2019; Sun et al., 2019). This produces overlapping chunks that jointly cover the entire post. Each chunk is forwarded through DeBERTa independently during training and inference. At inference time, chunk-level logits that share the same example identifier are averaged, and the class with the highest averaged logit is taken as the post-level prediction; this mean-pooling aggregation over chunk logits corresponds to the strategy reported by Sun et al. (2019) as the most stable among simple aggregation schemes.

- Layer freezing: To preserve the linguistic knowledge acquired during pre-training and to reduce the number of trainable parameters, the input embedding layer and the bottom 16 of the 24 encoder layers of DeBERTa-v3-Large are frozen. Only the top 8 encoder layers together with the pooler and the classification head remain trainable. This design follows the common practice of partial fine-tuning (Howard & Ruder, 2018), and empirical trials in our preliminary experiments confirmed that freezing the bottom two-thirds of the encoder stabilised optimisation and reduced GPU memory pressure without harming generalisation.

- Layer-wise Learning Rate Decay (LLRD): Within the trainable portion of the model, a Layer-wise Learning Rate Decay schedule is applied. Let η_0 denote the base learning rate assigned to the classification head; encoder layer i (indexed from top to bottom) then receives a learning rate of $\eta_0 \cdot \gamma^{(L-i)}$, where L is the index of the top-most trainable layer and $\gamma = 0.95$ is the decay factor. This encodes the prior that layers closer to the classification head require stronger task-specific adaptation, whereas deeper (more general) layers should be perturbed less aggressively. LLRD was originally proposed in the context of ULMFiT (Howard & Ruder, 2018) and is now standard practice in Transformer fine-tuning.

- R-Drop regularisation: R-Drop (Liang et al., 2021) is applied during training to reduce the inconsistency that stock dropout introduces between stochastic sub-networks. For each mini-batch, two forward passes are performed with independent dropout masks, yielding two logit tensors. The final training loss combines the mean of the two cross-entropy losses with a symmetric Kullback–Leibler divergence between the two softmax distributions, weighted by a scalar α. The objective can be written as;

$$L = 0.5 \cdot (\mathrm{CE}(p_1, y) + \mathrm{CE}(p_2, y)) + \alpha \cdot \frac{1}{2} \cdot (\mathrm{KL}(p_1 \parallel p_2) + \mathrm{KL}(p_2 \parallel p_1))$$

  with α set to 0.5 in the main experiments.

- Class-Balanced Loss Weighting. To address the inter-class imbalance, class weights (β=0.999) calculated based on the effective number of samples of each class were integrated into the cross-entropy loss function (Cui et al., 2019).The training hyperparameters of the model are summarized in Table 3.7.

Table 3.7 Training hyperparameters of the proposed DeBERTa-v3-Large model

Hyperparameter	Value
Backbone	microsoft/deberta-v3-large
Tokenizer	DeBERTa-v3 SentencePiece
Max length / stride (chunking)	384 tokens/ 192 tokens
Frozen layers	Embeddings + first 16 encoder layers
Optimiser	AdamW
Base learning rate (η_0)	2e-5
LLRD decay (γ)	0.95
Warm-up ratio	0.10 (linear scheduler)
Weight decay	0.01
R-Drop (α)	0.50
Class-balanced loss β	0.999
Per-device batch size	8
Epochs / early stop patience	up to 10 / 3
Early Stopping Metric	Validation Macro F1-Score

3.2.5 Stage 3: model validation and evaluation

To test the stability of the proposed model and the accuracy of the hyperparameter selections, hyperparameter optimization and 5-fold cross-validation processes were carried out extensively.

- Hyperparameter Optimization (HPO): To assess whether the baseline configuration could be improved further, a systematic Bayesian hyperparameter optimization (HPO) study was conducted using the Optuna framework (Akiba et al., 2019). Sampling was handled by a Tree-structured Parzen Estimator across ten trials, each capped at six epochs so the overall GPU budget stayed manageable. The best configuration is then retrained from scratch for 10 epochs on the full training set and

evaluated on the held-out test set. The HPO protocol executed on an NVIDIA A100 GPU (total wall time: ~ 11 hours). The search space is presented in Table 3.8.

Table 3.8 Optuna HPO search space

Hyperparameter	**Search space**
frozen_layers	{0, 4, 8, 12, 16, 20}
base_lr	$[1 \times 10^{-5}, 5 \times 10^{-5}]$ (log-uniform)
lr_decay (γ)	[0.85, 0.99]
rdrop_alpha (α)	[0.0, 1.0]
label_smoothing	{0.0, 0.05, 0.10}
warmup_ratio	[0.05, 0.20]

- Five-fold cross-validation and ensembling: To gauge how stable the model is from one split to another, a five-fold cross-validation was set up along the lines of the low-bias estimation framework of Kohavi (1995). Prior to cross-validation, a hold-out test set comprising 10% of the complete dataset was isolated using a fixed random state and reserved for the final ensemble evaluation. The remaining 90% of the data was subjected to a stratified 5-fold splitting strategy, as shown in Figure 3.2, ensuring that the original class distribution (Depression, Normal, Suicidal) was preserved across all folds.

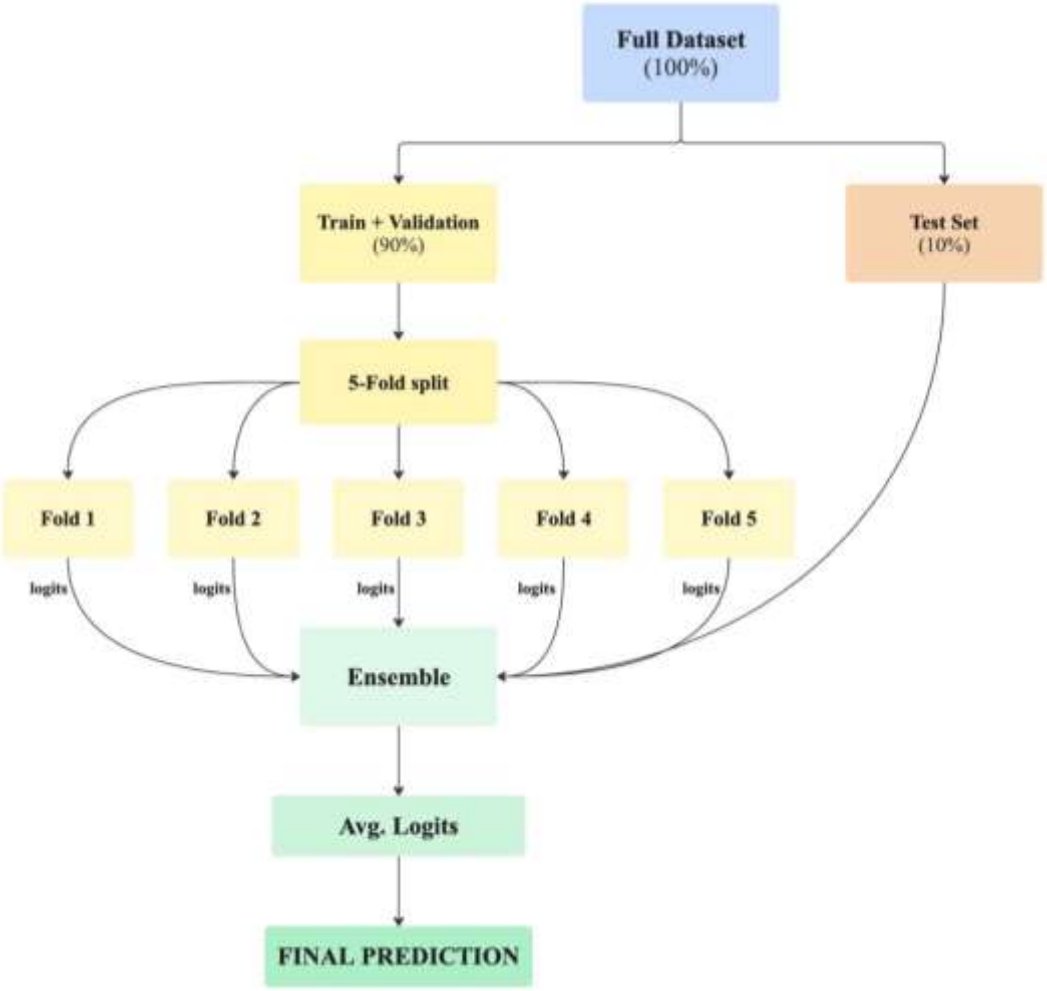

Figure 3.2 Workflow diagram of Five-fold cross-validation

In every fold the cross-validation data is split 80/20, with a rotating fifth serving as the validation set (Table 3.9).

Table 3.9 Stratified 5-Fold split scheme

	Part 1	Part 2	Part 3	Part 4	Part 5
Fold 1	Validation	Training	Training	Training	Training
Fold 2	Training	Validation	Training	Training	Training
Fold 3	Training	Training	Validation	Training	Training
Fold 4	Training	Training	Training	Validation	Training
Fold 5	Training	Training	Training	Training	Validation

At the beginning of each fold, an untrained DeBERTa-v3-Large model (He et al., 2023) is loaded. As illustrated in Figure 3.3, the fold's data is divided into training and validation subsets, both of which are tokenized using the DeBERTa tokenizer. The model is then trained for 10 epochs using the training data, while validation macro-F1 is monitored for early stopping.

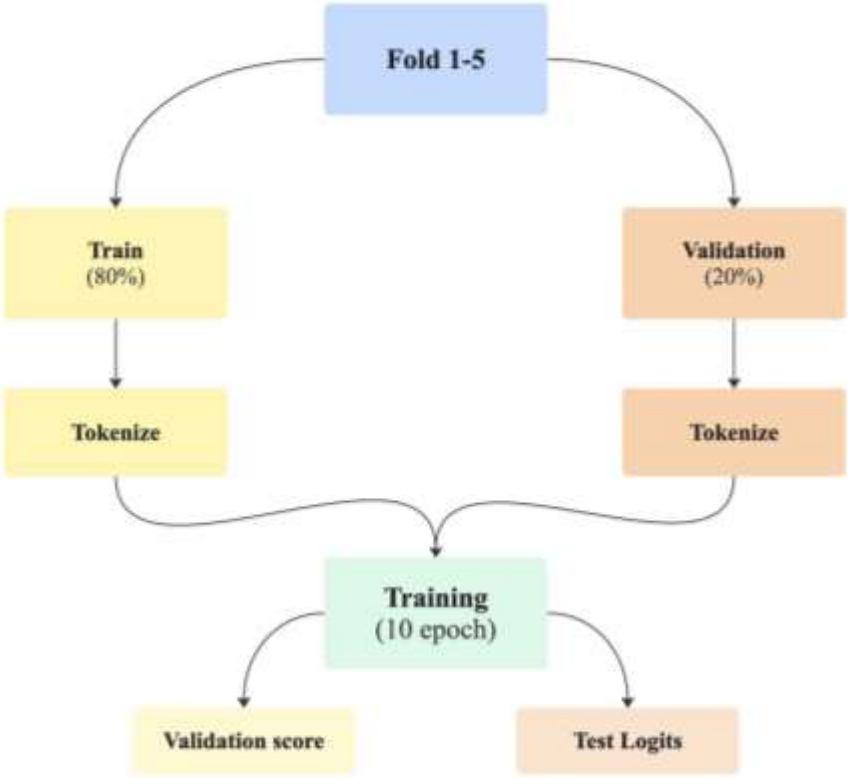

Figure 3.3 Internal workflow of a single fold

Once training is complete, the trained model make predictions on the fixed test set and produces a 3-dimensional logit vector for each test example, corresponding to the three classes (Depression, Normal, Suicidal).

For example, an output vector [2.8, 0.5, 1.1] indicates raw class scores for Depression, Normal, and Suicidal respectively. After these scores are saved, the model is released from GPU memory, and a new model is loaded for the next fold. Once all 5 folds are complete, there are 5 separate 3-dimensional logit vectors from 5 different models for each test example.

Figure 3.4 illustrates the process of averaging these five vectors element-wise into one mean vector. This technique mirrors the sum-rule combination scheme, which Kittler et al. (1998) found to be highly resilient against realistic estimation noise. The step also reflects Dietterich's (2000) principle that combining different predictors successfully lowers variance. To get the final prediction, an argmax operation is used on the averaged vector.

