



**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

Nowadays, people talk online more than face-to-face. For a long time, we've spotted emotional or mental problems by interacting with people in person and noticing how they act at work, school or in public. We could figure out how they feel from talking to them, watching their body language and interacting with them.

But then since people are interacting online more than in person this old way of finding mental distress isn't working anymore in many communities, including our community in the Middle East.

Today, social media and meta versa world where people chat with each other, share comments and respond to each other, express and share personal stories and so now or later we physical contact will be limited and people will be able to work, play, shop and even build emotional bonds completely online.

Whin this trend of the virtual world, online users generate huge amount of text every day, those text posts show their emotional and mental state.

and even

Checking for mental distress manually is difficult on such big scale of data and regular sentimental analysis often only sort into two categories; good or bad, it usually can't sort in more complicated states like depression, anxiety, stress, or bipolar disorder.

So, if we could find a way to check for mental distress automatically instead of checking manually it would be big advantage.

And building such smart, automatic system that can analyze text and figure out people's hidden emotional and mental health state is a need for today's circumstance of human interaction.

A system like this could help figuring out emotional patterns across online users and enhance research and understanding in digital mental health, especially where it's hard to see people or talk to them directly.

1.3 Thesis Objectives

The subject of thesis is to develop and evaluate transformer-based text classification models capable of detecting multiple mental health categories from textual datasets.

Objective steps:

- Preprocessing and preparing dataset "Sentiment Analysis for Mental Health" dataset from Kaggle.
- Fine tuning a RoBERTa model for classification.
- Evaluating model's performance using accuracy, precision, recall, and F1-score.
- -Analyzing classification results and model's limitations.

1.4 Thesis Contribution

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

From a technical perspective, this study develops a transformer-based natural language processing capable of detecting emotional and mental health states from textual data and evaluating it.

The developed system goes deeper than basic sentiment polarity classification and focuses on identifying more emotional categories

From a practical perspective, this research explores the ability of implementation AI-driven systems that are capable of state detection within digital service ecosystem, particularly in Middle East where government platforms adapt digital transformation widely.

Many countries in the region, such as the United Arab Emirates and the Kingdom of Saudi Arabia, have developed centralized digital platform for their citizens to facility issuing personal ID cards, passports, transactions such as buying selling, and registering contracts, etc.

Integration of such automated system for an emotional state detection module with government platform could potentially help in early risk identification and fast intervention strategies for individuals going through psychological distress. Although this thesis does not deploy in real-world governmental environment, it provides a foundational technical framework that could collaborate with digital public service for enhancing mental health awareness and early detection mechanisms.

Chapter 2 Background and Related Work

2.1 Introduction

Mental health disorders have been a significant health challenge globally and are affecting millions of individuals across the world that impose increasingly load on healthcare systems and society. The World Health Organization estimates that depression affects over 300 million people, while anxiety disorders impact around 264 million individuals.

Traditional methods of mental health assessment included clinical interviews, questionnaires consume time, and resources intensively.

In addition, mental health issues prevent individuals from asking for or seeking professional help and that leads to delayed diagnosis and intervention.

The increasing use of social media platforms has created huge data for researchers in field of mental health to support their studies.

Individuals express their thoughts, emotions and experiences via platforms such as Twitters, Reddit and Facebook, ang generating vast amounts of text that can indicate mental health status. This digital footprint provides researchers with data that reflects emotional states and psychological condition of the social media users.

Natural Language Processing NLP is a powerful tool for analyzing textual data and extracting patterns related to mental condition. Recent revolution in machine learning led to develop transformer-based language models such as BERT (Bidirectional Encoder Representation from Transformers) that has accurate and high reliability in detecting mental health problems including depression, anxiety, stress and suicidal ideation.

This chapter provides a comprehensive review of literature.

It starts by outlining the concepts of NLP in mental health and then dives into a critical analysis of previous work, exploring evolution from traditional machine learning to transformer-based models and explaining the need for explainable AI (XAI). Finally, it concludes by spotting a clear research gap that this thesis aims to fill.

