NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Developing AI for Mental Health

Recognizing Psychological Disorders from Social Media Posts

Helena-Leila Marie Mashayekhi

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Developing AI for Mental Health:

Recognizing Psychological Disorders from Social Media Posts

by

Helena-Leila Marie Mashayekhi

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Vitor Santos, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 2025]

Helena-Leila Marie Mashayekhi

DEDICATION

This thesis is dedicated to my father who made all of this possible through his hard work to make everything possible for me and his constant belief in me. And also to my brother who inspired and encouraged me to start this master's degree in the first place.

ABSTRACT

In recent years, mental health issues have surged globally, while on the contrary, the availability of qualified professionals has not kept pace, hindering timely and comprehensive care. Consequently, many psychological disorders go undetected, often resulting in poorer long-term outcomes. To address this gap, this study investigates the impact of RAG-based classification on detecting psychological disorders from X (formerly Twitter) posts. Therefore, this study formulated both binary (disorder vs. no disorder) and multiclass (three distinct psychological states) classification tasks and compared baseline LLMs, few-shot RAG-based classification and a RAG-based classification setup. Tweets were used to retrieve relevant external context from a vector database, which, together with few-shot examples, was included in a prompt for LLM-based classification. Results indicated that binary classification outperforms multiclass tasks, with macro-F_1 scores up to 0.95. Document retrieval alone produced mixed effects while it improved suicide detection but reduced performance on depression. However, combining it with few-shot prompting led to consistent gains across all tasks. Few-shot prompting reduced classification error rates by up to 5.6 percentage points, especially for challenging cases with subtle or ambiguous signals. However, multiclass classification remained difficult due to lexical overlap between depressive and suicidal language, with macro-F_1 scores plateauing around 0.76-0.78 even in the best configurations. These findings suggest that while RAG-based prompting improves classification performance, especially when augmented with few-shot examples, further refinement is needed for reliable multiclass mental health screening.

KEYWORDS

Disorder Detection; Retrieval-Augmented Generation; Large Language Model; Clinical Psychology

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BI-LSTM	Bidirectional Long-Short Term Memory
CBT	Cognitive Behavioural Therapy
CNN	Convolutional Neural Networks
DSM-5	Diagnostic and Statistical Manual of Mental Disorders (5th ed)
FNR	False Negative Rate
GenAI	Generative Artificial Intelligence
GAD	General Anxiety Disorder
GPT	Generative Pre-Trained Transformers
GRU	Gated Recurrent Unit
ICD-11	International Classification of Diseases 11th Revision
LLM	Large Language Model
LM	Language models
LSTM	Long-Short Term Memory
MDD	Major Depressive Disorder
ML	Machine Learning
NLP	Natural Language Processing
NSSI	Non-Suicidal Self-Injury
PPV	Positive Predictive Value
RAG	Retrieval-Augmented Generation
SVM	Support Vector Machine
UNICEF	United Nations International Children's Emergency Fund
US	United States
WHO	World Health Organization

1. INTRODUCTION

1.1. CONTEXT AND PROBLEM IDENTIFICATION

The global mental health crisis has put significant pressure on healthcare systems worldwide. According to the World Health Organization (2022), there has been a marked increase in mental health conditions, placing unprecedented demands on services already strained by chronic underfunding and limited human resources. In the United States, for instance, a 2022 survey by the American Psychological Association found that 60% of psychologists were unable to accept new patients, and two-thirds observed more severe symptoms among those they were treating (APA, 2022).

This crisis extends beyond adult populations. In the US, suicide is ranked the second leading cause of death among adolescents aged 10 to 14, highlighting the urgency for early interventions (Centers for Disease Control and Prevention, 2022). Researchers have also moved their attention to the potential impact of excessive social media use on young people's mental health. For example, Twenge and Campbell (2018) identified correlations between frequent social media engagement and elevated risks of depression, anxiety, and sleep disruption in adolescents.

While organizations such as UNICEF have called for greater protections for children online, including policy-level interventions to ensure safer digital environments, there remains a lack of consensus on how best to regulate access to social media platforms (UNICEF, 2021). Compounding these issues is a global shortage of mental health professionals, increasing the difficulties in providing timely and comprehensive care (World Health Organization, 2022). Although telehealth solutions have improved access, particularly for individuals in underserved or rural areas, these services have not resolved staffing shortfalls (Molfenter et al., 2021).

Consequently, there has been a surge in the use of digital mental health tools, such as mobile apps offering CBT exercises, mood tracking, and mindfulness resources (Powell et al., 2020). These applications enable on-demand and location-independent support without immediate clinician involvement (Torous et al., 2018). Despite their accessibility, user retention remains an ongoing challenge, as engagement tends to drop significantly over time (Torous et al., 2018).

Recently, AI, particularly LLMs, has emerged as a potential tool for scalable mental health screening and support. Studies suggest that LLMs, including ChatGPT, can assist in identifying early indicators of mental health disorders by analysing language in clinical notes, online forums, and social media posts (Montejo-Ráez et al., 2024). However, concerns remain around the reliability of these models, especially in the light of "hallucinations", where outputs include inaccurate or misleading information (Rane et al., 2023).

One approach that shows promise is RAG. By fusing retrieval-based and generative models, RAG can produce contextually rich and more accurate responses (Zhao et al., 2024). A related concept is RAG-based classification which supports similar mechanisms to enhance classification tasks. This architecture mitigates the production of spurious or misleading content by grounding the model's output in a specific retrieval process (Ghali et al., 2024). For mental health applications, such an approach could help in analysing social media posts to detect signals of psychological distress or disorders, thereby improving the precision and reliability of automated detection tools (Zhao et al., 2024).

Although the potential of AI-driven, RAG-based classification solutions for mental health is increasingly recognized, research focusing specifically on social media contexts remains limited. Given the urgency of the mental health crisis particularly among younger demographics and the ongoing challenges of staff shortages and telehealth limitations, there is a clear incentive to investigate how these innovative approaches might aid in early detection and leading the path for early interventions.

Therefore, the primary research question guiding this thesis is:

What is the impact of Large Language Models (LLMs) combined with Retrieval-Augmented Generation (RAG) based classification on the detection of psychological disorders in text data from X (formerly Twitter)?

To address this question, this study develops an AI model proficient in identifying psychological disorders from conversational data, with a specific emphasis on social media posts. By examining how RAG-based classification can enhance detection accuracy, the findings aim to contribute both to scientific work in AI-driven mental health diagnostics and to practical interventions that support digital platforms for early identification and support.

1.2. OBJECTIVES

The goal of the research is to find out if the RAG-based classification can reliably recognize psychological disorders from social media texts.

In order to achieve the main goal, the following objectives were defined:

- Analyse the intersection of mental health diagnostic and GenAI
- Integrate retrieval-based document augmentation into psychological text classification models
- Evaluate multiple LLM and RAG-based classification approaches to identify the most effective configurations
- Conduct empirical testing on annotated social media datasets to assess performance in detecting distinct psychological disorders

1.3. STUDY RELEVANCE AND CONTRIBUTIONS

The outcome of this research holds significant societal value, particularly within the fields of healthcare, mental health and well-being. Social media offers individuals a powerful platform for self-expression and learning; however, it also presents risks such as reinforcing negative emotional spirals (Naslund et al., 2020; Twenge et al., 2018).

This thesis aims to mitigate these risks by enabling early detection of psychological states and disorders from conversational data. By aiding with timely recognition and intervention, this research addresses an important gap in mental health care.

Furthermore, a successful implementation of this RAG model can lay the groundwork for AI-supported therapeutic interventions. Although AI is not expected to replace human therapists, it could serve as an adjunctive tool in therapy by reliably interpreting psychological language. This could enhance accessibility to mental health services, especially in underserved regions or among individuals hesitant to pursue traditional therapy.

Additionally, incorporating AI into the mental health assessment process could reduce the stigma associated with psychological disorders and psychotherapy. By increasing accessibility and normalizing the use of mental health diagnostic tools, individuals may become more inclined to seek assistance, further promoting mental well-being.

Scientifically, this research advances the field of computational psychology by applying LLMs and RAG-based classification techniques to analyse unstructured conversational data from social media. Unlike previous studies that mainly rely on structured inputs, this approach uses LLMs and their flexibility to process natural language. This study could provide new insights into the linguistic markers of psychological states, contributing to a deeper understanding of how language reflects mental health.

2. LITERATURE REVIEW

This chapter reviews the scientific background that supports this study drawing on relevant literature and related use cases. It begins with a review of the literature on mental disorder, then turns to developments in generative AI in the health sector and finally concludes with a systematic review that integrates the most recent research relevant to mental health detection on social media.

In the following sections, the acquired knowledge is presented through a structured and comprehensive review.

2.1. Mental Disorders

2.1.1. Overview

Mental health disorders are prevalent globally, affecting nearly half of the population by the age of 75, according to McGrath et al. (2023). Typically emerging in childhood, adolescence, or young adulthood, these disorders can profoundly affect cognitive, emotional, and behavioural domains, leading to impairments in personal, social, and occupational functioning (McGrath et al., 2023; Hyman et al., 2006). Severe mental disorders are strongly associated with increased suicide risk, with psychiatric disorders notably raising suicide attempts by sixteen times compared to the general population (Sutar et al., 2023). Timely intervention and early detection, however, can significantly reduce impairments and suicide risks (Stene-Larsen & Reneflot, 2019).