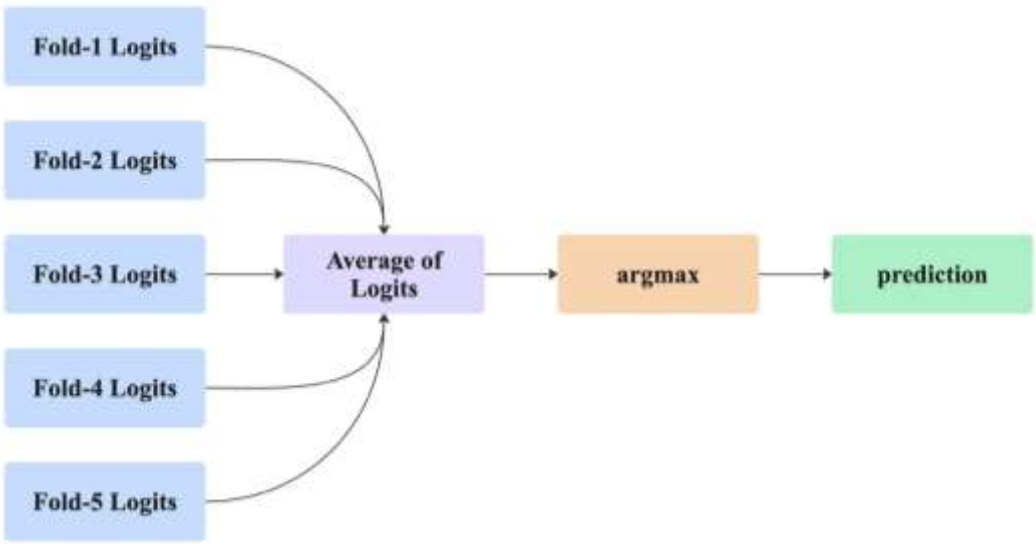

Figure 3.4 The ensemble prediction pipeline

Final prediction is mathematically formulated as follows:

$$\text{Ensemble Logits} = \frac{1}{5}\sum_{k=1}^{5} Logits_k$$

$$\hat{y} = \text{argmax (Ensemble Logits)}$$

Finally, the ensemble predictions are compared with the actual labels, and the final metrics are calculated. Because evaluation is spread over several partitions, this stratified k-fold scheme (Kohavi, 1995) keeps the reported performance from hinging on any one random split.

3.2.6 Evaluation metrics

Four fundamental metrics, widely accepted as standard in the classification literature, were used to evaluate the performance of the models. The formal definitions of these metrics follow the systematic framework presented by Sokolova and Lapalme (2009). In the multi-class setting with l classes, these counts are computed individually for each class C_i.

- Accuracy: Accuracy reports the fraction of all samples that the model labels correctly (Sokolova & Lapalme, 2009). For a multi-class problem with l classes and N test samples it is computed as the ratio of correctly classified samples to the total number of samples:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{l} \text{tp}_i = \frac{\text{number of correctly classified samples}}{\text{total number of samples}}$$

  Useful as a summary, accuracy can mislead when the classes are imbalanced, since the majority class tends to drive it

- Precision: Precision reflects how well a classifier's positive predictions match the true labels (Sokolova & Lapalme, 2009).

$$Precision_i = \frac{tp_i}{tp_i + fp_i}$$

  When precision is high, false positives are rare: the model seldom pins a positive label on the wrong item.

- Recall: Recall captures how thoroughly a classifier finds every positive instance of a class (Sokolova & Lapalme, 2009).

$$Recall_i = \frac{tp_i}{tp_i + fn_i}$$

- F1-Score: The F1-score folds precision, and recall into one figure through their harmonic mean, balancing the two (Sokolova & Lapalme, 2009).

$$\mathrm{F1}_i = \frac{2 \times \mathrm{Precision}_i \times \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

- To obtain a single aggregate score across all classes, two averaging strategies are commonly employed. Under macro-averaging the metric is computed separately for each class and then averaged without weighting, so that every class contributes equally regardless of its size:

$$\mathrm{Precision}_M = \frac{1}{l}\sum_{i=1}^{l} \mathrm{Precision}_i, \qquad \mathrm{Recall}_M = \frac{1}{l}\sum_{i=1}^{l} \mathrm{Recall}_i$$

$$\mathrm{F1}_M = \frac{1}{l}\sum_{i=1}^{l} \mathrm{F1}_i$$

As Sokolova and Lapalme (2009) note, macro-averaging treats all classes equally, whereas micro-averaging favors larger classes. Given the inherent class imbalance in the mental health dataset used in this study, the Macro F1-Score was adopted as the primary evaluation metric to ensure that the model's performance on minority classes (e.g., Suicidal) is given equal consideration alongside majority classes. This choice aligns with the evaluation strategy employed in the reference study (Aksoy, 2025).

3.2.7 Explainability analysis

The proposed DeBERTa-v3-Large model is evaluated for predictive performance in Sections 4.1 to 4.6. To complement these aggregate metrics with token-level evidence of how the model arrives at its decisions, an explainability analysis is added. The method of gradient based is applied.

The gradient based method used here is Integrated Gradients (Sundararajan et al., 2017), implemented with Captum: an axiomatic attribution technique that quantifies how much each input feature adds to a model's prediction. For an input x, a baseline x', and a model F, the attribution of the i-th feature is the integral of the gradient of Falong the straight-line path from $x'$to x:

$$IG_i(x) = (x_i - x'_i) \cdot \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha$$

In practice the integral is approximated by a Riemann sum over m equally spaced points along the path:

$$\mathrm{IG}_i^{\mathrm{approx}}(x) = (x_i - x'_i) \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m}(x - x')\right)}{\partial x_i}$$

with $m = 200$ integration steps in this study. Intuitively, the formula asks the following question for each token: as the input is gradually transformed from a neutral baseline to the actual post, how much does this particular token push the model's output toward the target class? A larger positive value means the token supports the target class, and a larger negative value means the token opposes it. The implementation in this thesis uses the LayerIntegratedGradients class of the Captum library (Kokhlikyan et al., 2020) at the embedding layer of DeBERTa-v3-Large, with a baseline that replaces only the content tokens with the [PAD] token while keeping [CLS] and [SEP] in their original positions.

Its completeness axiom requires that the per-feature attributions sum exactly to the gap between the model's output on the real input and its output on the baseline. The convergence delta δ measures the numerical error of the Riemann approximation that is used to compute the integral in practice:

$$\delta = \left| \sum_i IG_i(x) - (F(x) - F(x')) \right|$$

A value of δ close to zero indicates that the attribution scores faithfully account for the model's decision, while larger values indicate that more integration steps would be required for a faithful approximation. In the present study, the empirically observed $|\delta|$ is below 0.1 for nine of the ten reported examples, which is the standard acceptability threshold reported in the original paper.

The token-level attribution scores reported in Section 4.7 are visualised as red–blue diverging heat-maps; the colour scale used for these heat-maps is shown in Figure 3.5.

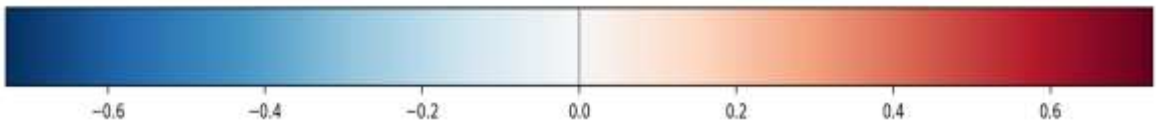

Figure 3.5 Heatmap colour band

Posts are selected from the qualitative error analysis in Section 4.6: five cases where DeBERTa is correct and the reproduced SVC is wrong (Table 4.14), and four cases with the opposite outcome (Table 4.15). Each post is located in the test set by substring matching and verified against its true label. For token-level reporting, special tokens ([CLS], [SEP], [PAD]) and standalone punctuation are excluded from attribution rankings because they do not represent post content. All experiments use the same checkpoint and tokenizer as Section 4.2 with a maximum sequence length of 384. Heat maps are generated with Matplotlib, while full attribution scores, example-level statistics, and aggregate token rankings are provided in the accompanying material.

4. RESULTS AND DISCUSSION

This section presents the results of the experiments described in Chapter 3 and discusses what they imply. Section 4.1 reports the reproduction of the baseline models used in the reference study (Aksoy, 2025). Section 4.2 reports the test-set performance of the proposed DeBERTa-v3-Large model. Section 4.3 places all models in a single comparative table. Sections 4.4 and 4.5 cover the validation experiments: hyperparameter optimisation and five-fold cross-validation. Section 4.6 contains the qualitative error analysis, Section 4.7 reports the token-level explainability analysis, and Section 4.8 closes with a discussion of mechanisms.

4.1 Reference (Baseline) Model Results

The classical ML and hybrid DL baselines from Aksoy (2025) were re-implemented under the same 80/10/10 stratified split and re-trained on the data partition described in Section 3.1.2. Tables 4.1 and 4.3 compare the test-set scores obtained in this thesis with those reported in the reference study.

In the three-class setting, reproduced models fall within 0.6 to 0.8 accuracy points of the published numbers. This margin is consistent with the run-to-run variance expected on a 42,399-row corpus when a single random seed is fixed, so the reproduction is therefore considered successful.

Table 4.1 Three-class baseline results: this study vs. Aksoy (2025)

Model	**Accuracy (this study)**	**Accuracy (Aksoy, 2025)**	**Macro F1 (this study)**	**Macro F1 (Aksoy, 2025)**	**Weighted-F1 (this study)**	**Weighted-F1 (Aksoy, 2025)**
SVC	0.8075	0.8140	0.7899	0.79	0.8061	0.81
SGD	0.8014	0.8062	0.7788	0.78	0.7966	0.81
DL	0.7884	0.7963	0.7704	0.78	0.7863	0.80

Table 4.2 SVC per-class performance on the three-class test set

Class	**Precision**	**Recall**	**F1-score**	**Overall Performance**
Normal	0.92	0.94	0.93	**Accuracy:** 0.81
Depression	0.76	0.77	0.77	**Macro-F1:** 0.79
Suicidal	0.69	0.66	0.67	**Weighted-F1:** 0.81
				Test samples: 4200

The three-class confusion matrices of the reproduced baselines are shown in Figures 4.1 to 4.3 SVC achieves the strongest balance among the three models, with 1543 of 1634 Normal posts (94.4%), 1181 of 1541 Depression posts (76.6%), and 700 of 1065 Suicidal posts (65.7%) classified correctly. The Suicidal class is the weakest in all three baselines: SVC misclassifies 301 Suicidal posts as Depression and 64 as Normal; SGD pushes an even larger fraction (333) to Depression; the USE+BiLSTM hybrid produces a different but equally pronounced confusion in the opposite direction, sending 398 Depression posts to the Suicidal column. This is the empirical gap that the proposed model is designed to close.