2.2 Background: Natural Language Processing and Mental Health

The NLP combined with mental health is predicted on the “language as window” hypothesis, where the individual’s linguistic patterns can reflect their cognitive and emotional state (Pennebaker, 2011).

Early work in this domain was heavily relied on lexicon-based approaches, like the Linguistic Inquiry and Word Count (LIWC), which use set of predefined words categories (pronouns, affective words, etc.) to infer psychological states, but these methods lacked the ability to capture context, semantics and complex linguistics.

The NLP has revolutionized using machine learning (ML) and deep learning.

Traditional ML models like Support Vector Machine (SVMs) and Logistic Regression were trained on features such as n-grams, part of speech tags, and sentiment scores, represented a step forward to achieving more adaptable model. However, the most shift has emerged after implementation of transformer-based language models, like BERT and its variants (RoBERTa, MentalBERT, DistilBERT). Those models can understand complex linguistic structures, sarcasm, and expressions which are critical for accurate detecting mental health conditions from social media texts.

This thesis aims to leverage this advanced model for multi-class classification of mental health.

2.3 Evolution of Methodologies: From Traditional NLP to Transformer Models

The most recent literature is focused on comparison evaluation of traditional NLP methods against modern transformer-based models for mental health classification.

It demonstrates the superior performance of transformer models, although with increased computational cost.

In their study (Köckritz, 2025) “paper 11” provide a direct comparison in performance for multi-model evaluation on mental health classification and provide a benchmark for traditional models like Logistic Regression and SVMs against transformer models such as BERT, RoBERTa, and MentalBERT.

The author has concluded that traditional models offer faster and interpretable classifications, but transformer models, especially those pre-trained on domain specific data (like MentalBERT) achieve higher accuracy, and F1-scores, and have more ability to capture complex linguistic markers of mental health conditions.

Also, this finding is reflected in context of student mental health by (Pria et al.) “paper 3” who implement a RoBERTa model for sentiment analysis.

Their work demonstrates the efficiency of using large scale transformer models to achieve high accuracy in classifying on depression severity. And how to utilize transformer-based model to dive beyond binary classification (depressed vs non-depressed) and toward more accurate understanding of severity levels of depression, Their result reinforces the transformer’s architectures to capture gradations in depressive language.

2.4 The Need for Explainability in Mental Health AI

Transformer model has become more complex like “black boxes” and that why the need for explainable AI (XAI) has become a need. Especially in a sensitive domain like mental health where clinical trust and acceptance are critical.

A clinician or user is unlikely to trust model’s prediction without understanding the behavior and reasoning behind this prediction.

Jahan (Paper 02) and Kers (Paper 10) both deal with this critical issue.

Jahan focus on depression detection and proposed a framework that combines high performance model “Black box model” with XAI techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations).

And continue demonstrating that it is possible to maintain high accuracy by providing explanations by highlighting specific words or phrases in a post that most influenced the model's decision.

Kerz (2023) takes a broader approach by providing a narrative review on the application of XAI for mental health detection. They argue that NLP models to be viable in clinical setting they must be accompanied by methods that explain their decision-making processes, by enabling error analysis, bias detection and encourage user trust.

2.5 Methodological Trends and Reviews

Several papers contribute valuable knowledge by focusing deeply on specific topics, summarizing and reviewing the progress in the field.

Chatterjee "Paper 5" present a literature review focused on examination of depression, their work is a good resource for understanding the specific linguistic markers associated with depression, like increasing of using first-person singular pronouns, negative emotion word and decreasing engagement in social terms. This review highlights the evolution from dictionary-based methods to complex deep learning models for this specific condition.

Zhang "Paper 9" has provided a broader review as they in their narrative review on NLP applied to mental illness detection provides a high-level overview and discussed data sources (Twitter, Reddit), key tasks (depression, schizophrenia, suicide risk detection), and route of the model.

In a survey Zhang "Paper 6" introduce the concept of "emotion fusion" for mental illness detection. This is an advanced concept where multiple emotional signals (text, sentiment, and potentially emojis) are integrated to improve accuracy of the detection and go away from simple sentiment polarity.

Patel et "Paper 8" come up with a practical tutorial on implementation of ML and deep learning models for this task, and it considered as useful methodological reference.