Social media platforms, such as X (formerly Twitter), provide abundant textual data through which psychological distress can be analysed early, offering opportunities for intervention before conditions worsen (Guntuku et al., 2017). AI-driven methods, particularly LLMs coupled with RAG-based classification, show promise in accurately detecting early signs of mental disorders from online textual interactions (Kermani et al., 2025).

2.1.2. Primary Disorders

Common mental health disorders include anxiety disorders (such as GAD), mood disorders (such as MDD), substance use disorders, and eating disorders (American Psychiatric Association, 2013). Each of these disorders manifest through a range of emotional, cognitive, and behavioural symptoms, impacting the daily lives of individuals affected (McGrath et al., 2023; Hyman et al., 2006).

MDD is characterized by persistent low mood, reduced interest in activities, disrupted sleep, changes in appetite, fatigue, feelings of worthlessness, and recurrent suicidal ideation (American Psychiatric Association, 2013). McGrath et al. (2023) identified depression as notably prevalent across genders, with additional specific patterns such as alcohol use disorder prominent in men and specific phobias in women. On social media, depression

manifests through reduced engagement, late-night postings, self-disclosure, negative sentiment and frequent mentions of symptoms and medication use (De Choudhury et al., 2013; Renjith et al., 2022).

While depression is a mood disorder, suicide is defined as a self-directed act in which an individual consciously intends to be fatal (Hawton & James, 2005). It is distinct from NSSI, which involves deliberate self-harm without intent to die; however, engaging in NSSI significantly elevates the risk of future suicide attempts (Hawton & James, 2005; Morgan et al., 2017; Hawton et al., 2012). In addition to individual risk factors, exposure to suicide - whether through personal connections, peer behaviour, or media reports - can increase suicidal behaviour in vulnerable populations, a phenomenon known as “suicide contagion” (Gould et al., 2003; Turecki et al., 2019).

2.1.3. Diagnostic Approach

Mental disorder diagnosis typically relies on clinical interviews, self-reports, and criteria from the DSM-5 or ICD-11 (Stein et al., 2022). DSM-5 criteria for conditions such as depression or anxiety depend on symptom presence and duration (American Psychiatric Association, 2013; Leichsenring et al., 2022).

Emerging methodologies support AI and NLP in analysing linguistic markers indicative of mental health disorders (e.g., the use of negative emotion words and first-person pronouns) to facilitate the early identification of psychological distress (Guntuku et al., 2017). Additionally, Hyman (2010) criticizes the DSM-5 for prioritizing reliability over biologically valid diagnostics. RAG-based classification methods mitigate this by integrating external knowledge bases to provide richer, contextually informed diagnostics (Hurtado, 2023).

2.2. GenAI

2.2.1. Concepts

GenAI, specifically LLMs such as GPT-3 and GPT-4, generates new textual content based on learned patterns from extensive textual datasets (Jovanovic & Campbell, 2022; Kumar et al., 2024). Transformers, the core architecture of LLMs, utilize self-attention mechanisms that enable them to capture contextual dependencies effectively, which is crucial for accurately interpreting the text (Vaswani et al., 2017; Kalyan, 2024).

Transformers are the backbone of LLMs consisting of encoders that process the input and decoders that generate the output (Kalyan, 2024). The key innovation of transformers is the self-attention mechanism, which allows the model to focus on different parts of the input text and handle long-range dependencies (Kalyan, 2024). Self-attention allows transformers to weigh the importance of different words in a sentence relative to one another, enabling models to understand context more effectively (Vaswani et al., 2017).

LLMs work by processing large corpora of text and learning the underlying language patterns (Kumar et al., 2024; Jeong, 2023). The core principle is to predict the next word in a sequence based on the previous ones, allowing the generation of coherent text (Kumar et al., 2024; Jeong, 2023). A challenge of LLMs is the hallucination problem, where LLMs generate plausible but incorrect information (Jeong, 2023). Methods like fine-tuning and RAG mitigate these errors (Jeong, 2023).

RAG-based classification combines retrieval systems with generative models, enhancing classification accuracy by retrieving contextually relevant external data before generating predictions (Jeong, 2023; Lewis et al., 2020). Forward-Looking Active Retrieval, an advanced retrieval augmentation approach, continually retrieves external documents during the entire classification process, increasing contextual accuracy (Jiang et al., 2023).

2.2.2. Developing Models for Psychological Disorder Detection

LLMs like GPT-3 and GPT-4 undergo pre-training on large corpora of text data, allowing them to learn the nuanced linguistic patterns and structures that signify natural language (Kalyan, 2024). Fine-tuning these models with targeted psychological datasets enables the recognition of linguistic markers associated with mental health issues, such as negative sentiment and first-person pronoun usage indicative of depression or anxiety (Yu & McGuinness, 2024).

RAG-based classification models employ tokenizers, transformers, retrievers, and classifiers to achieve accurate classification of psychological disorders from textual data. The retriever accesses relevant psychological literature or clinical information, allowing the classifier component of RAG-based classification to make informed and accurate predictions (Wolf et al., 2020; Lewis et al., 2020).

Instagram (Meta), for example, use convolutional neural networks to spot self-harm imagery (e.g. cutting) and BERT-style classifiers to flag captions suggesting suicidal ideation (Scherr et al., 2020). They decided to remove all graphic images of self-harm or related content (Instagram, 2019). Additionally, Facebook (now Meta) introduced machine learning structures and an NLP pipeline that scans posts and comments for high-risk language, triggering resources and, in severe cases, escalating to human review teams (Muriello et al., 2018).

2.2.3. The Role of RAG-based classification in Mental Health Detection

AI is also transforming healthcare by contributing to medical imaging, drug discovery, treatment planning, and clinical decision support (Kuzlu et al., 2023). AI has shown promising results in the early detection of diseases like breast cancer, skin cancer, eye disorders, and pneumonia using imaging technologies (Ahammad et al., 2020; Abdollahi et al., 2022; Esteva et al., 2017). However nowadays, AI is also able to analyse speech patterns to predict psychotic episodes (Kaywan et al., 2023; Bedi et al., 2015).

AI, specifically RAG-integrated LLMs, contributes significantly to healthcare through accurate early diagnosis, personalized treatment plans, and improved outcomes (Kuzlu et al., 2023).

RAG-based classification methodologies demonstrated substantial improvements over conventional classification models, offering capabilities previously reserved for advanced models like GPT-4 (Xiong et al., 2024).

Therefore, deploying LLMs and RAG-based classification models for mental health detection holds promise for automated, scalable mental health assessments from social media data, particularly in X (formerly Twitter).

2.3. GenAI in the detection of Mental Disorders

The following part takes advantage of the PRISMA protocol. PRISMA stands for "Preferred Reporting Items for Systematic Reviews and META-Analysis". It consists of a four-phase protocol and a 27 items checklist that helps guide systematic literature reviews (Moher, 2009; Page et al., 2021).

2.3.1. PRISMA Protocol

The PRISMA checklist ensures comprehensive and transparent reporting of a systematic review, while the four steps (Identification, Screening, Eligibility, Inclusion) guide the process of systematically identifying and selecting relevant studies for inclusion. Both the checklist and steps are essential for ensuring a strict and transparent systematic review (Moher, 2009).

The steps can be understood in the following way:

1. **Identification**: Relevant studies are identified using a systematic search strategy across key academic databases to compile a comprehensive pool of potential sources.
2. **Screening**: The identified studies are then reviewed against predefined inclusion and exclusion criteria to filter out those that do not meet basic relevance and quality standards.
3. **Eligibility**: The remaining studies undergo a detailed evaluation to ensure alignment with the research objectives and methodological rigor.
4. **Inclusion**: Finally, the studies that meet all criteria are included for full analysis and form the basis of the research.

2.3.2. PRISMA Execution

In this section the focus is moved from the theoretical background to the execution of the PRISMA model. This review investigates the following questions for the systematic review of the literature concerning GenAI and the detection of mental disorders.

Table 1 – Systematic review's research questions

SLRQ1	What are the current methodological trends in detecting mental health conditions from social media using AI?
SLRQ2	How are GenAI techniques (e.g. RAG) applied in this domain, and what tasks do they most effectively support?
SLRQ3	What are the advantages and disadvantages of applying RAG-based classification techniques in this field?

The following keywords were used to find the relevant articles to answer these questions.