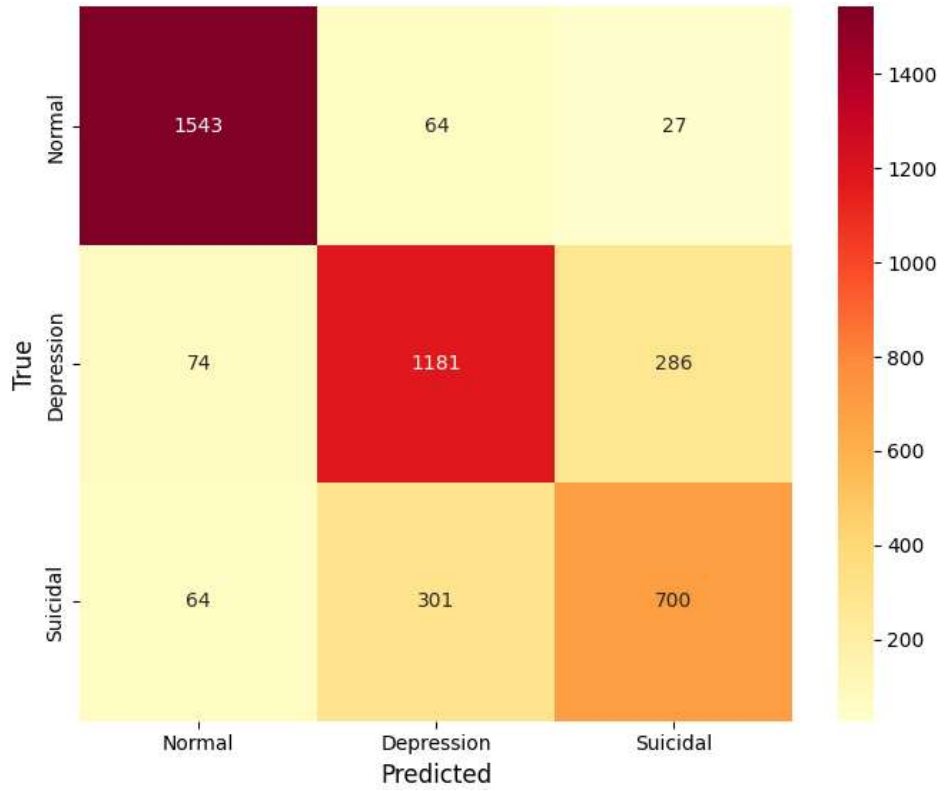

Figure 4.1 Confusion matrix of the SVC (TF-IDF) baseline, three-class test set

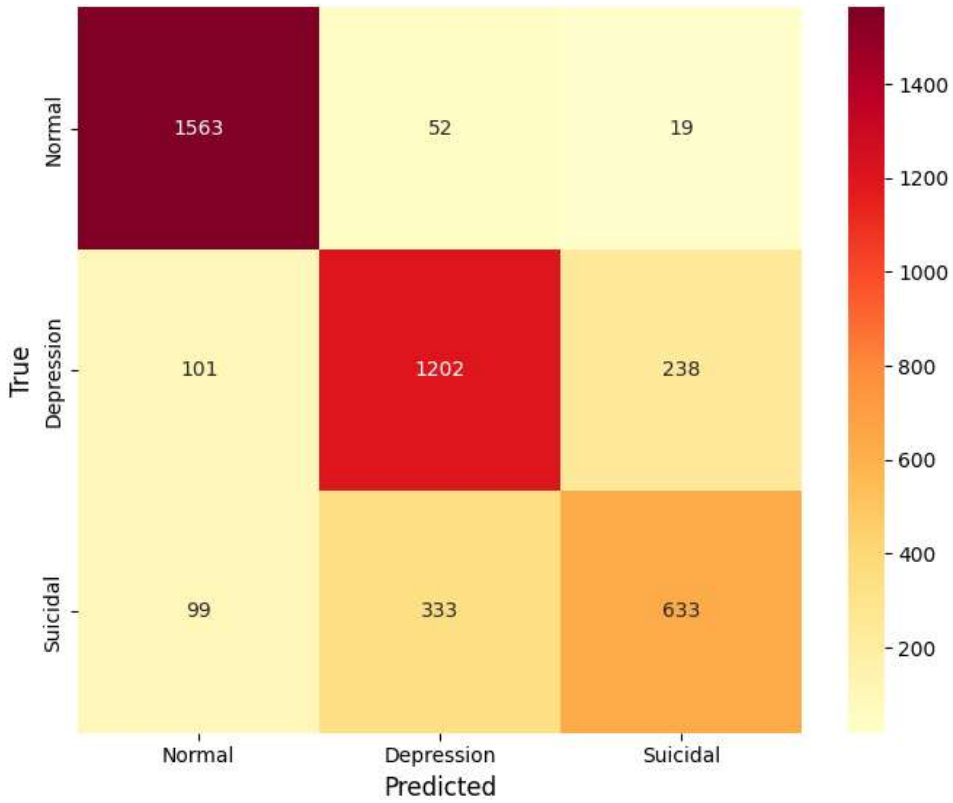

Figure 4.2 Confusion matrix of the SGD (TF-IDF) baseline, three-class test set

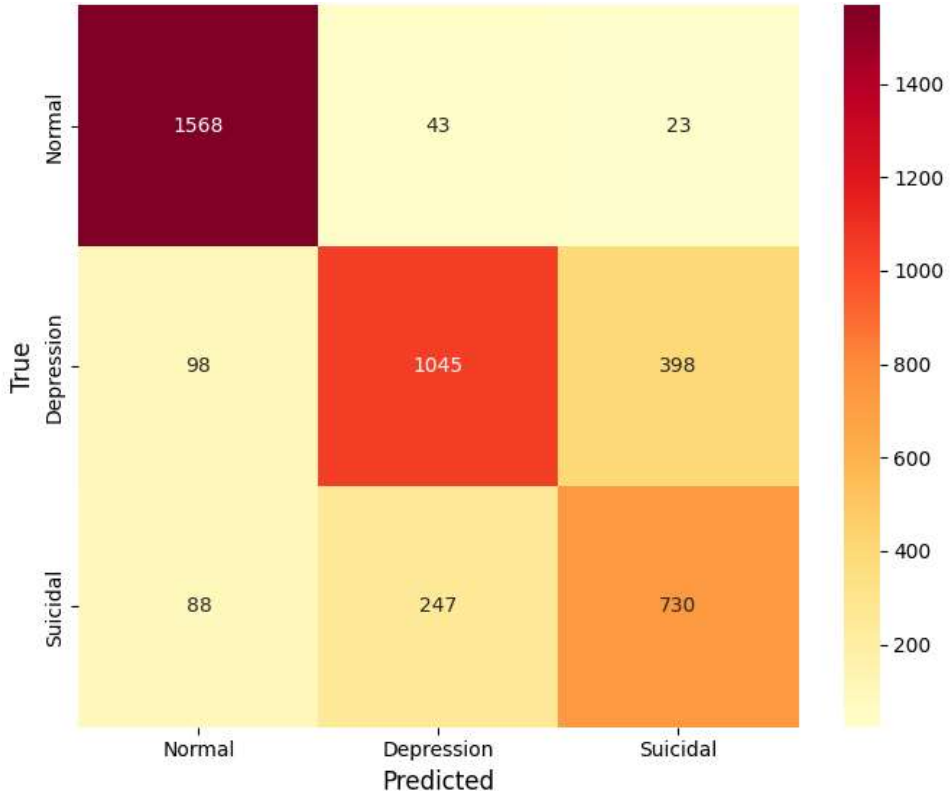

Figure 4.3 Confusion matrix of the DL baseline (USE + BiLSTM), three-class test set

In the binary setting, all reproduced classifiers exceed 0.92 accuracy. SVC and the deep-learning hybrid are practically tied at about 0.946 accuracy and 0.944 macro F1, with Random Forest about 1.8 accuracy points behind. The binary task is much easier than the three-class task because the two pathological classes that are linguistically hard to separate are merged into a single label, removing the dominant source of error.

Table 4.3 Binary baseline results: this study vs. Aksoy (2025)

Model	**Accuracy (this study)**	**Accuracy (Aksoy, 2025)**	**Macro F1 (this study)**	**Macro F1 (Aksoy, 2025)**
SVC (TF-IDF)	0.9455	0.9533	0.9424	0.95
Random Forest (TF-IDF)	0.9278	0.9368	0.9240	0.93
DL (USE + BiLSTM)	0.9469	0.9495	0.9443	0.95

The binary confusion matrices (Figures 4.4 to 4.6) reflect the same picture. SVC misclassifies 120 Normal posts as Depression/Suicidal and 110 Depression/Suicidal posts as Normal, an almost symmetric error structure of 230 misclassifications. The deep-learning hybrid produces a slightly tighter pattern (101 + 120 = 221 errors), and Random Forest the largest of the three (145 + 162 = 307 errors). Across the three baselines the errors are symmetric between the two classes, which means the binary task is well-posed for classical classifiers and leaves little headroom for further improvement at this representational level.

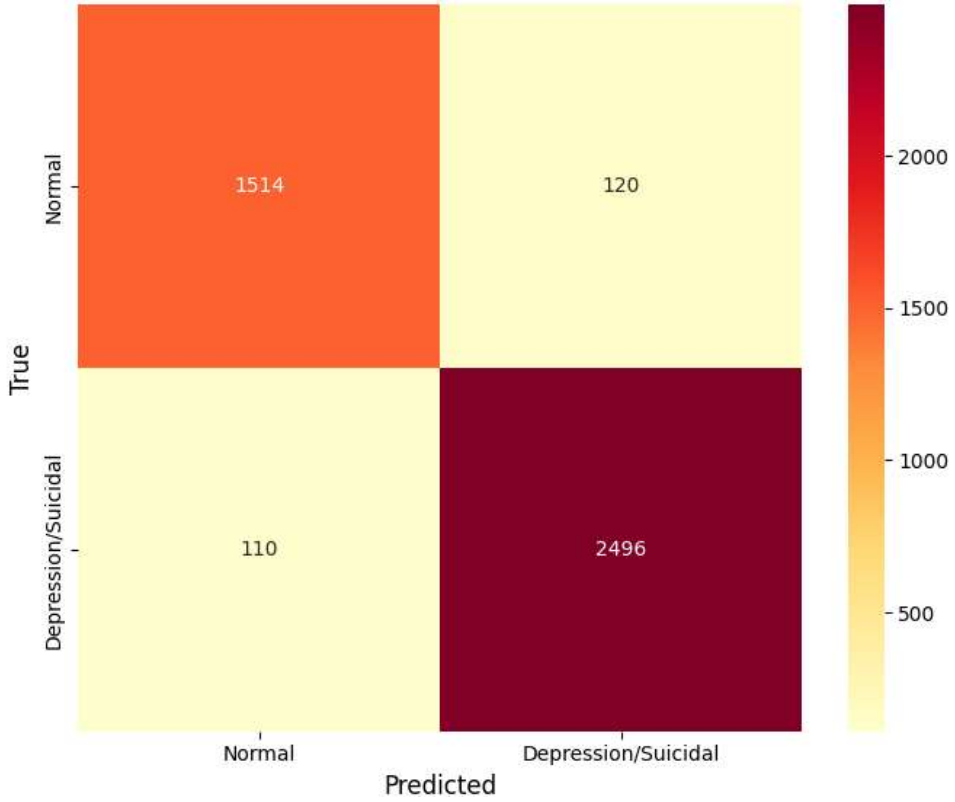

Figure 4.4 Confusion matrix of the SVC (TF-IDF) baseline, binary test set

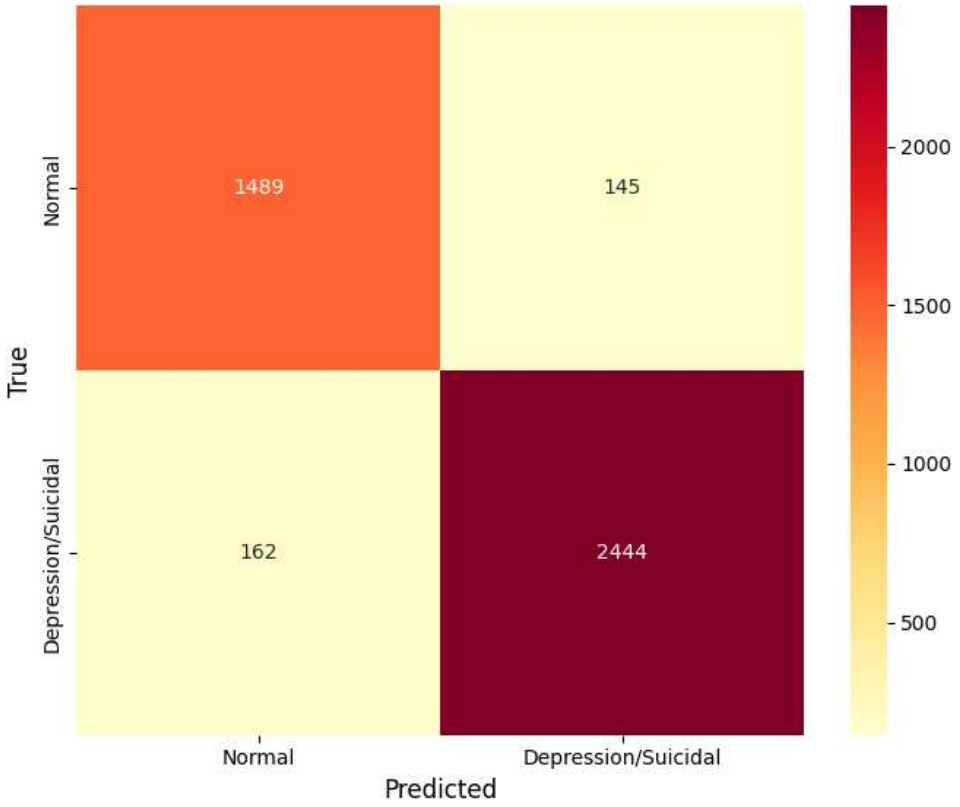

Figure 4.5 Confusion matrix of the Random Forest (TF-IDF) baseline, binary test set

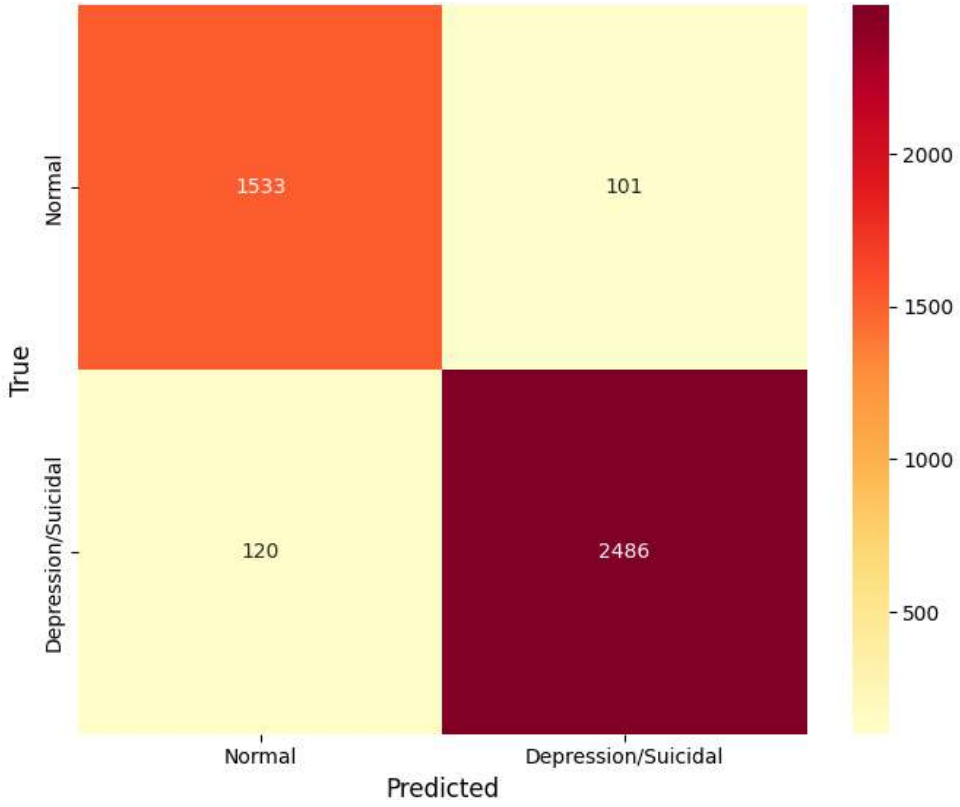

Figure 4.6 Confusion matrix of the DL baseline (USE + BiLSTM), binary test set

4.2 DeBERTa-v3-Large Results

The proposed DeBERTa-v3-Large model was trained on the deduplicated corpus described in Section 3.1.2 with the configuration listed in Table 3.7.