Lastly, Muhammed "Paper 7" extend the debate to multilingual and multi-emotional detection, and spotted the challenges and opportunities in those models based on non-English-language data

2.6 Research Gap and Contribution of This Thesis

After reviewing previous work three important gaps remain:

1. Most studies detect only one mental health condition (e.g, depression) or two classes (depressed vs normal). A few models can classify several states at once, such as anxiety, depression, stress, and suicidal ideation.
2. Lack of explanation, even model works well it's like a black box. Users cannot understand why the model labeled some post as "anxious" instead of "stressed". This lack of transparency makes models less trustworthy for real world use,
3. There is no pipeline that combines multi-class classification with explanations. Existing tutorials provide classification or explanation but not both.

Thesis contributions:

To fill those gaps, the following simple and practical contributions will be provided:

- A model that classifies five mental states.
Will train a RoBERT model to label social media post such as: Anxiety, depression, stress, Normal, or Suicidal Ideation.

The model is trained on the public Kaggle dataset "Sentiment Analysis for Mental Health."

- Clear explanation for prediction: SHAP "a popular explainability tool" will be used to show and explain why some post was called "depressed" and not "anxious"
- Python modular scripts will be used for all steps, and the code will be shared for easily adapt to other datasets or to add more mental health categories.

- Evaluation will be reported contain accuracy, precision, recall and F1-score for each of the five classes, this will show where the model best performs and where struggles.

2.7 Summary of Literature Review

The following table summarizes the reviewed papers :

No.	Author(s) & Year	Title	Key Contribution / Focus	Methodology / Models	Key Findings / Limitations
1	Thomas Kallstenius ² , Andrea Johansson Capusan ¹ , Gerhard Andersson ^{3,4} & Adam Williamson (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	Sidra Hameed ¹ , Muhammad Nauman ¹ , Nadeem Akhtar ² , Muhammad A. B. Fayyaz ^{3*} and Raheel Nawaz ⁴ . (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	Hikmat Ullah Khan ^{1*} , Anam Naz ¹ , Fawaz Khaled Alarfaj ^{2*} and Naif Almusallam ² (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	Amna Qasim [†] , Gull Mehak, Nisar Hussain [†] , Alexander Gelbukh [*] and Grigori Sidorov (2023)	Detection of Depression Severity in Social Media Text	Moves beyond binary classification	Transformer-based models (e.g., BERT)	Successfully classifies depression severity, a nuanced and clinically relevant task.

		Using Transformer-Based Models	to detect severity levels of depression.		
5	Bazen Gashaw Teferral ¹ , BSc, MSc, PhD; Alice Rueda ¹ , PhD; Hilary Pang ¹ , MSc, MD; Richard Valenzano ² , PhD; Reza Samavi ³ , PhD; Sridhar Krishnan ³ , PhD; Venkat Bhat ^{1,4} , MSc, MD. (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	Shahidul Islam Khan (2023) FaisalBinAziz MdMisbah Uddin	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	Yeyubei Zhang ¹ , Zhongyan Wang ² , Zhanyi Ding ² , Yexin Tian ³ , Jianglai Dai ⁴ , Xiaorui Shen ⁵ , Yunchong Liu ¹ , and Yuchen Cao ⁵ (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	Tianlin Zhang ¹ , Annika M. Schoene ¹ , Shaoxiong Ji ² and Sophia Ananiadou (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	Elma Kerz ^{1*} , Sourabh Zanwar ¹ , Yu Qiao ¹ and Daniel Wiechmann (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	Jannis Köckritz ^{1,2} , Bahar İlgen ¹ , Caroline Cohrdes ³ & Georges Hattab (2024)	Current applications and future directions in natural language processing for news media and mental health	Explores NLP applications for mental health in the context of news media.	Potentially topic modeling and sentiment analysis	Extends the application domain to news media, suggesting new data sources and research directions.
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)	Provides explainability via symptom-based justifications. Validated on social media data with strong results. Limitations include computational cost of LLMs.

				classifier for prediction.	
--	--	--	--	-------------------------------	--

Chapter 4 Proposed Approach (Methodology)

Chapter 5 :Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References