Table 2 – Systematic review's keywords

	GenAI	**Mental Disorder**	**Social Media**
Keywords	Large Language Model	Mental Disorder	Social Media
	Knowledge-enhanced	Anxiety	Twitter
	Machine Learning	Depression	Reddit
	Retrieval-Augmented	Mental Health	X
	Retrieval-Based Question Answering	Disorder Detection	
	Text Classification	Suicide	

Based on the search strings listed in Table 2, a targeted search string was developed to incorporate relevant terms and keywords, aiming to locate them within the abstracts, titles, or keywords of scientific articles and papers. Given the rapid advancements in the field, a specific focus was placed on recent publications. To ensure up-to-date and accurate information on the use of GenAI in detecting mental disorders, a filter was applied to include only articles published between 2020 and 2025. Therefore, the search string used was: **("Large language model" OR "Knowledge-enhanced" OR "Machine Learning" OR "Retrieval-Augmented" OR "Text classification" OR "Retrieval-Based Question Answering") AND ("Mental disorder" OR "Anxiety" OR "Depression" OR "Mental Health" OR "Disorder Detection" OR "Suicide") AND ("Social Media" OR "Twitter" OR "Reddit" OR "X").**

The search was conducted in March 2025 on the following scientific information resource databases:

Table 3 – Systematic review's resource databases

Resource Database	Resource URL
Web of Science	https://www.webofknowledge.com/
ACM	https://dl.acm.org
IEEE	https://ieeexplore.ieee.org
Google Scholar	https://scholar.google.com
Science Direct	https://sciencedirect.com

Following the PRISMA methodology, the subsequent step involved establishing the inclusion and exclusion criteria for the articles retrieved through the search process.

Table 4 – Systematic review's inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Each paper must be a paper written in English	Articles not in English and duplicate papers
Paper is published between 2020 and 2025	Articles published before 2020
	Non-academic or non-scientific papers (e.g., websites, magazines reports, newspapers, consulting articles, books, citations)
	Papers with titles outside the scope of this work

Following the PRISMA methodology, a series of steps were undertaken to refine the pool of relevant articles. Initially, the search led to a total of 2,189 articles across the selected databases. After removing duplicates, the dataset was reduced to 1,537 unique records. Next, the inclusion and exclusion criteria were applied, and articles that could not be accessed were excluded, resulting in 598 eligible articles.

Subsequently, a title screening process was conducted to ensure alignment with the research objectives, further narrowing the selection to 170 articles. A more in-depth review of abstracts followed, which reduced the number to 86 articles deemed potentially relevant. Finally, a

thorough evaluation of the full texts is done to determine the final set of studies to be included in the analysis. This led to a total of 20 studies.

This leaves the review at a number of 20 selected articles. These selected articles are presented in Table 5.

Table 5 – PRISMA results table - included articles

#	Authors	Article	Contribution	Publication Type
[1]	Alghazzawi et al. (2025)	Explainable AI-based suicidal and non-suicidal ideation detection from social media text with enhanced ensemble technique.	Proposed an ensemble of transformers (RoBERTa, ALBERT, etc.). These transformers were combined with SHAP interpretation to classify posts into suicidal vs non-suicidal.	Journal Article
[2]	Buddhitha & Inkpen (2023)	Multi-task learning to detect suicide ideation and mental disorders among social media users.	This study introduced a multi-task transformer identifies users with a mental disorder that were self-reported and those with suicidal ideation. It uses Twitter and Reddit datasets.	Journal Article
[3]	Cha et al. (2022)	A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community	Examined a two-step approach for early-stage depression detection. First, a depression post-classification model was proposed using multiple languages Twitter datasets. Moreover, a depression lexicon was built for each language, which mental health experts verified.	Journal Article
[4]	Ghanadian et al. (2023)	ChatGPT for Suicide Risk Assessment on Social Media: Quantitative Evaluation of Model Performance, Potentials and Limitations	This study evaluates ChatGPT in the context of suicide assessment from Reddit data. Results show that fine-tuned transformer-based models performed better than ChatGPT.	Workshop Paper
[5]	Guo et al. (2024a)	Evaluating large language models for health-related text classification tasks with public social media data.	This study systematically tested GPT-3.5 & GPT-4 against fine-tuned transformers on 6 social media classification tasks.	Journal Article
[6]	Guo et al. (2024b)	Large language models for mental health	Reviewed 40 studies that used LLMs in mental health. The study	Systematic Review

#	Authors	Article	Contribution	Publication Type
		applications: Systematic review.	noted that 15 studies (38%) since 2020 focused on social media mental health or suicide detection. The key findings were that transformer models are effective in detecting mental health issues but also pose risks (e.g. inconsistent outputs, lack of interpretability).	
[7]	Hossain et al. (2025)	Multi-task opinion enhanced hybrid BERT model for mental health analysis.	This study proposed Opinion-BERT. It combines BERT with sentiment and opinion features for simultaneous sentiment analysis and mental status classification on social media posts.	Journal Article
[8]	Kerasiotis et al. (2024)	Depression detection in social media posts using transformer-based models and auxiliary features.	Proposes a DistilBERT-based architecture augmented with user metadata and linguistic markers for Reddit depression detection.	Preprint
[9]	Kermani et al (2025)	A Systematic Evaluation of LLM Strategies for Mental Health Text Analysis: Fine-tuning vs. Prompt Engineering vs. RAG (CLPsych 2025)	Systematically compare RAG against zero-/few-shot prompting and supervised fine-tuning on emotion and mental-health condition tasks (DAIR-AI and SWMH datasets).	Workshop Paper
[10]	Lin et al. (2020)	SenseMood: Depression detection on social media.	Demonstrates a multimodal approach for depression detection. It uses a Twitter dataset where users' tweets and profile images are jointly analysed.	Conference Paper
[11]	Mazhar et al. (2025).	Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification (WWW 2025)	Retrieve figurative and commonsense knowledge (via ConceptNet/COMET) to enrich a multimodal meme representation.	Conference Paper
[12]	Metzler et al. (2022)	Detecting potentially harmful and protective suicide-related content	Uses BERT and XLNet to classify tweets into 12 categories of suicide-related content (e.g.	Journal Article

#	Authors	Article	Contribution	Publication Type
		on Twitter: Machine learning approach.	personal suicidal ideation, recovery, news reports, jokes).	
[13]	Muraka et al. (2021)	Classification of mental illnesses on social media using RoBERTa.	First known multi-class transformer model for mental illness detection on Reddit.	Workshop Paper
[14]	Ogunleye et al. (2024)	Sentiment informed sentence-BERT ensemble algorithm for depression detection	Proposes an ensemble of Sentence-BERT classifiers augmented with sentiment features to detect early-stage depression in a Twitter dataset.	Journal Article
[15]	Qasim et al. (2025)	Detection of depression severity in social media text using transformer-based models.	Focuses on classifying depression severity levels (mild, moderate, severe) from Reddit posts. Compares content-based n-gram features versus context-based Sentence-BERT embeddings and then fine-tunes three transformer models.	Journal Article
[16]	Ravenda et al. (2025)	Are LLMs Effective Psychological Assessors? Leveraging Adaptive RAG for Interpretable Mental Health Screening through Psychometric Practice (2025)	Their adaptive RAG approach retrieves the most relevant user posts to answer each item of a standardized psychological questionnaire (e.g., BDI-II).	Preprint
[17]	Vajre et al. (2021)	PsychBERT: A mental health language model for social media mental health behavioral analysis.	Introduces PsychBERT, a BERT model pre-trained on ~1.2M mental health-related posts (Reddit and mental health forum data). When fine-tuned on downstream tasks PsychBERT outperformed general BERT.	Conference Paper
[18]	Wang et al. (2020)	Depression risk prediction for Chinese microblogs via deep-learning methods: content analysis.	One of the first studies applying BERT to mental health in Chinese. Uses posts from Weibo labelled with a 3-level depression risk.	Journal Article
[19]	Yang et al. (2024)	MentalLLaMA: Interpretable mental health analysis on social media with large language models	Introduces an interpretable LLM-based classifier fine-tuned from LLaMA-7B on mental health data.	Conference Paper

#	Authors	Article	Contribution	Publication Type
[20]	Yang et al. (2022)	A mental state knowledge-aware and contrastive network for early stress and depression detection on social media.	This paper exemplifies retrieval-augmented ideas: it retrieves commonsense knowledge about speakers' mental states using the COMET knowledge base and infuses this into a dual-input GRU-based classifier to detect early stress and depression.	Journal Article

DISCUSSION OF THE RESULTS

SLRQ1 - What are the current methodological trends in detecting mental health conditions from social media using AI?

Recent research has shown a growing emphasis on applying LLMs to detect mental health conditions using social media data (Guo et al., 2024b). The trend has shifted significantly from traditional models toward more advanced transformer-based methods.

Earlier methods predominantly relied on sparse feature extraction techniques, such as TF-IDF, combined with classical classifiers, including SVM and logistic regression. These techniques, although computationally efficient, cannot capture semantic and contextual nuances in language (Metzler et al., 2022; Alghazzawi et al., 2025). The field evolved with the adoption of deep learning architectures, such as CNNs and LSTMs, which enabled models to learn more abstract linguistic features and better capture emotional nuances (Buddhitha & Inkpen, 2023).

Currently, transformer-based models, such as BERT, are prevalent in this domain. Early works indicate that fine-tuned BERT and its variants significantly outperform traditional classifiers. For example, Wang et al. (2020) applied BERT, RoBERTa, and XLNet to Chinese microblog posts achieving a micro-F_1 of 0.856 for depression detection, which is substantially better than earlier CNN/LSTM approaches. Similarly, Metzler et al. (2022) demonstrated that BERT and XLNet can reliably categorize suicide-related tweets into nuanced risk categories, with an accuracy rate of around 73% across six content types.