4.2.1 Three-class setting

On the three-class test set the proposed model reaches an accuracy of 0.8835 and a macro F1 of 0.8712. The losses are shown in Figure 4.7. Both curves drop quickly during the first 1500 optimisation steps. After this point the validation loss stabilises near 0.33 while the training loss continues to fall slowly to about 0.20. The gap between the two curves is narrow and does not widen, which indicates that the combination of layer freezing, LLRD and R-Drop (Liang et al., 2021) controls overfitting effectively. Validation accuracy and validation macro F1 (Figure 4.8) reach a plateau around step 3000; further training adds less than half an F1 point. EarlyStopping selects a checkpoint near this plateau on the validation macro F1.

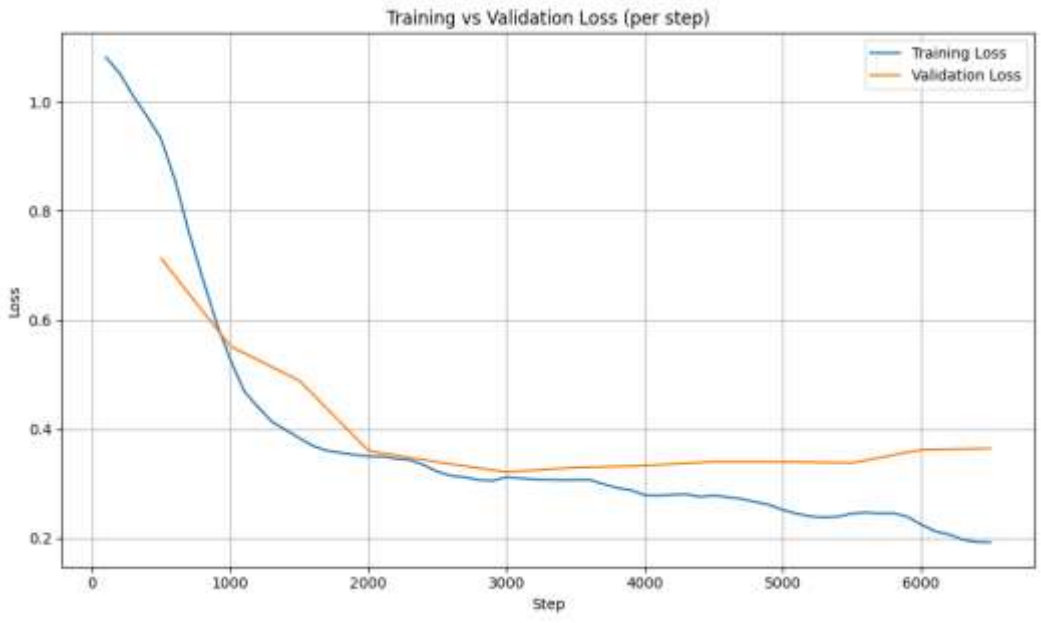

Figure 4.7 Training and validation loss across optimisation steps, three-class

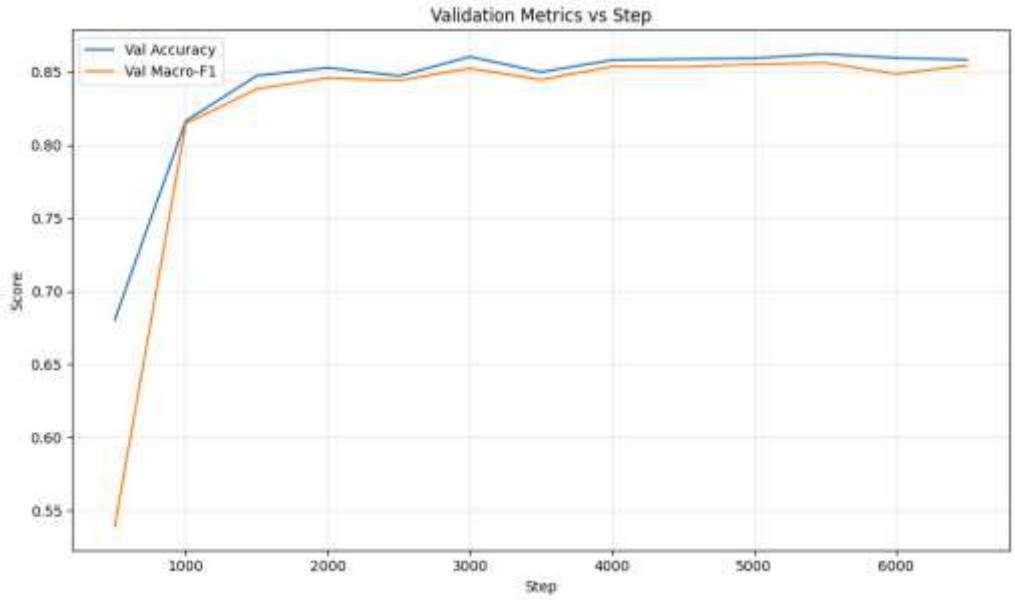

Figure 4.8 Validation accuracy and macro F1 across training steps, three-class

The example-level test confusion matrix is shown in Figure 4.9. Of 1604 Normal test posts the model classifies 1594 correctly and confuses only 10 with the two pathological classes. Of 1509 Depression posts, 1257 are classified correctly and 237 are predicted as Suicidal. Of 1064 Suicidal posts, 839 are classified correctly and 212 are predicted as Depression. The dominant residual error is therefore the Depression to Suicidal confusion: 237 + 212 = 449 of the 487 total errors lie on the off-diagonal between these two clinical classes.

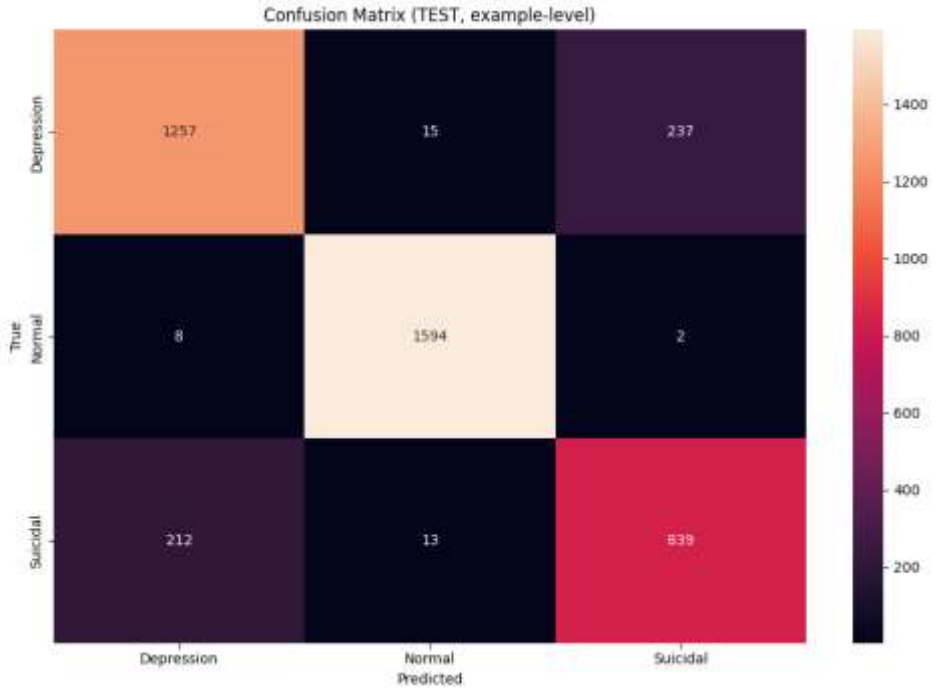

Figure 4.9 Example-level test confusion matrix of DeBERTa-v3-Large, three-class

Table 4.4 DeBERTa-v3-Large per-class performance on the three-class test set

Class	**Precision**	**Recall**	**F1-score**	**Overall Performance**
Depression	0.851	0.833	0.842	**Accuracy:** 0.8835
Normal	0.983	0.994	0.988	**Macro-F1:** 0.87
Suicidal	0.778	0.789	0.783	**Weighted-F1:** 0.88
				Test samples: 4188

4.2.2 Binary setting

The same model trained for the binary task reaches an accuracy of 0.9897 and a macro F1 of 0.9892 on the held-out evaluation set of 4177 examples. The training and validation losses (Figure 4.10) decrease monotonically with a small and almost constant gap between the two curves; validation accuracy and macro F1 (Figure 4.11) cross 0.99 by step 4000 and remain there for the rest of training. The example-level confusion matrix (Figure 4.12) shows 2544 of 2573 Depression/Suicidal posts and 1590 of 1604 Normal

posts classified correctly, which corresponds to a total of only 43 misclassified examples. Per-class metrics are reported in Table 4.5.

