



**Mental Health Sentiment & Emotional State Detection
Using Natural Language Processing**

Prepared By

Name: Radwan Haj Masoud

ID No.: [253000985](#)

Submitted in fulfillment of the requirements for the degree Master
of Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Amjad Qatish

Collage of Informatics, Midocean University, Comoros

Email: 253000985@st.midocean.edu.km

Acknowledgements

Special thanks to Dr. .

Abstract

List of Publications

Table of Contents

List of Abbreviations	vii
List of Figures	viii
List of Tables.....	ix

Chapter 1: Introduction and

Overview.....	1
1.2 Problem Definition	2
1.3 Thesis Objectives	3
1.4 Thesis Contribution	4
1.5 Thesis Organization.....	5

Chapter 2: Background and Related Work

.....	31
-------	----

Chapter 3: Proposed Approach

4.1 Introduction	78
------------------------	----

Chapter 4: Experimental Results and Discussion

Chapter 5: Conclusion and Future Work

7.1 Important Features	127
7.3 Future Work	129

References	130
------------------	-----

List of Abbreviations

- **Acc.:** Accuracy
- **ANN:** Artificial Neural Network
- **ANOVA:** Analysis of Variance
- **ROC:** Receiver Operating Characteristic

List of Figures

2.1 Educational Data Mining Process	7
2.2 Applications of Educational Data Mining	8
2.3 Distribution of Papers that used different attributes for PSP problem	19
2.4 Student Performance Data Set	21
2.5 Educational Process Mining (EPM): A Learning Analytics Data Set ...	23

List of Tables

2.1	Student Performance Data Set Description	21
2.2	Students' Academic Performance Dataset Description	23

Chapter 1 Introduction

Chapter 1: Introduction and Overview

In recent years, digital communication platforms have really taken off, changing how people show what they're thinking, feeling, and their mental state. Social media posts, online forums, and digital chats have become key places where people freely talk about their stress, anxiety, depression, and other mental health struggles. This means tons of text get created every day, showing how people feel and think.

At the same time, Artificial Intelligence (AI) and Natural Language Processing (NLP) have gotten much better, letting computers understand and look at human language more accurately. Newer models, like DistilBERT, which are based on 'transformers,' have shown they're good at sorting text. They get the meaning from the whole context; not just how often certain words appear.

Bringing mental health research and NLP together gives us a big chance: we can build automatic systems that can spot people's mental and emotional states just from the text they write.

1.2 Problem Definition

These days, we're talking online increasingly, instead of face-to-face. For ages, we've spotted emotional or mental problems by seeing people in person – noticing how they act at work, school, or out in public. We'd figure out how they felt from talking to them, watching their body language, and just interacting directly.

But since people are interacting online a lot more now, and less in person, this old way of finding mental distress isn't working as well for many communities, including ours in the Middle East.

Today, social media and virtual worlds (like what people call the metaverse) are where folks chat, respond to each other, show feelings, and share personal stories. Some even think that soon, we won't need any physical contact at all, and people will be able to work, play, shop, and even form emotional bonds completely online.

With this trend growing, online users create huge amounts of text every day, which shows their emotional and mental states. Trying to find mental distress by checking all this data by hand isn't practical and just won't work on a big scale.

Plus, regular sentiment analysis often only sorts of things into simply 'good' or 'bad.' It usually can't pick up on more complicated mental health issues like depression, anxiety, stress, or bipolar disorder.

So, there's a clear need for a smart, automatic system that can look at text and figure out people's hidden emotional and mental health states. A system like this could help spot emotional patterns across many people and boost research and understanding in digital mental health, especially in places where it's hard to see people in person or talk to them directly.

1.3 Thesis Objectives

The aim of this thesis is to develop and evaluate a transformer-based text classification model capable of detecting multiple mental health categories from textual statements.

The objectives of this study include:

- Preprocessing and preparing textual data “Sentiment Analysis for Mental Health” dataset from Kaggle.
- Fine-tuning a DistilBERT model for classification.
- Evaluating model performance using accuracy, precision, recall, and F1-score.
- Analyzing classification results and debuting model limitations.

1.4 Thesis Contribution

This thesis contributes both technically and practically to the field of artificial intelligence–based mental health analysis.

From a technical perspective, the study develops and evaluates a transformer-based natural language processing framework capable of detecting emotional and mental health states from textual data. The proposed system goes beyond basic sentiment polarity classification and focuses on identifying higher-risk emotional categories that may indicate psychological distress.

From a practical perspective, this research explores the feasibility of implementing AI-driven emotional state detection systems within digital service ecosystems, particularly in Middle East where government digital platforms are widely adopted.

Many countries in the region, such as the United Arab Emirates and the Kingdom of Saudi Arabia, have developed centralized digital service platforms (e.g., government e-services portals and integrated citizen applications).