Overall, transformer models fine-tuned on social media data have become state-of-the-art, enabling the detection of depression, anxiety, PTSD, and other conditions with far higher accuracy than methods before 2018 (Qasim et al., 2025). Researchers have expanded to multiclass classifications - e.g. Murarka et al. (2021) introduced a RoBERTa-based model that classifies posts into five mental illness categories (depression, anxiety, bipolar, ADHD, PTSD) using 17k Reddit posts. There is also increasing focus on multilingual and cross-platform detection. Cha et al. (2022) explored a lexicon-assisted model using X (formerly Twitter) data in Korean, English, and Japanese, revealing effective early detection of depression across languages.

In summary, current research is vibrant, with numerous studies fine-tuning LLMs (e.g., BERT, RoBERTa) on platforms such as Reddit, X (formerly Twitter), and others to detect mental health issues.

SLRQ2 - How are GenAI techniques (e.g. RAG) applied in this domain, and what tasks do they most effectively support?

While fine-tuned transformers have established themselves as high-performing solutions in this space (e.g., Quasim et al., 2025; Vajre et al., 2021), there is increasing interest in leveraging more flexible or interpretable GenAI methods, such as Retrieval-Augmented Generation (RAG), hybrid sentiment models, and multi-task transformers. Domain-specific models like PsychBERT have also emerged. Vajre et al. (2021) introduced PsychBERT, a pre-trained model on mental health forums and literature texts, which outperformed CNN/LSTM baselines (F_1 = 0.63 vs. 0.51) in classifying mental health statuses (Vajre et al., 2021). Similarly, Quasim et al. (2025) reported that RoBERTa outperformed BERT and DeBERTa in depression severity classification, achieving an F_1 score of 0.91.

Researchers are also exploring multi-task learning and sentiment-informed models. Buddhitha and Inkpen (2023) proposed a multi-task transformer that jointly detects mental disorders and suicide ideation, yielding strong performance by leveraging shared latent features. Hossain et al. (2025) developed Opinion-BERT, which integrates sentiment and opinion features, reaching 94.22% classification accuracy. By jointly classifying sentiment and user mental state, their model achieved 94.22% accuracy in mental status categorization, outperforming vanilla BERT and RoBERTa (Hossain et al., 2025). These hybrid approaches show that combining affective and clinical features can enhance performance and interpretability.

Cross-modal methods have further expanded the design space. Lin et al. (2020) presented SenseMood, a system that combines tweet text with image analysis to detect depression on X (formerly Twitter). Such cross-modal approaches reflect that depressed users' posted images can provide complementary signals, which enhance detection robustness.

Large generative models are also emerging. Ghanadian et al. (2023) evaluated ChatGPT (GPT-3.5) for suicide risk assessment on social media, finding it can classify risk with promising recall and outperform baseline SVMs in zero-shot settings (Gui et al., 2024a). However, fine-tuned models still had an edge in precision. Kermani et al. (2025) systematically compared fine-tuning, RAG, and few-shot prompting, finding that fine-tuned models still yielded the highest classification accuracy (91%), with RAG trailing at 40–68%.

Another advancement is interpretable LLMs. Yang et al. (2024) introduced MentalLLaMA, a fine-tuned LLaMA-7B model for interpretable symptom identification, showing how transparency can be integrated into deep classifiers.

Ensemble techniques are yielding top performance. Alghazzawi et al. (2025) combined multiple transformers in an explainable ensemble, achieving an impressive F_1 score of 95.5%

for suicidal ideation, compared to 99% for non-suicidal content (Alghazzawi et al., 2025), on several benchmark datasets. In summary, current GenAI techniques range from fine-tuned transformers (BERT, RoBERTa, GPT-3, GPT-4) to hybrid models (combining LLM embeddings with sentiment, multimodal data, or multi-task learning). These approaches consistently report high accuracy and F_1 (often +0.85), showing the effectiveness of GenAI in parsing language for mental health signals.

Overall, GenAI models are diversifying, with promising results from RAG, sentiment augmentation, and multi-task setups. These methods offer more adaptability and explanation potential but are not yet consistently outperforming fine-tuned baselines in raw accuracy.

SLRQ3 - What are the advantages and disadvantages of applying RAG-based classification techniques in this field?

Although few studies have explicitly applied RAG-based classification pipelines to social media mental health detection, the emerging evidence suggests both opportunities and challenges. RAG offers a promising alternative by integrating external knowledge into the classification process, potentially enriching sparse or ambiguous social media text (Kermani et al., 2025). Yang et al. (2022) introduced KC-Net, which retrieves commonsense mental-state knowledge from COMET and integrates it into a contrastive model. This approach improved the accuracy of stress and depression detection by modelling implicit psychological context. RAG-based classification can provide richer context that a standalone post may lack.

A major advantage of RAG is its flexibility and interpretability. Ravenda et al. (2025) introduced Adaptive RAG to answer items from psychological questionnaires by retrieving the most relevant user posts, showing the potential for personalized, psychometrically grounded analysis. Similarly, sentiment-augmented approaches (e.g., Hossain et al., 2025) can explain predictions via recognizable emotion signals.

On the other hand, there are disadvantages and challenges. Kermani et al. (2025) found that RAG-based models had lower classification accuracy (40–68%) than fine-tuned transformers, highlighting difficulties in effective retrieval, document relevance scoring, and integration. Moreover, RAG pipelines are computationally more complex and often require well-curated retrieval corpora to succeed.

Generalization and robustness are theoretical benefits of RAG, its ability to access updated knowledge on-the-fly can help models remain effective even as user language evolves. But this benefit is only realized if the retrieval mechanism is accurate and the external content is relevant and trustworthy. If retrieval is noisy or too narrow, it can degrade performance (Mazhar et al., 2025).

Finally, privacy concerns must be considered. Retrieving external posts or personal user data, even for improved context, may raise ethical flags in mental health analysis, especially when posts are highly sensitive (Kermani et al., 2025).

In practice, few social media mental health studies explicitly deploy RAG-based classification pipelines yet; instead, many use static knowledge integration (e.g., training on combined data sources or adding expert lexicons) (Cha et al., 2022). This shows that RAG-based classification holds promise for improving context awareness and interpretability in AI-based mental health systems.

3. METHODOLOGY

3.1. OVERVIEW

The objective of this study was to develop a RAG-based classification model that accurately identifies psychological disorders. The research was structured into four main phases: exploration, conceptualization, exploratory setup, and conclusive analysis. The initial phase involved a systematic literature review to understand current advancements in RAG-based classifications and their components for mental health applications, identifying gaps and laying the theoretical foundation for future research. Key models used were GPT-4o-mini and Mistral-7B-Instruct-v0.1, chosen for their effectiveness.

3.2. CONCEPTUALIZATION

3.2.1. RQ AND HYPOTHESIS & TASK DEFINITION

The systematic review of recent work in mental health detection from social media revealed that fine-tuned transformer models consistently achieve state-of-the-art performance across a variety of tasks, including binary classification (e.g., suicidal vs. non-suicidal), severity estimation, and multiclass mental disorder detection (e.g., Alghazzawi et al.; 2025; Qasim et al.; 2025; Vajre et al.; 2021; Kermani et al., 2025). While these results establish fine-tuning as a robust model, they also highlight a relative under-exploration of more flexible or interpretable alternatives, such as RAG and few-shot prompting.

Specifically, RAG has been addressed in only a handful of studies in that field (e.g., Ravenda et al., 2025; Yang et al., 2022), most of which emphasize its explainability and adaptability rather than its raw classification performance. Meanwhile, few-shot prompting - though resource-efficient and promising in general, NLP - is rarely evaluated within the mental health domain. This suggests a research gap regarding the comparative potential of these techniques, especially when applied to nuanced psychological tasks that benefit from context-rich reasoning or domain-specific knowledge grounding.

Consequently, this study compares baseline LLMs, RAG and few-shot prompting strategies for mental health classification. The aim is not only to benchmark performance across these methods but also to explore whether RAG can successfully perform in comparison a baseline LLM. Alternatively, RAG and few-shot can even offer complementary strengths - e.g., interpretability, sample efficiency, or generalizability - relative to traditional fine-tuning approaches. By exploring these less thoroughly studied paradigms, the study contributes to diversifying methodological pathways for scalable, robust, and ethically grounded mental health AI systems.

The central question driving this thesis is: How does the integration of RAG-based classification with the state-of-the-art LLMs affect the accuracy of psychological disorder detection in X (formerly Twitter) posts?

To investigate this, three classification tasks of increasing complexity were defined:

Binary Task A - Distinguish Suicidal from Normal posts

Binary Task B - Distinguish Depression from Normal posts

Three-way Task - Distinguish Depression, Suicidal and Normal posts simultaneously

By structuring the study in this way, it is demonstrated how model performance evolves as the diagnostic distinctions become more fine-grained. Grounded in the intuition that coarser distinctions are easier to learn, the following hypotheses are expected:

H_1: The suicidal vs standard classifier will achieve the highest accuracy and F_1-score, reflecting the model's ability to capture the most pronounced linguistic signals.

H_2: The depression vs standard classifier will perform slightly below the suicidal binary task, as depressive language patterns can overlap more subtly with normative expressions.

H_3: The ternary classifier will yield the lowest overall performance due to the added complexity of simultaneously disambiguating three classes and the increased potential for label confusion.