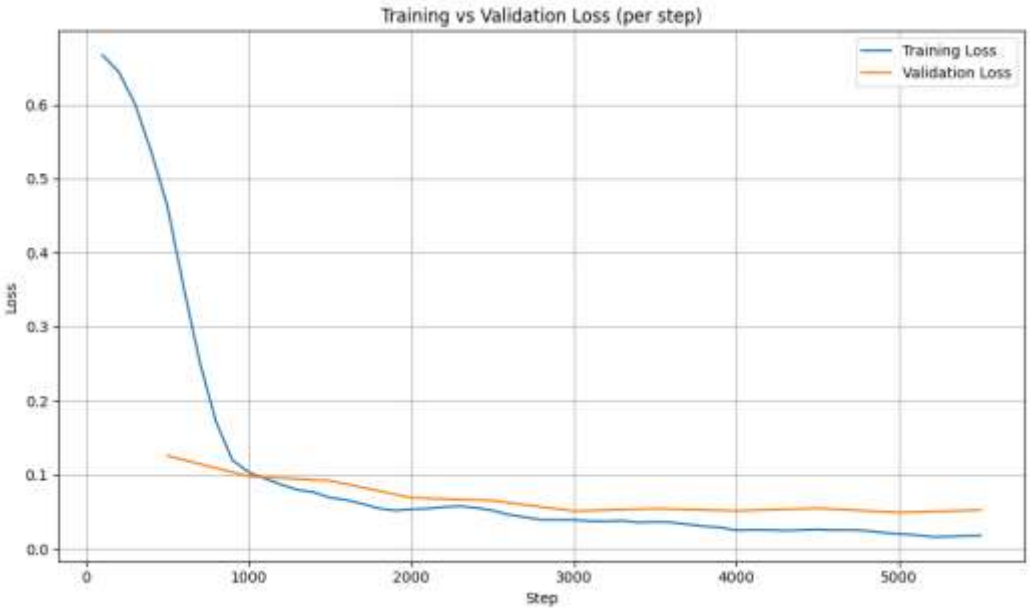

Figure 4.10 Training and validation loss across optimisation steps, binary-class

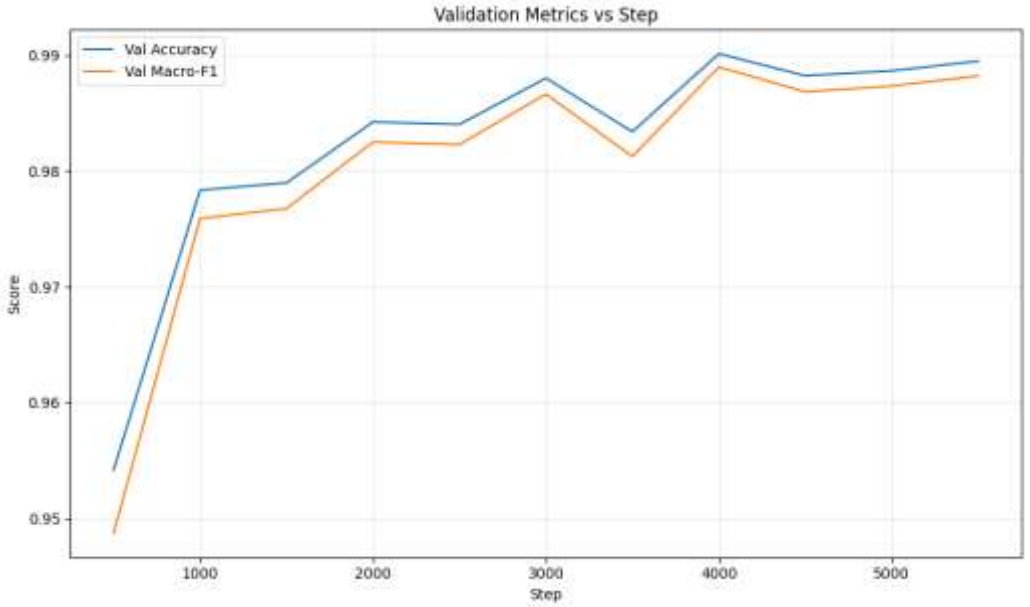

Figure 4.11 Validation accuracy and macro F1 across training, binary-class

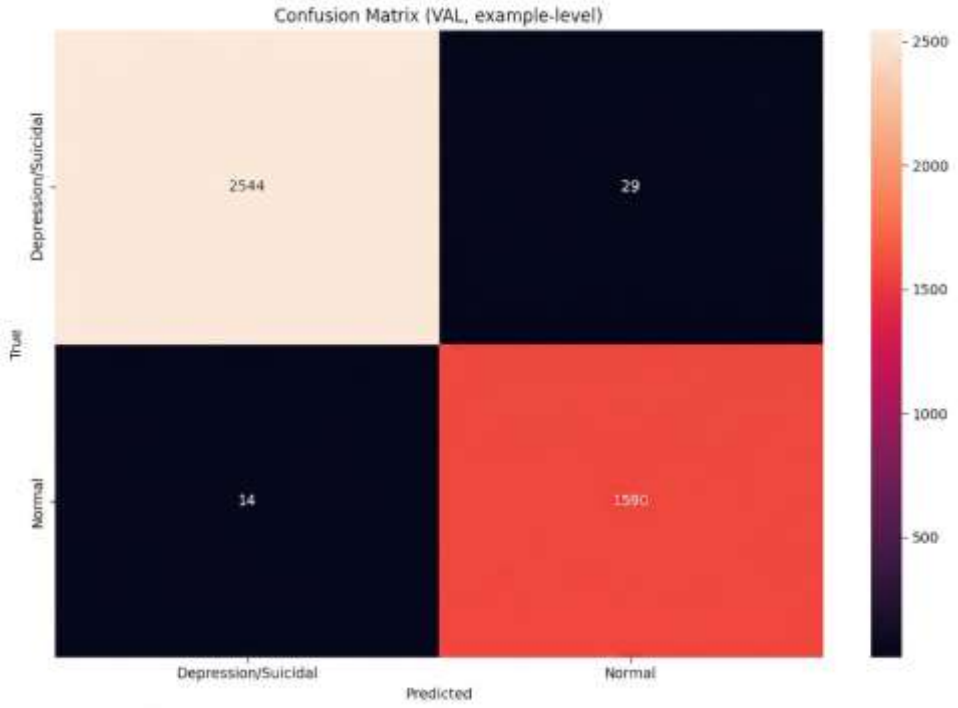

Figure 4.12 Example-level confusion matrix of DeBERTa-v3-Large, binary-class

Table 4.5 DeBERTa-v3-Large per-class performance, binary setting

Class	**Precision**	**Recall**	**F1-score**	**Overall Performance**
Depression/Suicidal	0.99	0.99	0.99	**Accuracy:** 0.99
Normal	0.98	0.99	0.98	**Macro-F1:** 0.99
				Weighted-F1: 0.99
				Test samples: 4177

4.3 Comparative Performance Analysis

Tables 4.6 and 4.7 place the proposed DeBERTa-v3-Large model alongside the baseline scores reported by Aksoy (2025), for the three-class and binary tasks respectively. The Improvement column is computed as the difference between the proposed model and each baseline on the same metric (Accuracy, Macro F1, and Weighted F1 where reported); the row corresponding to the proposed model is therefore left empty.

In the three-class task (Table 4.6), the proposed model exceeds the strongest reference baseline (SVC) by 7.0 accuracy points, 8 macro F1 points, and 7 weighted F1 points. The improvement against the SGD and the USE+BiLSTM baselines is larger, reaching 8.7 accuracy points against the deep-learning hybrid. In the binary task (Table 4.7), the absolute gain is smaller because the reference baselines are already close to or above 0.95 accuracy, but DeBERTa still adds 3.6 accuracy points and 4 macro F1 points over the strongest reference baseline (SVC), 5.3 accuracy points and 6 macro F1 points over Random Forest, and 4.0 / 4 points over the deep-learning hybrid.

Table 4.6 Comparative three-class test-set results

Model	**Accuracy**	**Macro avg. F1**	**Weighted avg. F1**	**Improvement**
SVC (TF-IDF)	0.8140	0.79	0.81	+0.0695 / +0.08 / +0.07
SGD (TF-IDF)	0.8062	0.78	0.81	+0.0773 / +0.09 / +0.07
DL (USE + BiLSTM)	0.7964	0.78	0.80	+0.0871 / +0.09 / +0.08
DeBERTa-v3-Large (proposed)	0.8835	0.87	0.88	

Table 4.7 Comparative binary results

Model	**Accuracy**	**Macro avg. F1**	**Improvement**
SVC (TF-IDF)	0.9533	0.95	+0.0364 / +0.04
Random Forest (TF-IDF)	0.9368	0.93	+0.0529 / +0.06
DL (USE + BiLSTM)	0.9495	0.95	+0.0402 / +0.04
DeBERTa-v3-Large (proposed)	0.9897	0.99	

The class-wise picture explains where the gain comes from. In the three-class setting the per-class F1 of the SVC baseline reproduced under the same protocol as Aksoy (2025) is 0.93 / 0.77 / 0.67 for Normal, Depression and Suicidal, while the proposed model reaches 0.99 / 0.84 / 0.78 on the same test posts. The largest absolute improvement is therefore on the Suicidal class (+11 F1 points), followed by Depression (+7), with Normal already close to ceiling for both models. In the binary setting both classes improve by about 4 F1 points. The proposed model is not only more accurate on aggregate; it concentrates its improvement on the class where the reference model is weakest and where the clinical stakes are highest.

4.4 Hyperparameter Optimisation Results

Ten Optuna trials were executed on the three-class task under the search space defined in Table 3.8. Table 4.8 lists the parameters and validation macro F1 of each trial. All ten configurations land within a 0.7 F1 point band on the validation set (0.864 to 0.871), which indicates that the proposed model is not highly sensitive to the search dimensions within the explored ranges. The best trial configuration is shown in Table 4.9.

Table 4.8 Optuna trials on the three-class task (10 trials, validation macro F1)

Trial	Frozen	base_lr	lr_decay	rdrop_α	LS	warmup	Val F1
0	4	1.10×10^{-5}	0.971	0.60	0.10	0.17	0.8669
1	16	1.60×10^{-5}	0.936	0.14	0.10	0.17	0.8640
2	16	1.11×10^{-5}	0.983	0.97	0.00	0.15	0.8672
3	16	2.90×10^{-5}	0.894	0.52	0.10	0.17	0.8689
4	0	1.08×10^{-5}	0.896	0.39	0.05	0.09	0.8689
5	16	1.38×10^{-5}	0.851	0.82	0.10	0.06	0.8693
6	8	1.65×10^{-5}	0.896	0.73	0.05	0.07	0.8671
7	12	1.99×10^{-5}	0.854	0.11	0.05	0.13	0.8688
8	0	1.59×10^{-5}	0.873	0.93	0.10	0.17	0.8710
9	16	1.19×10^{-5}	0.882	0.43	0.05	0.13	0.8674

Table 4.9 Best Optuna trial (Trial 8) configuration

Hyperparameter	Best value
frozen_layers	0
base_lr	1.594×10^{-5}
lr_decay (γ)	0.8726
rdrop_alpha (α)	0.9297
label_smoothing	0.10
warmup_ratio	0.1706
Validation macro F1	0.8710

Optuna also reports the relative importance of each hyperparameter in explaining the variance of the validation objective across trials. Figure 4.13 shows these scores. The single most influential coordinate is lr_decay (importance 0.36), followed by rdrop_alpha (0.24) and warmup_ratio (0.20). frozen_layers, base_lr and label_smoothing each contribute less than 0.12 of the explained variance. The ordering matches the prior on partial fine-tuning of DeBERTa: once a sensible learning-rate range is fixed, the schedule shape and the regularisation strength matter more than the exact number of frozen layers (Howard & Ruder, 2018; He et al., 2021).

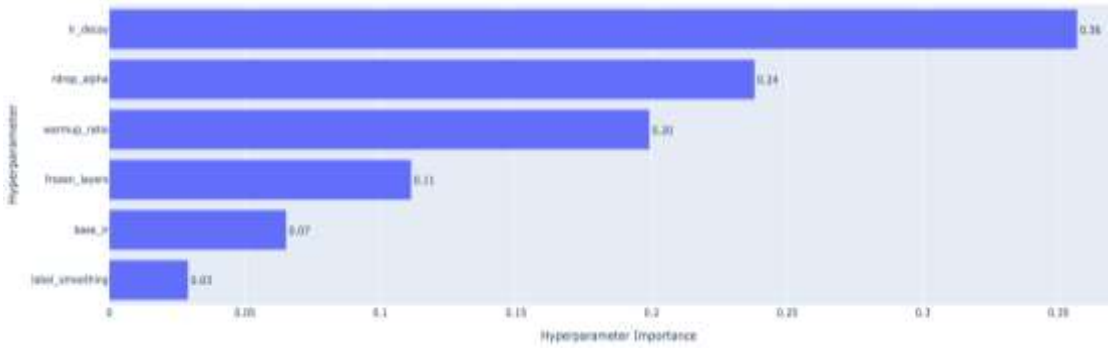

Figure 4.13 Hyperparameter importance scores estimated by Optuna over 10 trials

The Optuna-best configuration was then retrained for 10 epochs on the full training set. Test accuracy reached 0.8743 and macro F1 reached 0.8612, which is statistically indistinguishable from the DeBERTa-v3-Large configuration. Table 4.10 reports this

comparison. The training and validation loss of the retrained run is shown in Figure 4.14, and the corresponding test confusion matrix in Figure 4.15. The loss curves show the same controlled behaviour as in Section 4.2: a smooth decrease with a small gap between training and validation, and stabilisation by about step 4000.

Table 4.10 Stage-2 configuration vs. Optuna-optimal retraining (three-class)

Configuration	**Accuracy**	**Macro F1**
Stage-2 configuration (Section 3.2.4)	0.8835	0.8712
Optuna best (retrained with Trial 8)	0.8743	0.8612
Difference	−0.0092	−0.0100

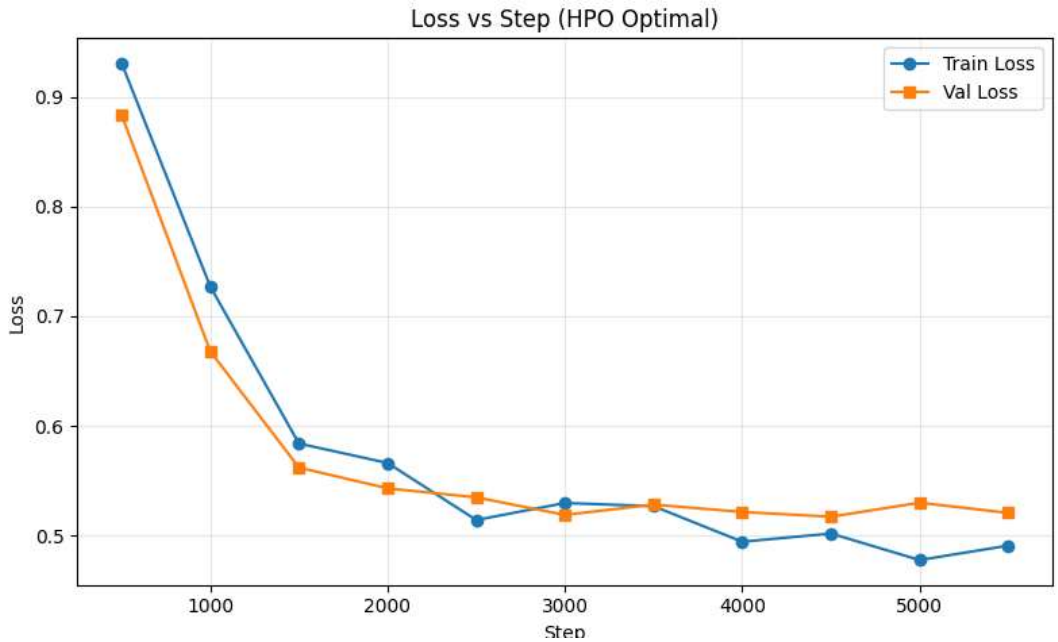

Figure 4.14 Training and validation loss for the retrained model

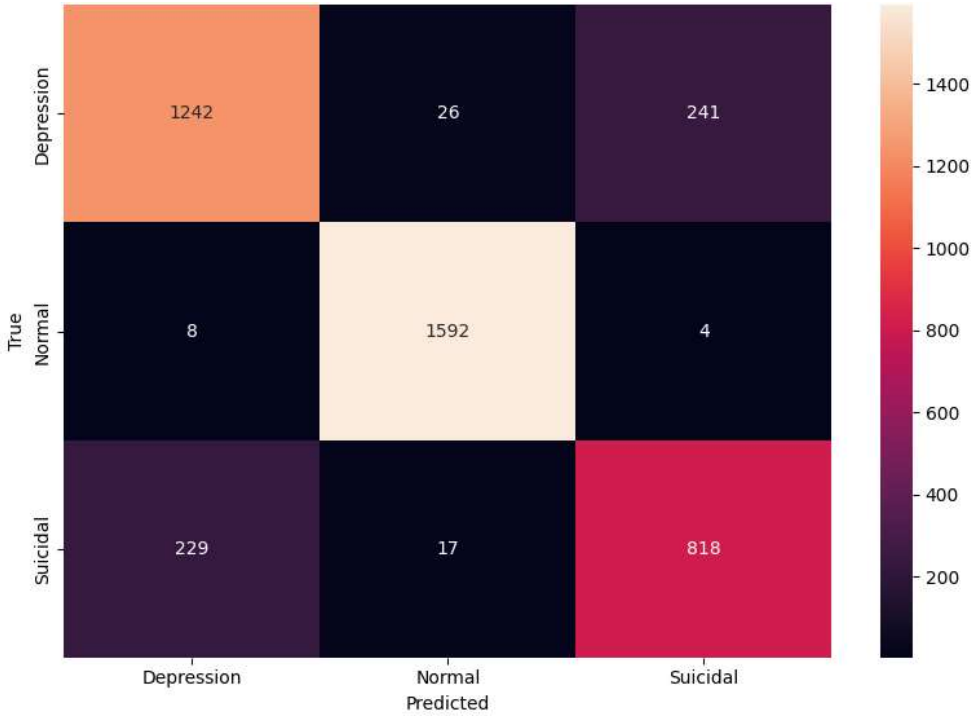

Figure 4.15 Test confusion matrix of the retrained model

The HPO experiment did not yield a meaningful improvement over the manually selected Stage-2 configuration; the retrained model performed about one F1 point lower on the test set. This null result is best interpreted as a positive statement about the Stage-2 configuration rather than a failure of the search. The choices made on the basis of prior work on DeBERTa fine-tuning (He et al., 2021; Howard & Ruder, 2018) already place the model in a flat region of the validation landscape, so small perturbations are absorbed by the remaining coordinates. The fact that a systematic search produced a slight decrease rather than an increase suggests that the model already sits close to its performance ceiling on this dataset, and that further gains would more likely come from changes at the data or architecture level than from additional hyperparameter tuning.