Integrating an automated emotional state detection module within such platforms could potentially support early risk identification and proactive intervention strategies for individuals experiencing psychological distress.

Even this thesis does not deploy the system in a real-world governmental environment, but it provides a foundational technical framework that could support future collaboration with digital public service entities aiming to enhance mental health awareness and early detection mechanisms.

Chapter 2 Background and Related Work

2.1 Introduction

Mental health disorders represent a significant global health challenge, affecting millions of individuals worldwide and imposing substantial load on healthcare systems and society. The World Health Organization estimates that depression alone affects over 300 million people globally, while anxiety disorders impact approximately 264 million individuals. Traditional methods of mental health assessment, including clinical interviews and standardized questionnaires, are often time-consuming, resource intensive. Moreover, the stigma associated with mental health issues frequently prevents individuals from seeking professional help, leading to underdiagnosis and delayed intervention.

The increasing of social media platforms has created unprecedented opportunities for mental health research and early detection. Individuals increasingly express their thoughts, emotions, and experiences through digital platforms such as Twitter, Reddit, and Facebook, generating vast amounts of user-generated content that can serve as valuable indicators of mental health status. This digital footprint provides researchers with access to naturalistic language data that reflects genuine emotional states and psychological conditions, offering a complementary approach to traditional clinical assessment methods.

Natural Language Processing (NLP) has emerged as a powerful tool for analyzing textual data and extracting meaningful patterns related to mental health conditions. Recent advances in machine learning and deep learning, particularly the development of transformer-based language models such as BERT (Bidirectional Encoder Representations from Transformers) and its variants, have significantly improved the accuracy and reliability of automated mental health detection systems. These models can capture complex linguistic patterns, semantic relationships, and contextual nuances that are indicative of various mental health states, including depression, anxiety, stress, and suicidal ideation.

This chapter provides a comprehensive review of the existing literature, establishing the theoretical and methodological background for this thesis. It begins by outlining the foundational concepts of NLP in mental health, then dives into a critical analysis of prior work, categorizing it into key thematic areas: the evolution from traditional machine learning to transformer-based models, the crucial need for explainable AI (XAI), and domain-specific applications. The chapter concludes by identifying a clear research gap that this thesis aims to fill, gathered from a detailed literature review table.

2.2 Background: Natural Language Processing and Mental Health

The application of NLP to mental health is predicated on the "language as a window" hypothesis, which situation that an individual's linguistic patterns can reflect their cognitive and emotional state (Pennebaker, 2011). Early work in this domain relied heavily on lexicon-based approaches, such as the Linguistic Inquiry and Word Count (LIWC), which quantified the use of predefined word categories (e.g., pronouns, affective words) to infer psychological states. While foundational, these methods lacked the ability to capture context, semantics, and complex linguistic nuances.

The field has since been revolutionized using machine learning (ML) and, more recently, deep learning. Traditional ML models like Support Vector Machines (SVMs) and Logistic Regression, trained on features such as n-grams, part-of-speech tags, and sentiment scores, represented a significant step forward, offering more data-driven and adaptable models.

However, the most transformative shift has been the introduction of transformer-based language models. Models like BERT (Bidirectional Encoder Representations from Transformers) and its variants (RoBERTa, MentalBERT, DistilBERT) have set new state-of-the-art benchmarks by learning deep, bidirectional contextual representations of text. These models can understand complex linguistic structures, sarcasm, and expressions, which are often critical for accurately detecting mental health conditions from informal social media texts. This thesis is situated within this modern context, aiming to leverage these advanced models for a multi-class mental health classification task.

2.3 Evolution of Methodologies: From Traditional NLP to Transformer Models

A central insight in recent literature is the comparative evaluation of traditional NLP methods against modern transformer-based models for mental health classification. This body of work consistently demonstrates the superior performance of transformer models, although with increased computational cost. In their 2025 study, Köckritz et al. (Paper 11) provide a timely and direct comparison, performing a multi-model evaluation on mental health status classification. Their work is particularly relevant as it benchmarks traditional models like Logistic Regression and SVMs against transformer models such as BERT, RoBERTa, and MentalBERT. The authors conclude that while traditional models offer a faster, more interpretable baseline, transformer models, especially those pre-trained on domain-specific data (like MentalBERT), achieve significantly higher accuracy and F1-scores, effectively capturing the complex and often implicit linguistic markers of mental health conditions.

This finding is reflected in the context of student mental health by Priya et al. (Paper 03), who employed a RoBERTa-Large model for sentiment analysis. Their work demonstrates the efficiency of large-scale transformer models in parsing the unique language patterns found in student social media posts, achieving high accuracy in classifying a spectrum of emotional states. Similarly, Kumar et al. (Paper 04) focus specifically on depression severity, utilizing transformer-based models to move beyond binary classification (depressed vs. non-depressed) towards a more accurate understanding of severity levels. Their results reinforce the transformer architecture's capability to capture subtle gradations in depressive language.