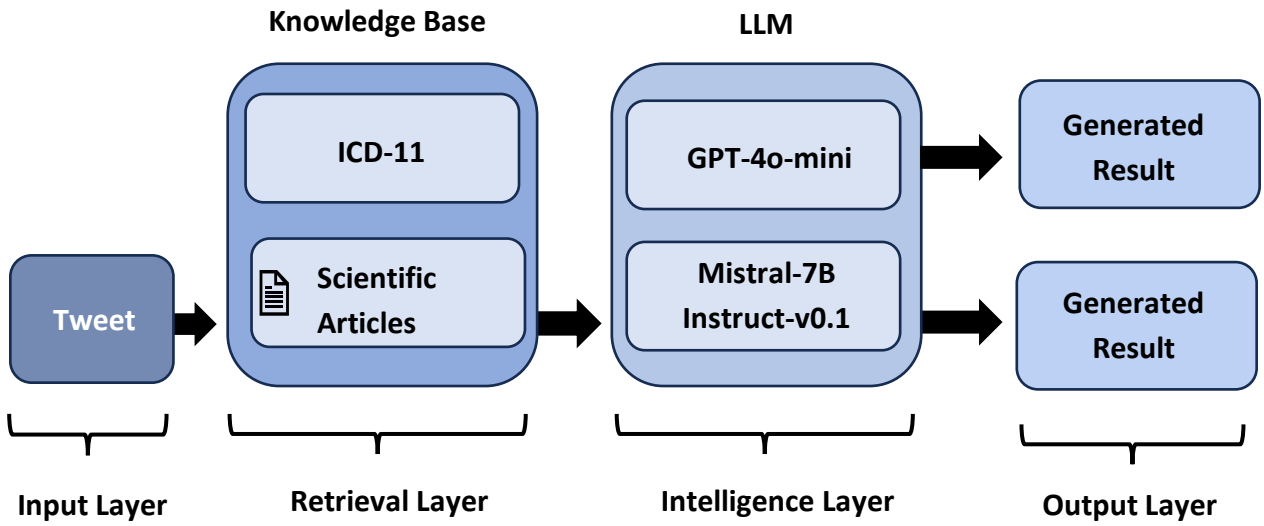

Figure 1 – Disorder detection framework

3.2.2. Component Choices

The model was divided into two key components: the retrieval layer, which provides the clinical context and the intelligence layer, which is composed of LLMs to produce diagnostic labels. Together, these components balance domain expertise with the generative reasoning capabilities of modern LLMs.

3.2.2.1. RETRIEVAL LAYER

To enhance classification accuracy with domain-specific knowledge, a retrieval component was integrated into the model's architecture. This component retrieved relevant texts (such as clinical definitions or scholarly findings) from external knowledge repositories to provide context during classification.

Source texts were drawn from two curated corpora: (1) the ICD-only corpus, comprising formal definitions and diagnostic criteria for the two psychological disorders obtained via the official ICD-11 API, and (2) the scientific articles corpus, containing abstracts and key excerpts from twenty and forty peer-reviewed studies on 'depression' and 'suicide' selected through targeted Google Scholar queries. However, in an ablation study, the best-performing source text was chosen to be continued with.

Prior to indexing, each document was segmented into overlapping 500-character chunks to preserve localized meaning while respecting LLM token limits. The chunk overlap was set to 50 after an ablation study. Each chunk was embedded using a pre-trained sentence-transformer model optimized for short texts. Four distance metrics were evaluated to identify the optimal similarity measure for disorder-label embeddings: cosine similarity, which measures the angle alignment between vectors (ignoring magnitude) with values near 1 indicating identical directionality; dot-product, which computes the sum of element-wise products and reflects magnitude-sensitive alignment; Manhattan distance (L1), calculated by summing absolute dimensional differences to quantify total positional deviation; and Euclidean distance (L2), which computes straight-line distance through Pythagorean summation and amplifies significant dimensional differences with its squaring operation.

The ablation study indicated that Euclidean distance maximized recall (weighted F_1 score), and thus, all vectors were stored in a FAISS inner-product index tuned for Euclidean searches.

At inference time, each user post was likewise embedded and used to retrieve the top three nearest chunks, which were concatenated into a context block and prepended to the prompt. This mechanism ensured that the downstream LLM classifier operated with both the user's input and domain-specific definitions or findings, thereby enhancing its ability to discern subtle linguistic indicators of mental health states.

3.2.2.2. INTELLIGENCE LAYER

In this study, two different LLMs were selected to evaluate the impact of RAG-based classification on detecting psychological disorders in X (formerly Twitter) text: Mistral-7B-Instruct-v0.1 and GPT-4o-mini.

Mistral-7B was chosen due to its open-source availability and its demonstrated effectiveness in recent studies involving classification tasks on social media data, such as fake news detection (Neralla & Vroe, 2024; Anonymous Authors, 2024). Its strong performance in prior NLP tasks, combined with its relatively small size and accessibility for local deployment, made

it an ideal candidate for testing retrieval-augmented strategies cost-effectively and transparently (Gül et al., 2024).

GPT-4o-mini, on the other hand, represents a state-of-the-art model developed by OpenAI (Badran et al., 2025). It is widely recognized for its robust generalization abilities and has become a standard baseline in both academic and industry NLP research (Roumeliotis et al., 2024). Including GPT-4o-mini provides a comparative benchmark against a leading commercial LLM, helping to situate the results of this study within a broader context and evaluate whether open-weight models like Mistral can achieve similar performance at lower computational or financial cost.

Together, these models offer a meaningful contrast between open-source and proprietary LLMs in the context of retrieval-augmented mental health classification. The full implementation details and model configuration for the prompt are described in Section 3.4.2.

3.3. Dataset & Preprocessing

For this study, a publicly available dataset from Kaggle was chosen. This dataset compiles 53,043 social media posts (primarily tweets, with some content from Reddit and Facebook) labelled into seven psychological categories: 'Normal', 'Depression', 'Suicidal', 'Anxiety', 'Bipolar', 'Stress', and 'Personality Disorder'. Each entry in the dataset is assigned a unique ID to ensure traceability and ease of handling during data processing. Furthermore, each entry consists of a short textual statement (the post content) and a corresponding status label indicating the author's mental state.

Upon initial analysis, it was observed that there were 362 posts contained missing text. These were removed, which left 52,681 labelled posts. Moreover, the dataset was unevenly distributed across the seven psychological states. Categories like 'Normal', 'Depression', and 'Suicidal' each contain approximately 10,000 tweets, while others, such as 'Bipolar' and 'Personality Disorder', have 2,000-3,000 tweets. No resampling or class-weighting strategies were applied, as the goal was to assess model performance on the proper distribution of social media posts, as would be encountered in real-world monitoring. Given that the primary evaluation metric is macro-averaged F_1, it was ensured that the performance of the minority suicidal class was measured equally alongside the majority classes. Moreover, oversampling or synthetic augmentation could introduce linguistic artefacts that interfere with retrieval and prompt-engineering effects. A more detailed discussion can be found in the limitations section. It was decided to continue with the three most substantial classes of the dataset - 'Normal', 'Depression' and 'Suicide' - for more advanced experimentation.

Table 6 – Data label distribution prior to split

Label	Row Count
Normal	16,351
Depression	15,404
Suicidal	10,653

On these labels a random train/validation split with a test size of 20% was performed for each the depression binary task, the suicidal binary task and the ternary task. Which left the dataset distribution at:

Table 7 – Data label distribution for ternary task

Label	Row Count
Normal	13,065
Depression	12,345
Suicidal	8,516

Table 8 – Data label distribution for depression binary task

Label	Row Count
Normal	13,104
Depression	12,300

Table 9 – Data label distribution for suicidal binary task

Label	Row Count
Normal	13,132
Suicidal	8,471

3.4. Model and Retrieval Implementation

Before building the RAG pipelines, it was first established that basic text features vary systematically by risk status. Therefore, Welch's ANOVA and Bonferroni-corrected pairwise t-tests on statement length and word count (Section 4.1) confirm a graded linguistic signal supporting H_1-H_3.

3.4.1. RETRIEVER SETUP

First, definitions for 17 ICD-11 entities were fetched via the WHO ICD-11 REST API. The first 20 and then 40 scientific abstracts were curated on depression and suicide, yielding 57 source documents through the Semantic Scholar Academic Graph API. To prefilter, multiple similarity measures were tested to explore the best possible. For this, an analysis of cosine, dot-product, Manhattan, and Euclidean distances over a sample of 200 posts shows that Euclidean distance maximized recall.

Table 10 – Similarity check

	Macro-F_1	**Weighted-F_1**	**Accuracy**
Cosine	0.6192	0.6453	0.6350
Euclidean	0.6737	0.7071	0.7075
Manhattan	0.6620	0.6963	0.6950
Dot-Product	0.6715	0.7061	0.7050

Because Euclidean distance outperformed all other metrics on macro-F_1, weighted-F_1, and overall accuracy, the FAISS index was fixed with L2 (Euclidean) similarity.

Furthermore, the 'search_kwargs' parameter was fixed to k equals 3 to retrieve the top three nearest neighbours for every query. This configuration ensured that each LLM prompt was augmented with the three most semantically relevant context snippets, balancing domain coverage against prompt length constraints.