4.5 Model Stability Test Results (5-Fold CV)

To check whether model success depends on a particular train/validation split, a stratified five-fold cross-validation was carried out on the 90 percent train + validation portion of the corpus, with the 10 percent test set held constant across folds (Kohavi, 1995).

Table 4.11 reports the per-fold validation and test scores. The mean test accuracy across the five folds is 0.8755 (std. 0.0020) and the mean test macro F1 is 0.8644 (std. 0.0023). Fold-level test macro F1 is bounded between 0.8607 (Fold 2) and 0.8668 (Fold 5), a range of 0.6 F1 points.

Table 4.11 Five-fold cross-validation results of the proposed model (three-class)

Fold	**Val. Accuracy**	**Val. Macro F1**	**Test Accuracy**	**Test Macro F1**
1	0.8666	0.8552	0.8755	0.8646
2	0.8810	0.8703	0.8726	0.8607
3	0.8759	0.8654	0.8743	0.8641
4	0.8732	0.8619	0.8767	0.8660
5	0.8784	0.8681	0.8781	0.8668
Mean	0.8750	0.8642	0.8755	0.8644
Std.	0.0055	0.0059	0.0020	0.0023

This is shown graphically in Figure 4.16, where each bar represents one fold and the dashed lines mark the single-model and ensemble references.

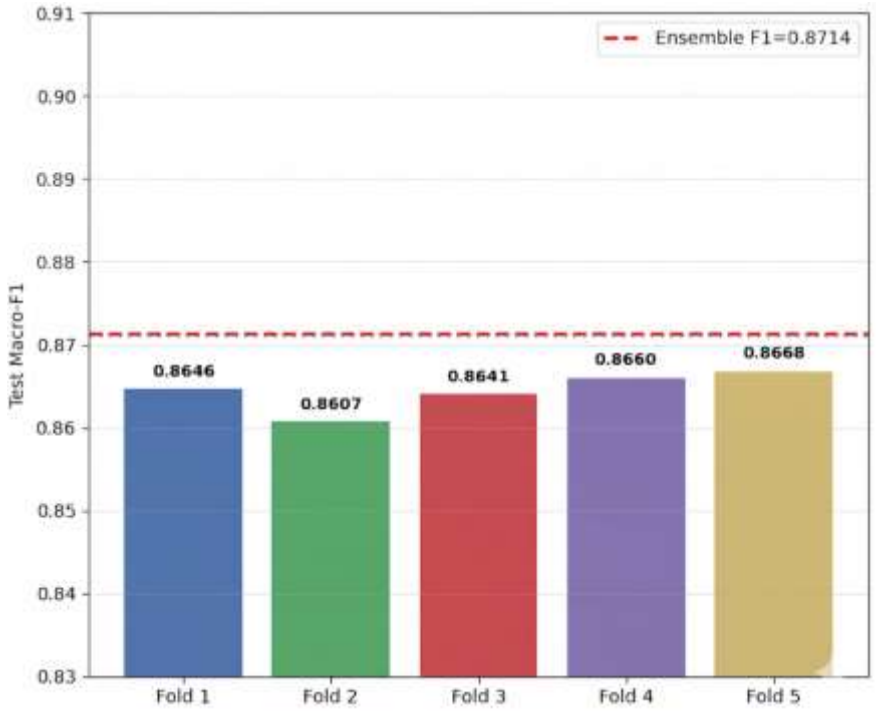

Figure 4.16 Per-fold test macro F1 of the proposed model

A standard deviation below 0.003 on the test set means that the reported accuracy of the proposed model is not the consequence of a favourable single split. The ensemble reaches a test accuracy of 0.8820 and a macro F1 of 0.8714.

Table 4.12 Five-fold ensemble: per-class test performance (three-class)

Class	**Precision**	**Recall**	**F1**	**Overall Performance**
Normal	0.985	0.990	0.987	**Accuracy:** 0.8820
Depression	0.876	0.797	0.835	**Macro-F1:** 0.871
Suicidal	0.750	0.839	0.792	

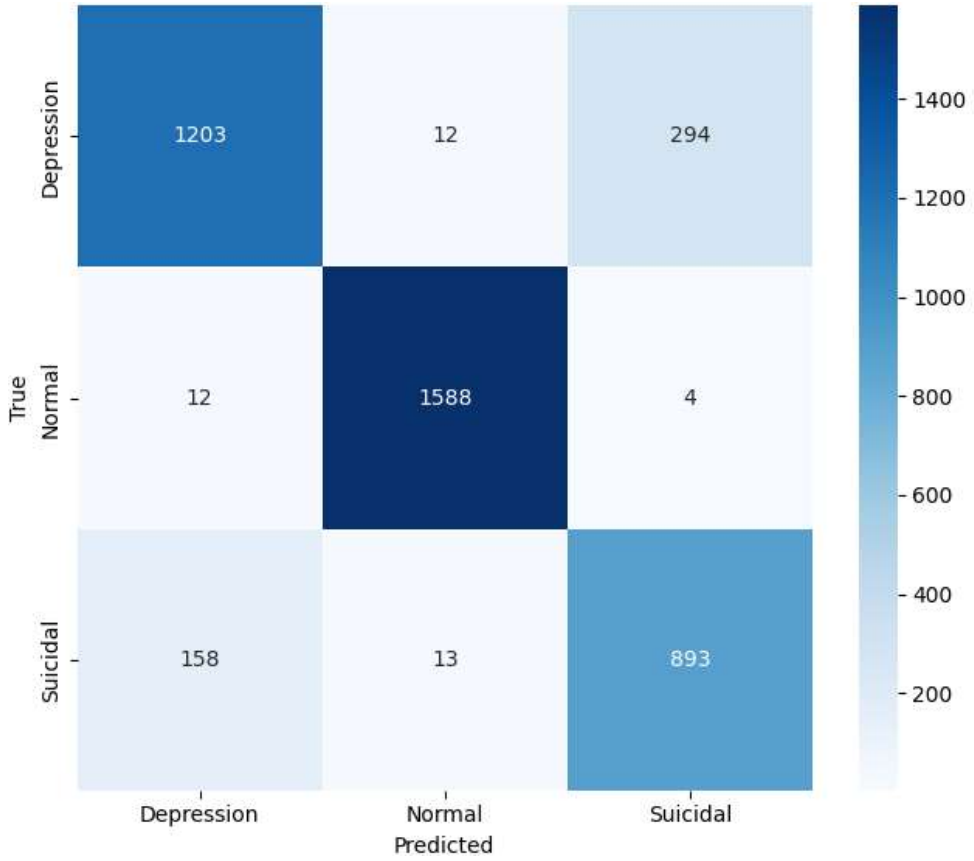

Figure 4.17 Test confusion matrix of the five-fold ensemble

Compared with the single Stage-2 model (0.8835 / 0.8712), the ensemble produces a limited change on accuracy (−0.0015) and a limited gain on macro F1 (+0.0002), at the cost of fivefold inference. This is summarised in Table 4.13. The very small ensemble gain confirms that the five fold-models converge to similar solutions on the test set, which is the expected behaviour of a well-regularised model whose performance does not depend on a single random split.

Table 4.13 Single Stage-2 model vs. five-fold ensemble on the test set

Configuration	**Accuracy**	**Macro F1**
Single model (Stage-2)	0.8835	0.8712
Five-fold ensemble	0.8820	0.8714
Δ (ensemble − single)	−0.0015	+0.0002

4.6 Error Analysis Results

The error analysis contrasts the proposed DeBERTa model with the reproduced SVC baseline on the common three-class test set. Three disjoint groups of test samples are considered: (i) DeBERTa correct and SVC wrong (224 samples); (ii) SVC correct and DeBERTa wrong (58 samples); and (iii) both models wrong (123 samples). The asymmetry between (i) and (ii) already reflects the aggregate advantage of DeBERTa, but the qualitative content of each group is more informative than the counts. Tables 4.14 to 4.16 give examples from each group.

Table 4.14 DeBERTa is correct and SVC is wrong

Excerpt	Ground truth	DeBERTa	SVC
"A minute too late and a dollar too short… I feel like I dodged so many bullets in life then the last one finally got me."	Suicidal	Suicidal ✓	Depression ✗
"Goodnight everyone I hope I do not wake up tomorrow."	Suicidal	Suicidal ✓	Normal ✗
"I feel like my life's over, I am not smart and I just suck at everything. I am not strong enough. I keep failing."	Suicidal	Suicidal ✓	Depression ✗
"and I am only 20. I am ready to give up, i just want to get away from all of this."	Suicidal	Suicidal ✓	Normal ✗
"So where to start, I am currently 18 made 3 attempts and failed each time …"	Suicidal	Suicidal ✓	Depression ✗
"Used to want to soar through the sky like a bird … Now I just want to go real high and dive down straight into the ground."	Depression	Depression ✓	Normal ✗

The dominant pattern in this group is that DeBERTa correctly recognises implicit, metaphorical or euphemistic expressions of suicidal ideation, such as "I hope I do not wake up tomorrow", "dive down straight into the ground" or "get away from all of this". SVC, which relies on surface n-gram statistics, labels such posts as Normal or Depression and so misses the clinically most serious cases. Of the 224 posts in this group, the majority belong to the Suicidal class. The pattern is consistent with the +11 F1 improvement on the Suicidal class reported in Section 4.3 and supports the view that DeBERTa (He et al., 2021) captures the context-dependent meaning of these expressions, while TF-IDF features cannot.

Table 4.15 SVC is correct and DeBERTa is wrong

Excerpt	**Ground truth**	**SVC**	**DeBERTa**
"If you knew me you would see me smiling or talking about how hard time working to get through my mental struggles but honestly what is the point?"	Suicidal	Suicidal ✓	Depression ✗
"because I always want to fucking kms I sexually identify as a unit of speed"	Depression	Depression ✓	Suicidal ✗
"My parents have supported me this past year and its dragging on them … I made some recent mistakes …"	Suicidal	Suicidal ✓	Depression ✗
"My life's only purpose seems to be to work to give into buying stupid pointless crap … I feel so empty …"	Suicidal	Suicidal ✓	Depression ✗
"that is it I do not even have the motivation to start whining I just want things to end …"	Depression	Depression ✓	Suicidal ✗

This group contains 58 samples. The striking feature is that almost every error made by DeBERTa in this group is a Depression to Suicidal confusion: 44 of the 58 posts are Suicidal that DeBERTa labelled Depression, and 13 are Depression that DeBERTa labelled Suicidal. Only one post in the entire group is a Normal vs. non-Normal confusion. The two classes are genuinely close in meaning (hopelessness, exhaustion, the wish for things to "end"), and DeBERTa's errors are therefore of a very different quality from the

SVC errors in Table 4.14: they are fine-grained mistakes along the Depression to Suicidal axis, not coarse failures on the Normal to non-Normal axis. The residual errors of DeBERTa reflect ambiguity in the phenomenon itself, while the residual errors of SVC include outright failures to detect suicidal content.

Table 4.16 Examples misclassified by both models

Excerpt	**Ground truth**	**DeBERTa**	**SVC**
"Is been a while since i posted and left the things halfways. I explained how I was forced to quit from a well paying work … have been unable to recover economically since …"	Suicidal	Depression ✗	Depression ✗
"9111 10335"	Suicidal	Normal ✗	Normal ✗
"Ugly.. depressed.. cannot attract anyone... unemployed.. college dropout … severely socially anxious … never had a …"	Suicidal	Depression ✗	Depression ✗
"my town is depressing as fuck like insane levels of boring … i ended up losing all …"	Suicidal	Depression ✗	Depression ✗
"wow. IF U REACH OUT TO SOMEONE TRYING TO HELP, DO NOT COMPARE THE SEVERITY OF YOURS AND THEIR MENTAL ISSUES …"	Depression	Normal ✗	Normal ✗

The examples that both models misclassify fall into two broad categories. The first is linguistically degenerate posts such as "9111 10335", essentially numerical noise carrying a Suicidal label, where no model can recover the label from the text alone. The second is posts in which the sharp clinical distinction between Depression and Suicidal is carried by a single sentence buried inside a long narrative of economic hardship, social difficulty or family conflict. Both categories indicate that text-only classification on this dataset is approaching its ceiling, and that further progress would likely require richer context such as user history, platform metadata or longer temporal windows.

4.7 Explainability Results

This section reports the token-level attribution analysis described in Section 3.2.7. Integrated Gradients was applied to nine posts drawn from the qualitative error analysis of Section 4.6 (Tables 4.14 and 4.15); the per-example top supporting and opposing tokens are reported in Table 4.17, and three representative posts are shown as heat-maps in Figures 4.18 to 4.20.