2.4 The Need for Explainability in Mental Health AI

As models become more complex "black boxes", the need for explainable AI (XAI) has become paramount, especially in a sensitive domain like mental health where clinical acceptance and trust are critical. A clinician or user is unlikely to trust a model's prediction without understanding the reasoning behind it.

Jahan et al. (Paper 02) and Kerz et al. (Paper 10) both tackle this critical issue. Jahan et al. focus on depression detection, proposing a framework that combines high-performing black-box models (like transformer networks) with XAI techniques such as LIME (Local Interpretable Model-agnostic Explanations) and

SHAP (SHapley Additive exPlanations). Their work demonstrates that it is possible to maintain high predictive accuracy while providing explanations, highlighting the specific words or phrases in a post that most influenced the model's decision. Kerz et al. (2023) take a broader view, providing a narrative review on the application of XAI for mental health detection. They argue that for NLP models to be viable in clinical or supportive settings, they must be accompanied by methods that elucidate their decision-making processes, thereby enabling error analysis, bias detection, and fostering user trust.

2.5 Methodological Trends and Reviews

Several papers provide valuable insights by focusing on specific domains or consolidating the field's progress.

Chatterjee et al. (Paper 05) present a literature review focused on screening for depression. Their work is a valuable resource for understanding the specific linguistic markers associated with depression, such as increased first-person singular pronouns, negative emotion words, and decreased social engagement terms. This review highlights the evolution from dictionary-based methods to complex deep learning models for this specific condition.

A broader perspective is offered by Zhang et al. (Paper 09) in their narrative review on NLP applied to mental illness detection. This paper provides a high-level overview of the field, discussing data sources (Twitter, Reddit), key tasks (depression, schizophrenia, suicide risk detection), and the trajectory of model architecture. In a separate survey, Zhang et al. (Paper 06) introduce the concept of "emotion fusion" for mental illness detection. This is a more advanced concept where multiple emotional signals (e.g., from text, sentiment, and potentially emojis) are integrated to improve detection accuracy, moving beyond simple sentiment polarity.

Patel et al. (Paper 08) provides a practical tutorial on implementing ML and deep learning models for this task, which is a useful methodological reference. Lastly, Muhammad et al. (Paper 07) extend the conversation to multilingual and multi-emotional detection, highlighting the challenges and opportunities in scaling these models beyond English-language data.

2.6 Research Gap and Contribution of This Thesis

After reviewing the existing work, three simple but important gaps remain:

1. Most studies only detect one mental health condition (e.g., depression) or use two classes (depressed vs. normal). Few models can distinguish between several related states at once, such as anxiety, depression, stress, and suicidal ideation.
2. Even when a model works well, it is often a "black box." Users cannot see why a post was labeled as "anxious" instead of "stressed." This lack of transparency makes the model less trustworthy for real-world use.
3. There is no simple, ready-to-use pipeline that combines multi-class classification with clear explanations. Existing tutorials show either classification or explanation, but not both together using a public dataset.

Contributions of This Thesis:

To fill these gaps, this thesis provides the following simple and practical contributions:

- A model that classifies five mental states. We train a RoBERTa model to label social media posts such as: *Anxiety*, *Depression*, *Stress*, *Normal*, or *Suicidal Ideation*. The model is trained on the public Kaggle dataset "Sentiment Analysis for Mental Health."
- Clear explanations for each prediction. We use SHAP (a popular explainability tool) to show, for every post, which words most influenced the model's decision. For example, you can see why a post was called "depressed" and not "anxious."
- A simple, reusable code. All steps (loading data, training the model, evaluating, and explaining predictions) are written as short, modular Python scripts. The code will be shared so others can easily adapt it to their own datasets or add more mental health categories.
- Complete evaluation per mental state. We report accuracy, precision, recall, and F1-score for each of the five classes. This shows exactly where the model performs well and where it struggles.

2.7 Summary of Literature Review

The following table summarizes the key papers reviewed, highlighting their primary contributions, methodologies, and limitations, which collectively inform the direction of this thesis.