Another filtering step was to see what retrieval documents would perform the best. Therefore, only ICD vs 20 scientific articles vs 40 scientific articles vs 80 scientific articles vs both ICD and all scientific articles were tested on a sample of 200 X (formerly Twitter) posts. This resulted in a continuation with only the ICD documents.

Table 11 – Database check

	Macro F_1	**Weighted F_1**	**Accuracy**
Mistral-7B-Instruct-v0.1 ICD	0.7379	0.7510	0.7463
Mistral-Instruct-v0.1 -7B SA (20)	0.3227	0.3198	0.3313
Mistral-7B-Instruct-v0.1 SA (40)	0.3314	0.3324	0.3374
Mistral-7B-Instruct-v0.1 SA (80)	0.3101	0.3118	0.3156
Mistral-7B-Instruct-v0.1 SA (40) & ICD	0.3297	0.3308	0.3355

3.4.2. LLM IMPLEMENTATION DETAILS

For the implementation of the LLM, two LLM classifiers were compared under identical prompt structures. These two models were selected based on the rationale described in Section 3.2.2.2. This section provides a detailed description of their configuration and deployment.

The Mistral-7B-Instruct-v0.1 model was implemented via the DeepInfra API endpoint. The 'max_new_token' parameter was set to 20 to optimize for short, label-only completions, and the temperature was set to 0.1.

The GPT-4o-mini model, accessed through OpenAI's API interface, was employed with the 'max_tokens' set to 20, as the output consists solely of a label, making a short response sufficient. Furthermore, the temperature was set to 0.1 to ensure deterministic and consistent outputs, which is crucial for reliable label generation.

The prompts were constructed as follows:

First, chain-of-thought prompting was applied to structure the model's reasoning into clear, sequential steps. The prompt template instructs the model to:

1. Analyse the input post, along with its contextual snippets.
2. Determine the single most appropriate diagnostic label.
3. Respond only with the label name in lowercase.
4. Include no additional commentary.

This explicit scaffold reduces the risk of spurious or overly verbose responses. Building on this foundation, two in-context learning regimes were then evaluated:

1. Zero-Shot Prompting: The model receives only the structured instructions and the retrieved context, relying solely on its pre-trained knowledge.
2. Few-Shot Prompting: Two random example triplets (post, context, and correct label) are embedded in the prompt to guide its classification reasoning.

Furthermore, after implementing the prompt few-shot with three random examples were prepended per target class (nine in total for the ternary task and six for the binary tasks), randomly drawn from the 20% random train/validation split data, to guide in-context learning.

All responses were post-processed using a fuzzy-matching label manager (threshold = 60 %) to convert raw text outputs into discrete label codes, ensuring robustness to minor formatting variations.

3.5. EXPERIMENTAL DESIGN

To evaluate the RAG-based classification framework across varying levels of label complexity, three parallel experiments were conducted: two binary distinctions and one three-class task, each following an identical data-splitting, training, and evaluation protocol. All experiments utilized the same train/validation sets described in Section 3.3, and three configurations were trained and evaluated: (1) the baseline LLM without retrieval or exemplars, (2) the LLM with retrieval, and (3) the LLM with retrieval and few-shot prompts.

3.5.1. BINARY TASK A

In this experiment, posts labelled with 'Depression' containing 12,345 examples and 'Normal' posts with 13,065 examples were isolated. Using the same three experimental configurations, model performance was evaluated in discriminating between depressive language patterns and normative expressions. The development split is used to select exemplars representative of each class.

3.5.2. BINARY TASK B

Analogously, the dataset was filtered to retain only posts labelled 'Suicidal' or 'Normal'. The resulting subset comprises 8,516 'Suicidal' and 13,065 'Normal' posts. This task was done to isolate the incremental gains from each augmentation stage.

3.5.3. THREE-CLASS TASK

For the ternary classification, all posts were merged from the three target labels 'Normal', including 13,065 posts; 'Depression', including 12,345 posts; and 'Suicidal', including 8,516 posts - totalling 33,926 entries. This was done while maintaining class proportions in the splits and extending the few-shot condition to include two exemplars per class - six total. This task tests the framework's ability to resolve finer-grained distinctions among three mental health states.

3.5.4. MODEL & RETRIEVAL PIPELINE

All experiments share a consistent inference pipeline (Figure 2), integrating the components described in Sections 3.4.1 and 3.4.2. Each input post is embedded and used to retrieve the top 3 most relevant chunks from the FAISS index. These are prepended to a structured chain-of-thought prompt and submitted to the LLM under one of two regimes: zero- vs. few-shot. Final outputs are normalized via fuzzy matching to ensure robust label assignment. To evaluate statistical reliability, each configuration was repeated across three random seeds, and performance metrics were reported as the mean and standard deviation on the validation set. This setup ensures comparability across tasks and isolates the contributions of retrieval, in-context learning, and model type.

ID	Text	Label
42346	monasmith sadly yes i think i need counciling now	Normal
10870	i seriously do not understand how telling someone that it does not solve the problem.I am going to prepare for my death. and just work on that. tired of living. death does solve pain	Suicidal
47870	I feel like you're supposed to grow as a person in life and I haven't at all I'm 23 and I am a former shell of who I was at like 16. I have no goals or ambitions anymore, I used to be ambitious. I'm trying to improve, like I am volunteering but I still don't have confidence or anything	Depression

Split & Extract

ID	Text
42346	monasmith sadly yes i think i need counciling now
10870	i seriously do not understand how telling someone that it does not solve the problem.I am going to prepare for my death. and just work on that. tired of living. death does solve pain
47870	I feel like you're supposed to grow as a person in life and I haven't at all I'm 23 and I am a former shell of who I was at like 16. I have no goals or ambitions anymore, I used to be ambitious. I'm trying to improve, like I am volunteering but I still don't have confidence or anything

ID	Original Label
42346	Normal
10870	Suicidal
47870	Depression

Label

ID	Text	RAG Label
42346	monasmith sadly yes i think i need counciling now	Normal
10870	i seriously do not understand how telling someone that it does not solve the problem.I am going to prepare for my death. and just work on that. tired of living. death does solve pain	Suicidal
47870	I feel like you're supposed to grow as a person in life and I haven't at all I'm 23 and I am a former shell of who I was at like 16. I have no goals or ambitions anymore, I used to be ambitious. I'm trying to improve, like I am volunteering but I still don't have confidence or anything	Depression

Compare Labels

ID	RAG Label
42346	Normal
10870	Suicidal
47870	Depression

ID	Original Label
42346	Normal
10870	Suicidal
47870	Depression

Figure 2 – Classification framework

3.5.5. METRICS

The classification performance of models was comprehensively assessed using multiple metrics to ensure robust and meaningful evaluation. These metrics included a classification report that provided detailed performance measures, including precision, recall, macro- and weighted-F_1-score for each psychological category, offering a nuanced view of model accuracy and sensitivity across classes. Additionally, a confusion matrix was generated and visualized through heatmaps, allowing for the intuitive identification of specific misclassification patterns between categories.

Clinical utility metrics were specifically calculated for critical categories such as the suicidal class, including FNR, PPV, and critical recall. Lastly, error analysis was conducted to identify misclassified samples and reveal recurring error patterns, potentially informing future model refinements and a deeper understanding of category distinctions

4. RESULTS

First, basic linguistic signals (4.1) were explored to verify the hypotheses (H_1–H_3). Then the reported model performance was evaluated under three settings: baseline (4.2), retrieval (4.3), and retrieval-integrated few-shot prompting (4.4). Moreover, it concluded with an error analysis (4.5).

4.1. DATA ANALYSIS RESULT

The exploration of the three risk states showed multiple differences. Pairwise Welch t-tests with Bonferroni correction revealed a clear separation in statement length: normal posts were on average 754 chars longer than depressive posts (t(15870) = –97.44, p-Bonf < 0.001, see Table 12) and 645 chars longer than suicidal posts (t(10863) = –67.42, p-Bonf < 0.001, see Table 12); suicidal posts were themselves 109 chars longer than depressive posts (t(22455) = 8.93, p-Bonf < 0.001, see Table 12). In word count, normal posts contained 151 more words than depressive (t(15828) = –98.74, p-Bonf < 0.001, see Table 12) and 129 more words than suicidal posts (t(10858) = –70.98, p-Bonf < 0.001, see Table 12), while suicidal posts out-worded depressive posts by 22 words (t(23004) = 9.14, p-Bonf < 0.001, see Table 12). All effects remain highly significant. The largest gaps (normal vs. suicidal) support H_1, intermediate gaps (normal vs. depression) support H_2, and the smallest gap (depression vs. suicidal) supports H_3, underscoring the graded linguistic signal.

Table 12 – Pairwise Welch t-test with Bonferroni correction for statement length and word count

Feature	Comparison	t	df	p-unc	p-Bonf
Statement Length	Normal vs. Depression	-97.44	15,870	<0.001	<0.001
Statement Length	Normal vs. Suicidal	-67.42	10,863	<0.001	<0.001
Statement Length	Depression vs. Suicidal	8.93	22,455	<0.001	<0.001
Number of Words	Normal vs. Depression	-98.74	15,828	<0.001	<0.001
Number of Words	Normal vs. Suicidal	-70.98	10,858	<0.001	<0.001
Number of Words	Depression vs. Suicidal	9.14	23,004	<0.001	<0.001

Normal posts (Fig. 3) were dominated by every day, social-oriented words such as “want”, “time”, “one”, “people”, “work”, “friend” and “day” – reflecting routine concerns, future planning, or simple social interactions. This suggests that non-at-risk users frame their experience around external activities and connections.