Table 4.17 Top supporting/opposing tokens for 9 examples

Excerpt	True/ Predicted	Top supporting tokens	Top opposing tokens	delta
A minute too late and a dollar too short...Ya know? I feel like I dodged so many...	Suicidal/ Suicidal	late(+0.22), short(+0.16), minute(+0.14), feel(+0.13), like(+0.12)	too(-0.09), I(-0.13), so(-0.20), A(-0.27), dodged(-0.28)	-0.0888
Goodnight everyone I hope I do not wake up tomorrow	Suicidal/ Suicidal	hope(+0.73), not(+0.51), everyone(+0.20), Goodnight(+0.18), wake(+0.15)	I(-0.04), tomorrow(-0.22), I(-0.23)	-0.0171
and I am only 20. I am ready to give up, i just want to get away from all of thi...	Suicidal/ Suicidal	from(+0.17), ready(+0.16), to(+0.15), get(+0.15), away(+0.13)	and(-0.10), am(-0.12), am(-0.15), i(-0.16), 20(-0.26)	0.0148
So where to start, I am currently 18 made 3 attempts and failed eatch time. I am...	Suicidal/ Suicidal	So(+0.24), just(+0.11), where(+0.10), attempt(+0.09), frequent(+0.09)	eat(-0.08), gailed(-0.08), a(-0.09), some(-0.15)	0.1932
Used to want to soar through the sky like a bird and see everything cool and stu...	Depression/ Depression	skydiving(+0.79), soar(+0.30), Maybe(+0.19), Used(+0.18), just(+0.12)	ground(-0.01), is(-0.02), cool(-0.02), always(-0.07), wanted(-0.11)	-0.0245
If you knew me you would see me smiling or talking about how hard time working	Suicidal/ Depression	If(+0.20), at(+0.17), the(+0.16), to(+0.14), the(+0.10)	of(-0.10), vid(-0.11), headlines(-0.16), the(-0.20), What(-0.38)	-0.0069
because I always want to fucking kms I sexually identify as a unit of speed	Depression/ Suicidal	kms(+0.85), of(+0.22), because(+0.14), always(+0.08), sexually(+0.06)	identify(-0.09), as(-0.11), to(-0.13), unit(-0.19), I(-0.35)	-0.0093
that is it I do not even have the motivation to start whining I just want things...	Depression/ Suicidal	whining(+0.48), psychological(+0.29), not(+0.19), seem(+0.18), the(+0.18)	and(-0.07), I(-0.07), do(-0.09), my(-0.12), I(-0.15)	0.0003
My life's only purpose seems to be to work	Suicidal/ Normal	work(+0.42), life(+0.22), to(+0.15)	seems(-0.08), purpose(-0.24), be(-0.35), only(-0.35)	0.0115

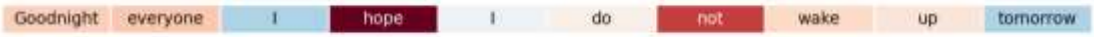

Figure 4.18 Token-level Integrated Gradients example 1 (correctly classified)

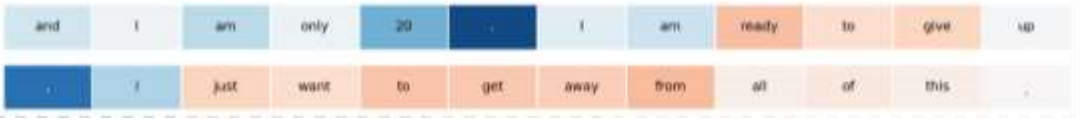

Figure 4.19 Token-level Integrated Gradients example 2 (correctly classified)

Figure 4.20 Token-level Integrated Gradients example 3 (misclassified)

In the correctly classified posts the highest-attribution tokens are short multi-word constructions that encode the implicit pathological cue rather than direct clinical vocabulary. In the post "Goodnight everyone I hope I do not wake up tomorrow" (Figure 4.18; true and predicted: Suicidal), the strongest positive attributions concentrate on the negation–verb construction "hope" (+0.73), "not" (+0.51) and "wake" (+0.15); the temporal marker "tomorrow" itself carries a negative attribution (−0.22), which means it is treated as neutral once the surrounding construction has already supplied the Suicidal cue. In the post "and I am only 20. I am ready to give up, i just want to get away from all of this" (Figure 4.19; true and predicted: Suicidal), the strongest attributions appear on the escape-related sequence "ready" (+0.16), "get" (+0.15), "away" (+0.13) and "from" (+0.17), which jointly encode a metaphor of permanent departure rather than any single class-defining word. The remaining correctly classified examples in Table 4.17 exhibit the same pattern of focal attribution on contextual rather than lexical cues.

The misclassified posts in Table 4.17 illustrate two error modes. The first is focal misattribution: in the post "because I always want to fucking kms I sexually identify as a unit of speed" (Figure 4.20; true: Depression, predicted: Suicidal), the abbreviation "kms" receives a very high attribution (+0.85) toward the Suicidal class. The ground truth is Depression because the post is a joke in which "kms" means "kilometres per second", but

the model does not recover the pragmatic frame supplied by the sequel. The second error mode is diffuse attribution under semantic ambiguity, illustrated in Table 4.17 by posts such as "If you knew me you would see me smiling…" where no single token receives more than +0.20 attribution and the prediction is unstable between Depression and Suicidal. Together these findings indicate that the model relies on context-dependent constructions rather than on a vocabulary of class-defining keywords, and that its residual errors are concentrated on posts whose pragmatic frame or available text cannot reliably separate Depression from Suicidal.

4.8 Discussion

4.8.1 Why DeBERTa outperforms the baseline

Two aspects of DeBERTa-v3-Large appear to drive most of the improvement. First, the disentangled attention introduced in Section 2.4 (He et al., 2021), which models token interactions through both content and relative position is structurally well matched to mental-health text. Posts in this domain rely on idiomatic and metaphorical expressions of emotional state, such as "I hope I do not wake up" or "get away from all of this", whose meaning cannot be recovered from unigram or bigram statistics and whose polarity depends on the surrounding clause. Table 4.14 shows that DeBERTa correctly recognises such posts as Suicidal while SVC does not. Second, the ELECTRA-style pre-training of DeBERTa-v3 with gradient-disentangled embedding sharing (He et al., 2023) yields representations that adapt well under partial fine-tuning, which is consistent with the observation that only the top eight encoder layers needed to remain trainable to reach the reported accuracy.

4.8.2 HPO and stability

The HPO experiment of Section 4.4 did not yield a meaningful improvement over the DeBERTa-v3-Large configuration. The ten Optuna trials all fall within a 0.7-point F1 band on the validation set, and retraining with the best trial gives a test macro F1 that is

one point lower than the DeBERTa-v3-Large configuration. This null finding is itself informative. It shows that the combination of frozen bottom layers, LLRD, R-Drop at $\alpha = 0.5$ and a 1×10^{-5} base learning rate already sits in a flat region of the validation landscape, so small perturbations along any one coordinate are absorbed by the remaining ones. The five-fold cross-validation supports the same conclusion from a different angle: the standard deviation of test macro F1 across the five folds is below 0.003, so the reported accuracy is not an artefact of a particular split.

4.8.3 What the error analysis tells us

The asymmetry between the two error groups in Section 4.6 is the most informative qualitative finding. Where SVC errs, the typical failure is a missed Suicidal post, often because the cue is a metaphor that does not match any informative TF-IDF feature. Where DeBERTa errs, the typical failure is a confusion along the Depression to Suicidal axis, that is, between two semantically adjacent mental-health states that even human annotators would disagree on in short excerpts. The residual errors of the proposed model are therefore of an intrinsically harder kind, and further progress on this dataset is likely to require either richer context (longer threads, user-level features) or a re-examination of the label boundaries themselves.

4.8.4 What the explainability analysis adds

The token-level attributions reported in Section 4.7 do not change the aggregate result but they refine the interpretation of two earlier findings of this chapter.

The first finding is the +11 macro-F1 improvement on the Suicidal class reported in Section 4.3. Section 4.6 attributed this gain qualitatively to metaphor recognition, on the basis of a visual reading of the error analysis examples. The Integrated Gradients results quantify that claim: in the correctly classified posts the highest-attribution tokens are short metaphorical constructions ("hope … not … wake", "ready to get away from") rather than direct clinical vocabulary, and this focal-attribution is consistent with the disentangled

attention mechanism of DeBERTa (He et al., 2021), which encodes content and relative position separately and can therefore assign high attribution to otherwise neutral tokens when their surrounding context activates a class-specific cue.

The second finding is the concentration of residual errors on the Depression–Suicidal axis. The attribution analysis identifies two distinct error modes underlying these errors. Focal misattribution occurs when the model concentrates attribution on a salient surface cue whose pragmatic interpretation conflicts with the literal one, as in the joke involving the abbreviation "kms". Diffuse attribution under genuine ambiguity occurs when no single token carries enough class-specific information to resolve the Depression–Suicidal distinction. Together these patterns support the qualitative claim of Section 4.6 that the residual errors of the proposed model are not failures of feature recognition but cases in which the available text is either pragmatically misleading or lexically inseparable between the two pathological classes.

Taken together, the results support the two hypotheses set out in Chapter 1. The reference baselines of Aksoy (2025) are reproducible within natural variance, and a carefully fine-tuned DeBERTa-v3-Large model improves over them by approximately 7.6 accuracy points and 8.1 macro F1 points in the three-class setting and by about 4.3 / 4.5 points in the binary setting, with stable performance across splits and a qualitatively different and clinically more favourable error profile.

5. CONCLUSION

This thesis investigated whether a fine-tuned Transformer encoder can outperform the classical and hybrid baselines used in prior mental-health text classification work, and whether the resulting gain is stable enough to warrant the additional computational cost. The investigation followed the three-stage protocol described in Chapter 3: a controlled reproduction of the baselines reported by Aksoy (2025), the fine-tuning of DeBERTa-v3-Large under a fixed configuration, and a validation stage consisting of an Optuna-based hyperparameter search and five-fold stratified cross-validation. The dataset, preprocessing, and stratified 80/10/10 partition were held constant across the comparison, so any difference in test-set performance can be attributed to the model rather than to a difference in protocol.

5.1 Summary of Findings

The reproduction of the four reference baselines (SVC, SGD, Random Forest, USE+BiLSTM) yielded test scores within one accuracy point of the values published by Aksoy (2025), which confirms that the comparison rests on a verified reference rather than on figures recomputed under a different protocol. With the baselines anchored, fine-tuning DeBERTa-v3-Large on the same partition produced the three-class and binary test scores detailed in Section 4.2. The improvement of about 7 to 8 macro F1 points over the strongest baseline in the three-class task and about 4 macro F1 points in the binary task is therefore the central empirical result of the thesis.

The class-wise picture is more informative than the aggregate. The largest single gain on the three-class task is on the Suicidal class, whose F1 rises from 0.67 (reproduced SVC) to 0.78 (DeBERTa). The Normal class is already close to ceiling for both models, and Depression improves by about 7 F1 points. The proposed model therefore concentrates its improvement on the under-represented and clinically most consequential category, which was the principal motivation for adopting a contextualised encoder.

The validation stage confirms that this improvement is not an artefact of a single split or of a particular hyperparameter choice. Five-fold stratified cross-validation produced fold-level test macro F1 values in the range 0.8607 to 0.8668, with a standard deviation below 0.003. A ten-trial Optuna search over six hyperparameters produced no further test-time improvement; retraining from scratch with the best trial yielded a macro F1 of 0.8612, one F1 point below the manually selected configuration. This null finding is informative because it shows that the chosen configuration already lies in a flat region of the validation landscape, so the reported gains do not depend on aggressive hyperparameter tuning.

The qualitative error analysis explains the source of the improvement. The proposed model correctly classifies most posts in which suicidal ideation is expressed indirectly, through metaphor, euphemism, or extended narrative, that surface n-gram statistics cannot recognise. The residual errors of the model are concentrated along the Depression to Suicidal axis, that is, between two semantically adjacent categories, rather than between Normal and the two pathological classes. The remaining errors are therefore of an intrinsically harder kind than the errors of the classical baselines.

A token-level explainability analysis based on Integrated Gradients (Section 4.7) provides quantitative evidence for the qualitative claims of the error analysis. In correctly classified posts the highest-attribution tokens are short metaphorical constructions rather than direct clinical vocabulary, which is consistent with the disentangled attention mechanism of DeBERTa. In the misclassified posts the same analysis identifies two distinct error modes; focal misattribution on pragmatically misleading cues, and diffuse attribution under semantic ambiguity that together characterise the residual confusion on the Depression–Suicidal axis.

5.2 Answers to the Research Questions

RQ1 is answered affirmatively. Fine-tuning DeBERTa-v3-Large improves the three-class macro F1 from 0.79, the strongest baseline in Aksoy (2025), to 0.8712 in the present study, an absolute gain of about 8 F1 points. The corresponding gain in the binary task is about 4 F1 points.

RQ2 is supported by the qualitative analysis. Texts whose label depends on context-sensitive expressions, such as 'I hope I do not wake up' or 'dive down straight into the ground', are correctly classified by DeBERTa but missed by the SVC baseline. This pattern is consistent with DeBERTa, which captures the polarity of ambiguous tokens that n-gram features treat as neutral.