Table 2.1: Summary of Key Literature on NLP for Mental Health Detection

No.	Author(s) & Year	Title	Key Contribution / Focus	Methodology / Models	Key Findings / Limitations
1	Köckritz et al. (2025)	Comparing traditional NLP and language models for mental health status classification: a multi-model evaluation	Compares traditional and transformer models for mental health classification.	Logistic Regression, SVM, BERT, RoBERTa, MentalBERT	Transformers outperform traditional models. Domain-specific pre-training (MentalBERT) yields best results.
2	Jahan et al. (2023	Explainable AI-driven depression detection from social media using NLP and black box machine learning models	Combines high-performance black-box models with XAI for depression detection.	LSTM, CNN, BERT; XAI: LIME, SHAP	Achieves high accuracy with XAI, providing insights into model decisions. XAI is crucial for trust.
3	Priya et al. (2024)	Analyzing student mental health with RoBERTa-Large: a sentiment analysis and data analytics approach	Applies RoBERTa-Large to analyze student mental health from text.	RoBERTa-Large, Sentiment Analysis	Demonstrates high accuracy in classifying student mental health states. Focus on a specific demographic.
4	Kumar et al. (2023	Detection of Depression Severity in Social Media Text Using Transformer-Based Models	Moves beyond binary classification to detect severity levels of depression.	Transformer-based models (e.g., BERT)	Successfully classifies depression severity, a nuanced and clinically relevant task.

5	Chatterjee et al. (2022)	Screening for Depression Using Natural Language Processing: Literature Review	A focused literature review on NLP methods for depression screening.	Review of lexicon-based, ML, and DL methods	Summarizes key linguistic markers and the evolution of methods for depression detection.
6	Zhang, T. et al. (2022)	Emotion fusion for mental illness detection from social media: A survey	Introduces the concept of fusing multiple emotional signals for improved detection.	Survey of emotion fusion techniques	Proposes that integrating diverse emotional data improves detection accuracy over using text alone.
7	Muhammad et al. (2023)	Emotion detection from multilingual text and multi-emotional	Focuses on the challenge of multilingual and multi-emotional detection.	transformer models for multilingual tasks	Highlights the complexity and importance of moving beyond English-only, single-emotion tasks.
8	Patel et al. (2022)	Tutorial on Using Machine Learning and Deep Learning Models for Mental Illness Detection	A practical guide/tutorial for implementing ML/DL models for mental health tasks.	Tutorial covering ML (SVM) and DL (LSTM, CNN)	Serves as a methodological resource for building models in this domain.
9	Zhang, T. et al. (2022)	Natural language processing applied to mental illness detection: a narrative review	A broad narrative review of the entire field, covering data sources and models.	Narrative review	Provides a high-level overview of the field's landscape, progress, and challenges.
10	Kerz, E. et al. (2023)	Toward explainable AI (XAI) for mental health detection based on language behavior	Argues for the necessity of XAI in mental health NLP applications.	Narrative review and proposal of XAI frameworks	Emphasizes that model interpretability is as important as accuracy for clinical relevance.
11	Köckritz, J. et al. (2024)	Current applications and future directions in natural language processing for news	Explores NLP applications for mental health in the	Potentially topic modeling and sentiment analysis	Extends the application domain to news media, suggesting new data sources and research directions.

		media and mental health	context of news media.		
12	Islam et al. (2025)	multiMentalRoBERTa: A Fine-tuned Multiclass Classifier for Mental Health Disorder	Multi-class classification of mental health conditions (stress, anxiety, depression, PTSD, suicidal ideation, neutral).	Fine-tuned RoBERTa; comparison with traditional ML, domain-specific transformers (MentalBERT), and LLMs; explainability via Layer Integrated Gradients and KeyBERT.	Achieves macro F1-scores of 0.839 (6-class) and 0.870 (5-class), outperforming MentalBERT. Reveals strong correlations between depression/suicidal ideation and anxiety/PTSD.
13	Jampana et al. (2024)	Exploring Kolmogorov Arnold Networks for Interpretable Mental Health Detection and Classification from Social Media Text	Introduces Kolmogorov Arnold Networks (KANs) as an interpretable alternative to MLPs for mental health classification.	KANs with LIME for explainability; trained on social media text for mental health detection.	KANs achieve higher accuracy than MLPs with fewer parameters. Enables derivation of governing equations per class for deeper interpretability.
14	Karpontinis & Soufleri (2025)	DP-CARE: a differentially private classifier for mental health analysis in social media posts	Privacy-preserving framework for mental health text classification to prevent data leakage.	Frozen MentalBERT encoder + lightweight classifier trained with Differentially Private AdamW (DP-AdamW).	Achieves F1=78.08% and Recall=88.67% under privacy budget $\epsilon \approx 3$. Demonstrates that privacy and performance can be balanced effectively.
15	Lan et al. (2025)	Depression Detection on Social Media with Large Language Models	Uses LLMs to integrate clinical diagnostic criteria and temporal mood analysis for depression detection.	DORIS framework: LLMs annotate text against medical criteria and summarize mood courses; Gradient Boosting Tree (GBT) classifier for prediction.	Provides explainability via symptom-based justifications. Validated on social media data with strong results. Limitations include computational cost of LLMs.

Chapter 4 Proposed Approach (Methodology)

Chapter 5 :Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References