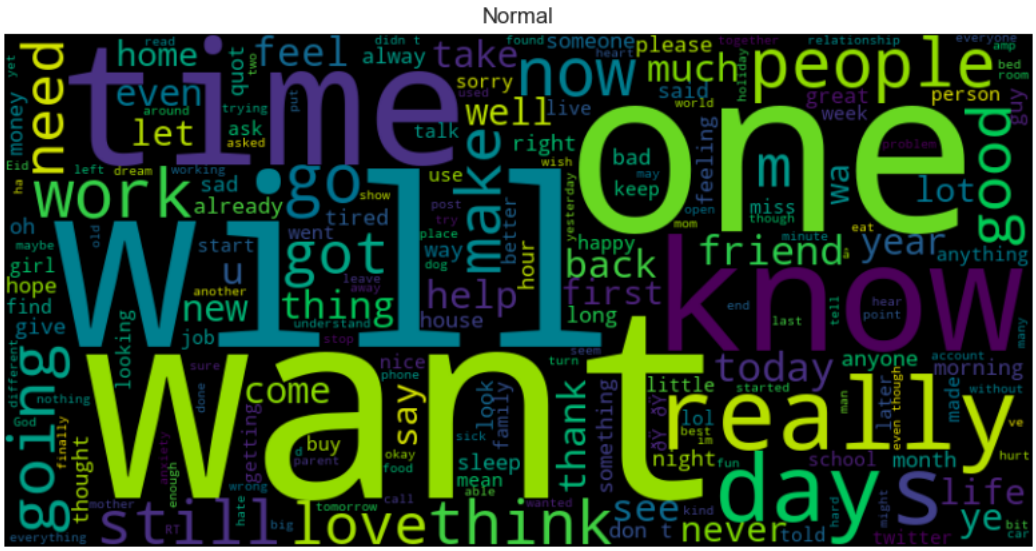

Figure 3 – Word cloud showing the top tokens for the normal status

Depression posts (Fig. 4) shifted toward more self-focused and negative-affect terms. “Feel”, “want”, “life”, “now”, “alone”, “can’t” and “make” become prominent, indicating rumination on personal mood, loss of agency, and emerging hopelessness. The persistence of “want” paired with “feel” and “life” underscores the gap between desires and emotional state.

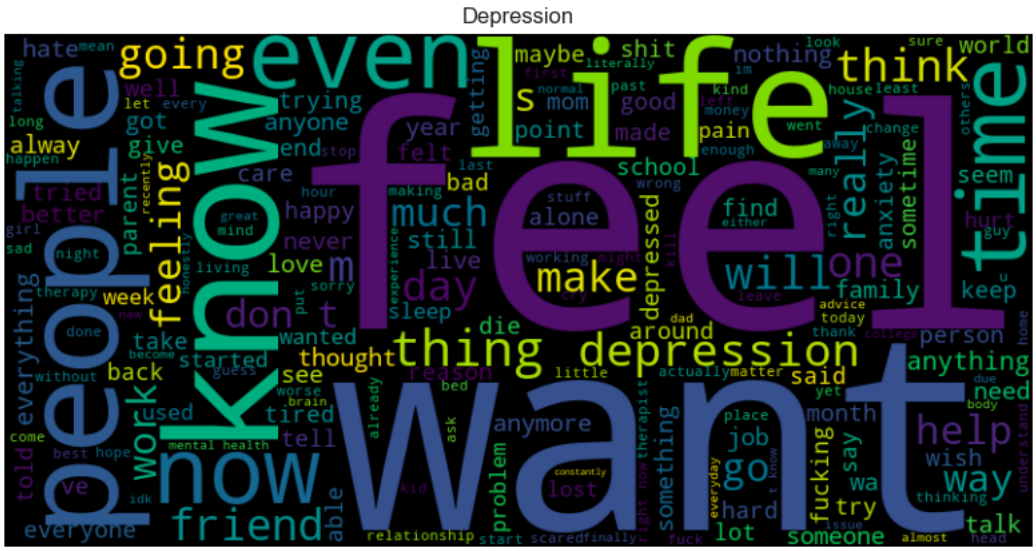

Figure 4 – Word cloud showing the top tokens for the depression status

Suicidal posts (Fig. 5) amplify death-related and urgency cues. While “want” and “feel” remain similar to in the depression posts, they now co-occur with stark terms like “die”, “death”, “anymore”, “take” and “know.” The prominence of “die” and “anymore” signals explicit intent and perceived finality, and “take” hints at planning. This lexical pattern, heavy on life-ending verbs and adverbs of time, directly flags acute suicidal ideation.

Figure 5 – Word cloud showing the top tokens for the suicidal status

This underscores that language not only becomes more negative but also more lexically complex as risk intensifies.

4.2. Baseline (No Retrieval)

In the baseline setting, both models received only the raw post text and the chain-of-thought scaffold without any external context. GPT-4o-mini outperformed Mistral-7B-Instruct-v0.1 in the binary classification however in the ternary Mistral-7B-Instruct-v0.1 outperformed GPT-4o-mini:

Table 13 – Base-model performance on binary depression detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.8957	0.8957	0.8957
Mistral-7B-Instruct-v0.1	0.8782	0.8779	0.8787

Table 14 – Base-model performance on binary suicide detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.9192	0.9238	0.9248
Mistral-7B-Instruct-v0.1	0.9129	0.9160	0.9152

Table 15 – Base-model performance on ternary mental health detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.7366	0.7495	0.7536
Mistral-7B-Instruct-v0.1	0.7474	0.7590	0.7582

Across the three core tasks, GPT-4o-mini and Mistral-7B-Instruct-v0.1 exhibit complementary strengths (Tables 13–15). GPT-4o-mini leads with a macro-F_1 of 0.896 (accuracy = 0.896), outperforming Mistral by ≈ 1.8 percentage-points (pp) on macro-F_1 and 1.7 pp on accuracy.

Weighted-F_1 mirrors this gap, implying the advantage is consistent across both majority and minority classes. GPT-4o-mini achieves macro-F_1 = 0.919 and accuracy = 0.925, only about 0.6–1.0 pp ahead of Mistral-7B-Instruct-v0.1 (see Table 14). Both models exceed 0.91 on every metric, indicating that suicide-related language is easier to separate in a two-class setting than depressive language. Here the ranking reverses. Mistral-7B-Instruct-v0.1 yields the highest macro-F_1 (0.747) and accuracy (0.758), surpassing GPT-4o-mini by roughly 1 pp on each measure (see Table 15). The drop in macro-F_1 for both models relative to the binary tasks (≈15–18 pp) underscores the added difficulty created by lexical overlap between the depression and suicidal classes.

4.3. RETRIEVAL

Adding domain knowledge via the FAISS-based retriever yields varying gains depending on corpus:

Table 16 – Retrieval-model performance on binary depression detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.8727	0.8725	0.8730
Mistral-7B-Instruct-v0.1	0.8192	0.8184	0.8223

Table 17 – Retrieval-model performance on binary suicide detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.9394	0.9423	0.9425
Mistral-7B-Instruct-v0.1	0.9098	0.9130	0.9122

Table 18 – Retrieval-model performance on ternary mental health detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.7379	0.7510	0.7461
Mistral-7B-Instruct-v0.1	0.7379	0.7510	0.7463

For the retrieval condition in Tables 16 it was observed that the retrieved documents lowered macro- F_1 for depression by 2–6 percentage-points, with the sharper drop for Mistral-7B-Instruct-v0.1. When the target concept is explicitly defined in the retrieved text (e.g., intent to self-harm), GPT-4o-mini gains two points, while Mistral-7B-Instruct-v0.1 remains virtually

unchanged (see Table 17). After retrieval, both systems stabilise at macro-$F_1 \approx 0.74$, indicating that external definitions alone cannot fully disentangle the lexical overlap between depression and suicidal posts (see Table 18). GPT-4o-mini is comparatively robust - one gain, one mild loss - whereas Mistral-7B-Instruct-v0.1 deteriorates on two out of three tasks, suggesting it is more easily "distracted" by additional context.

4.4. ZERO- VS. FEW-SHOT

Next in-context learning was accessed with few-shot prompts - both using ICD-11 retrieval.