RQ3 is supported by the cross-validation result (test macro F1 standard deviation below 0.003) and by the HPO null finding. The performance of the model is statistically stable across splits, and the systematic hyperparameter optimisation did not improve it.

RQ4 is answered by the error analysis. The principal limitation of the proposed model is its tendency to confuse Depression with Suicidal in posts that contain hopelessness or wishes for things to end but do not mention explicit ideation. Linguistically degenerate posts, such as numerical noise or very short statements, are misclassified by both DeBERTa and the SVC baseline.

5.3 Contributions

The empirical contribution of the thesis is the controlled comparison itself. By holding dataset, preprocessing, and split constant, the study isolates the effect of replacing a TF-IDF and classical-classifier pipeline with a fine-tuned DeBERTa-v3-Large encoder. The numerical improvement reported in Section 4.3 is therefore attributable to the model rather than to a different evaluation protocol.

The methodological contribution is the integration of layer-wise learning-rate decay, partial layer freezing, R-Drop regularisation, sliding-window chunking, and a class-balanced cross-entropy loss into a single fine-tuning protocol that is reported and tested as a unit. The HPO experiment shows that this combination already sits in a flat region of the validation landscape, which means that the reported improvement is robust to hyperparameter perturbation and does not depend on careful per-trial tuning.

5.4 Limitations and Future Work

Four limitations should be acknowledged. First, all results are obtained on a single English-language corpus assembled from Reddit and Twitter; cross-lingual and cross-platform generalisation is not tested, and a domain shift to clinical or non-English social media data could change the absolute numbers. Second, the compute budget restricted the HPO experiment to ten Optuna trials of six epochs each and prevented us from running larger architectures, longer training schedules, or HPO across all five folds simultaneously. A larger search budget could uncover gains that the present search did not find. Third, the deduplication step applied before DeBERTa training differs from the reference pipeline; although the impact on class counts is small, the two splits are not strictly identical, and a fully matched setting with a common split would tighten the comparison further. Fourth, and most importantly, the study does not include clinical validation: the labels are derived from public dataset metadata and have not been assessed by a clinician on individual posts. The proposed model should therefore be interpreted as a research artefact that demonstrates the advantage of modern Transformer architectures on a well-defined text-classification task, not as a diagnostic tool.

Future work could address these limitations along several axes. Cross-lingual experiments on Turkish or other non-English corpora would test whether the disentangled attention mechanism transfers under language shift. Domain-adapted pre-training on a mental-health corpus, in the spirit of MentalBERT (Ji et al., 2022), could be combined with the fine-tuning protocol of this thesis to test whether domain adaptation and architectural improvement are additive. The explainability analysis in this thesis is restricted to Integrated Gradients; future work could combine this attribution method with alternative families such as layer-wise relevance propagation or attention-head probing, in order to triangulate the model's decision process from multiple complementary perspectives. Finally, augmenting the input with user-level or temporal context, such as longer threads or posting history, could help the model resolve the residual Depression to Suicidal confusion that remains its main source of error.

5.5 Closing Remarks

The thesis demonstrates that a carefully fine-tuned DeBERTa-v3-Large model improves over the classical and hybrid baselines used in prior work on this dataset by roughly 7 to 8 macro F1 points in the three-class setting and 4 macro F1 points in the binary setting, with stable performance across data splits and a more favourable distribution of residual errors from a clinical perspective. The improvement comes at a higher computational cost, but the cost is bounded (a single A100-80GB GPU was sufficient for both training and the validation experiments), the gain is concentrated in the under-represented and most consequential class, and the validation stage shows that the gain does not depend on a fortunate split or on aggressive hyperparameter tuning. On the well-defined classification task addressed here, the answer to the question raised in Chapter 1 is therefore that the additional computational cost of fine-tuning DeBERTa-v3-Large is justified by the empirical performance gain, and the resulting model can use as a stronger baseline for subsequent research on mental-health classification.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623–2631). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330701

Aksoy, S. 2025. Textual sentiment analysis for mental health diagnosis. In A. Kumar, F. Ortiz-Rodriguez, J. B. De Vasconcelos, P. K. Dutta, H. K. Saini, & P. S. Rathore (Eds.), Adversarial deep generative techniques for early diagnosis of neurological conditions and mental health practises: Theoretical insights with practical applications (Information Systems Engineering and Management, Vol. 46). Springer. https://doi.org/10.1007/978-3-031-91147-7_15

Aldhyani, T. H. H., Alsubari, S. N., Alshebami, A. S., Alkahtani, H., and Ahmed, Z. A. T. 2022. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. International Journal of Environmental Research and Public Health, 19(19), 12635. https://doi.org/10.3390/ijerph191912635

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. 2013. How noisy social media text, how diffrnt social media sources? Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP 2013), 356-364.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), Proceedings of COMPSTAT'2010 (pp. 177–186). Physica-Verlag. https://doi.org/10.1007/978-3-7908-2604-3_16

Breiman, L. 2001. Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. 2017. Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649-685. https://doi.org/10.1017/S1351324916000383

Camacho-Collados, J., and Pilehvar, M. T. 2018. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (pp. 40–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5406

Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. 2018. Universal sentence encoder. arXiv. https://arxiv.org/abs/1803.11175

Chancellor, S., and De Choudhury, M. 2020. Methods in predictive techniques for mental health status on social media: A critical review. npj Digital Medicine, 3(1), Article 43. https://doi.org/10.1038/s41746-020-0233-7

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=r1xMH1BtvB

Coppersmith, G., Dredze, M., and Harman, C. 2014. Quantifying mental health signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (pp. 51–60). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3207

Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. 2018. Natural language processing of social media as screening for suicide risk. Biomedical Informatics Insights, 10, 1–11. https://doi.org/10.1177/1178222618792860

Cortes, C., and Vapnik, V. 1995. Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9268–9277). IEEE. https://doi.org/10.1109/CVPR.2019.00949

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. 2020. A survey of the state of explainable AI for natural language processing. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (pp. 447–459). Association for Computational Linguistics. https://aclanthology.org/2020.aacl-main.46/

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. 2013. Predicting depression via social media. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM), 7(1), 128-137. https://doi.org/10.1609/icwsm.v7i1.14432

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dietterich, T. G. 2000. Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol. 1857 (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

GBD 2019 Mental Disorders Collaborators. 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: A systematic analysis for the Global Burden of Disease Study 2019. The Lancet Psychiatry, 9(2), 137-150. https://doi.org/10.1016/S2215-0366(21)00395-3

Garg, M. 2023. Mental health analysis in social media posts: A survey. Archives of Computational Methods in Engineering, 30, 1819-1842. https://doi.org/10.1007/s11831-022-09863-z

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., and Dutta, R. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. Scientific Reports, 7, 45141. https://doi.org/10.1038/srep45141

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. 2017. Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43-49. https://doi.org/10.1016/j.cobeha.2017.07.005

He, P., Gao, J., and Chen, W. 2023. DeBERTa-v3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=sE7-XhLxHA

He, P., Liu, X., Gao, J., and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=XPZIaotutsD

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hossain, M. M., Sanjara, S., Hossain, M., Chaki, S., Rahman, M. S., and Ali, A. B. M. S. 2025. Revolutionizing mental health sentiment analysis with BERT-Fuse: A hybrid deep learning model. IEEE Access, 13, 1–15. https://doi.org/10.1109/ACCESS.2025.3568340

Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Islam, K. M. S., Fields, J., and Madiraju, P. 2025. multiMentalRoBERTa: A fine-tuned multiclass classifier for mental health disorder. 2025 IEEE International Conference on Big Data (BigData).

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 7184–7190). European Language Resources Association.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), Machine Learning: ECML-98 (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683

Juanita, S., Daeli, A. H., Syafrullah, M., Anggraeni, W., and Purnomo, M. H. 2025. A machine learning approach for text pattern diagnosis in mental health consultations. Decision Analytics Journal, 15, 100572. https://doi.org/10.1016/j.dajour.2025.100572

Kallstenius, T., Capusan, A. J., Andersson, G., and Williamson, A. 2025. Comparing traditional natural language processing and large language models for mental health status classification: A multi-model evaluation. Scientific Reports, 15, 15231. https://doi.org/10.1038/s41598-025-15231-1

Kamarudin, N. S., Beigi, G., Manikonda, L., and Liu, H. 2020. Social media for mental health: Data, methods, and findings. In M. A. Tayebi, U. Glasser, & D. B. Skillicorn (Eds.), Open source intelligence and cyber crime: Social media analytics. Lecture Notes in Social Networks (pp. 187-211). Springer. https://doi.org/10.1007/978-3-030-41251-7_8

Kaushik, P., and Sharma, P. 2024. Integrating machine learning and sentiment analysis: A comparative study on mental health classification from social media data. 2024 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 1–6). IEEE. https://doi.org/10.1109/DASA63652.2024.10836490

Khan, H. U., Naz, A., Alarfaj, F. K., and Almusallam, N. 2025. Analyzing student mental health with RoBERTa-Large: A sentiment analysis and data analytics approach.

Frontiers in Data Science, 1, 1615788. https://doi.org/10.3389/fdata.2025.1615788

Kim, Y. 2014. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–1751). https://doi.org/10.3115/v1/D14-1181

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 226–239. https://doi.org/10.1109/34.667881

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) (Vol. 2, pp. 1137–1143). Morgan Kaufmann.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv*. https://arxiv.org/abs/2009.07896

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. 2019. Text classification algorithms: A survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=H1eA7AEtvS

Lavanya, P. M., and Sasikala, E. 2021. Deep learning techniques on text classification using natural language processing (NLP) in social healthcare network: A comprehensive survey. 2021 3rd International Conference on Signal Processing and Communication (ICPSC) (pp. 603–609). IEEE. https://doi.org/10.1109/ICSPC51351.2021.9451752

Lee, J., Tang, R., and Lin, J. 2019. What would Elsa do? Freezing layers during transformer fine-tuning. arXiv. https://arxiv.org/abs/1911.03090

Li, W., Sun, K., Wang, S., Zhu, Y., Dai, X., and Hu, L. 2024. DePNR: A DeBERTa-based deep learning model with complete position embedding for place name recognition from geographical literature. Transactions in GIS, 28, 993–1020. https://doi.org/10.1111/tgis.13170

Liang, X., Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., and Liu, T.-Y. 2021. R-Drop: Regularized dropout for neural networks. In Advances in Neural Information Processing Systems (Vol. 34, pp. 10890–10905). Curran Associates.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Loshchilov, I., and Hutter, F. 2019. Decoupled weight decay regularization. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=Bkg6RiCqY7

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv. https://arxiv.org/abs/1301.3781

Müller, R., Kornblith, S., and Hinton, G. 2019. When does label smoothing help? Advances in Neural Information Processing Systems, 32, 4694–4703.

Murarka, A., Radhakrishnan, B., and Ravichandran, S. 2020. Detection and classification of mental illnesses on social media using RoBERTa. arXiv. https://arxiv.org/abs/2011.11226

Naslund, J. A., Bondre, A., Torous, J., and Aschbrenner, K. A. 2020. Social media and mental health: Benefits, risks, and opportunities for research and practice. Journal of Technology in Behavioral Science, 5, 245–257. https://doi.org/10.1007/s41347-020-00134-x

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. 2019. Hierarchical Transformers for long document classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 838–844). IEEE. https://doi.org/10.1109/ASRU46091.2019.9003958

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. 2018. Deep contextualized word representations. Proceedings of NAACL-HLT 2018 (pp. 2227–2237). https://doi.org/10.18653/v1/N18-1202

Rehman, A., Ahmed, M., Khan, H. U., Bukhari, A., Daud, A., and Dawood, H. 2025. Mental health sentiment analysis: Exploring an optimized BERT with deep encodings. Engineering, Technology & Applied Science Research, 15(5), 26242–26248. https://doi.org/10.48084/etasr.11265

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Sarkar, S. 2024. Sentiment analysis for mental health [Dataset]. Kaggle. https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health

Sebastiani, F. 2002. Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282.505283

Sokolova, M., and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Sun, C., Qiu, X., Xu, Y., and Huang, X. 2019. How to fine-tune BERT for text classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science, Vol. 11856 (pp. 194–206). Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Sundararajan, M., Taly, A., and Yan, Q. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML*)*, 70, 3319–3328.

Szymański, J. 2014. Comparative analysis of text representation methods using classification. Cybernetics and Systems, 45(2), 180–199. https://doi.org/10.1080/01969722.2014.874828

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Wagay, F. A., and Jahiruddin. 2026. An efficient approach to detecting mental illness on online social media platforms by using multidimensional textual features. Acta Psychologica, 262, 106191. https://doi.org/10.1016/j.actpsy.2025.106191

World Health Organization. 2021. Suicide worldwide in 2019: Global health estimates. WHO Press. https://www.who.int/publications/i/item/9789240026643

World Health Organization. 2025, September 2. Over a billion people living with mental health conditions – services require urgent scale-up. WHO News. https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up

Wu, Y., and Xing, Y. 2024. Efficient machine translation with a BiLSTM-Attention approach. arXiv. https://arxiv.org/abs/2410.22335

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 818–833). Springer. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, Y., Wang, Z., Ding, Z., Tian, Y., Dai, J., Shen, X., Liu, Y., and Cao, Y. 2023. Employing machine learning and deep learning models for mental illness detection. arXiv. https://arxiv.org/abs/2306.15497