Table 19 – Few-shot retrieval-model performance on binary depression detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.8946	0.8947	0.8950
Mistral-7B-Instruct-v0.1 Few Shot	0.9136	0.9137	0.9137

Table 20 – Testing of Few-shot retrieval-model performance on binary suicide detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.9480	0.9508	0.9513
Mistral-7B-Instruct-v0.1 Few-Shot	0.9379	0.9404	0.9402

Table 21 – Few-shot retrieval-model performance on ternary mental health detection

	Macro F_1	**Weighted F_1**	**Accuracy**
GPT-4o-mini	0.7615	0.7738	0.7775
Mistral-7B-Instruct-v0.1 Few Shot	0.7620	0.7793	0.7795

Both models gained on every task, confirming that two carefully chosen demonstrations are enough to steer the model toward more reliable label boundaries. The largest single improvement (+9.5 pp) occurs for Mistral-7B-Instruct-v0.1 on binary depression, pushing it past GPT-4o (0.914 vs 0.895) as can be seen in Table 19. This suggests Mistral-7B-Instruct-v0.1 benefits more from explicit examples when clinical and colloquial language diverge. GPT-4o-mini's macro-F_1 rose to 0.948, widening its lead over Mistral-7B-Instruct-v0.1 to 1 pp. In terms of recall, the gap grew from 1.3 pp in the baseline (92.5 % vs 91.2 %) to 2.3 pp with retrieval + few-shot (94.3 % vs 92.0 %), a clinically important margin. Few-shot guidance boosted both GPT-4o-mini and Mistral-7B-Instruct-v0.1 by 2.4 pp, yet both remain around 12 pp below their binary scores, underscoring the persisting lexical overlap between depression and suicidal posts (see Table 21). GPT-4o-mini shows modest, steady gains (+0.9–2.4 pp) but retained overall leadership only in one of three tasks, while Mistral-7B-Instruct-v0.1 exhibits a larger swing - dramatically better on depression, slightly better in the ternary detection but still trailing on suicide.

4.5. Error Analysis

The following table shows all models error rates.

Table 22 – Error Table

Model / Setting	Task	Total Samples	Misclassified	Error Rate
Mistral-7B-Instruct-v0.1	Binary (Depression)	25,404	3,067	12.07%
Mistral-7B-Instruct-v0.1 + Retriver	Binary (Depression)	25,404	4,506	17.73%
Mistral-7B-Instruct-v0.1 + Retriver + Few-Shot	Binary (Depression)	25,404	2,285	8.99%
GPT-4º-mini	Binary (Depression)	25,404	2,375	9.35%
GPT-4º-mini + Retriver	Binary (Depression)	25,404	2,909	11.45%
GPT-4º-mini + Retriver + Few-Shot	Binary (Depression)	25,404	1,996	7.85%
Mistral-7B-Instruct-v0.1	Binary (Suicidal)	21,603	1,831	8.48%
Mistral-7B-Instruct-v0.1 + Retriver	Binary (Suicidal)	21,603	1,895	8.77%
Mistral-7B-Instruct-v0.1 + Retriver + Few-Shot	Binary (Suicidal)	21,603	1,286	5.95%
GPT-4º-mini	Binary (Suicidal)	21,603	1,598	7.40%
GPT-4º-mini + Retriver	Binary (Suicidal)	21,603	1,233	5.71%
GPT-4º-mini + Retriver + Few-Shot	Binary (Suicidal)	21,603	1,053	4.87%
Mistral-7B-Instruct-v0.1	Ternary	33,926	8,186	24.13%
Mistral-7B-Instruct-v0.1 + Retriver	Ternary	33,926	8,547	25.19%
Mistral-7B-Instruct-v0.1 + Retriver + Few-Shot	Ternary	33,926	7,432	21.91%
GPT-4º-mini	Ternary	33,926	8,361	24.64%
GPT-4º-mini + Retriver	Ternary	33,926	8,208	24.19%
GPT-4º-mini + Retriver + Few-Shot	Ternary	33,926	6,880	20.28%

The error analysis confirms a clear and consistent benefit from few-shot prompting across all experimental tasks. Incorporating a small set of labelled examples reduced the overall error rate for both models, with the most pronounced improvement, a 5.6-percentage-point decrease, observed for Mistral-7B-Instruct-v0.1 in the binary depression task (see Table 22). In contrast, the retrieval component on its own proved ambivalent: while it enhanced

performance when combined with few-shot prompts, it tended to slightly increase errors when applied in isolation, suggesting that the additional context can over-extend the models' attention capacity. On the baseline configurations GPT-4o-mini outperformed Mistral-7B-Instruct-v0.1. However, once few-shot examples were introduced, that advantage diminished and, in some cases, even reversed, indicating that Mistral-7B-Instruct-v0.1 derives comparatively greater benefit from explicit in-context demonstrations. Despite these gains, ternary classification (normal, depression, suicidal) remains the most challenging scenario, with error rates persisting above 20 percent - mainly due to the lexical and conceptual overlap between posts labelled as depression and those labelled as suicidal (see Section 4.4). The following shows examples of misclassifications.

Table 23 – Qualitative inspection of misclassifications

ID	Text (truncated)	True Label	Predicted Label	Likely cause
14487	"(...) I wish I could watch them for the rest of my life (…) I am so ready to go (...)"	Suicidal	Depression	Metaphorical language; no direct intent cues
17639	"what is the point? (…) plan my next attempt but really, (...)"	Depression	Suicidal	Explicit self-harm plan triggered model
28293	"(…) I honestly dont think i can live long (…) I was pretty much done with my live (…)"	Normal	Suicidal	Label noise - text clearly self-harm-related - however in the past
5985	"I haven't opened it for 2 days, it's all over, it's really late"	Normal	Depression	Ambiguous context (no subject)
7279	"Do you think if you were born in another era, you would be happier? Different time period"	Suicidal	Normal	Rhetorical question; lacks self-referential cues
11256	"(...) I ATE A FULL MEAL TODAY (...) I FUVKING DID TI"	Depression	Normal	Positive sentiment overrides depressive label

A closer inspection of misclassified cases revealed four recurring error patterns (see Table 23). First, tweets labelled as depression or suicidal often share the same vocabulary of hopelessness, leading the models to confuse expressions of ideation that stop short of an explicit self-harm intent. Second, very short posts - such as ID 5985 - provide little syntactic structure (few pronouns, no temporal markers), depriving the model of cues it needs to detect the

author's mental state. Third, a degree of label noise limits the attainable ceiling: some messages tagged normal in the source dataset actually contain overt self-harm language (e.g., ID 28293), undermining both training and evaluation fidelity. Finally, the models sometimes over-weight surface sentiment; in ID 11256 the celebratory tone ("I ATE A FULL MEAL TODAY…") prompted a normal prediction even though the tweet belonged to a user with depressive symptoms.

5. DISCUSSION

This study offers valuable insights into the effectiveness and limitations of RAG and LLMs for classifying mental health issues in social media texts. In particular, it highlights how task formulation—binary vs. multiclass—significantly impacts performance outcomes.

Binary classifiers outperformed multiclass models - depression vs. normal (macro-F_1 = 0.87; Table 16) and suicidal vs. normal (macro-F_1 = 0.94; Table 17) - substantially outperformed the three-class model (weighted-F_1 ≈ 0.78; Table 18) in the RAG task, confirming H_1–H_3 and highlighting the escalating ambiguity of multiclass detection in short social-media posts. This demonstrates the inherent difficulty of multiclass mental health classification.

Furthermore, it was shown that adding ICD-11 snippets, which performed superior to scientific article snippets, helped the GPT-4o-mini model on the suicidal task but introduced noise to GPT-4o-mini on depression and ternary detection. However, Mistral-7B-Instruct-v0.1's performance degraded in all metrics for all three classification tasks, so GPT-4o-mini was overall more robust to retrieval. It is possible that the formal, clinical language of the ICD-11 definitions introduced a stylistic mismatch with the informal nature of social media posts, creating noise for the less robust model or in tasks where the linguistic boundary was already subtle (depression vs. normal).

In addition to retrieval, few-shot exemplars played a crucial role in the RAG pipeline. By embedding just three representative examples per class, Mistral-7B-Instruct-v0.1's depression-vs-normal F_1 improved by over 3 points - enabling it to surpass GPT-4o-mini in that task - while GPT-4o-mini also saw modest gains. This reveals that, in domains with subtle linguistic distinctions, few-shot prompting serves as a low-cost alternative to fine-tuning, guiding models toward the nuanced decision boundaries needed for accurate mental health classification.

To better understand the model performance, especially in the multiclass setting, a linguistic comparison across the three mental health states was conducted. This analysis revealed statistically significant differences in both statement length and word count, supporting the hypothesized continuum of risk (H_1–H_3). It was demonstrated that linguistically and clinically, depression and suicidal ideation share many features, which complicate their distinction in short, unstructured social media posts. Depression posts may contain expressions of hopelessness, fatigue, or sadness, while suicidal posts convey similar sentiments but with added intensity or finality. For a model, especially one trained on noisy or loosely annotated data, these nuanced differences are hard to discern. This class confusion is compounded by the fact that many features helpful in distinguishing depression from normal are not necessarily optimal for separating depression from suicidal ideation.

When comparing these results to the recent literature, it's clear that the approach chosen in this study falls short of the top benchmarks set by fine-tuned models on similar tasks. Fine-

tuned transformer models remain state-of-the-art, often reaching substantially higher accuracies. For example, Kermani et al. (2025) report that a fine-tuned LLaMA-based model achieved approximately 91% accuracy in a multiclass text classification (emotion recognition) task and about 80% accuracy in detecting specific mental health conditions. These figures exceed the 75% accuracy achieved by this study on the ternary mental health classification task. Likewise, other recent works using specialized or ensemble models have reported very high performance (Hossain et al., 2025; Alghazzawi et al., 2025).

That said, it is worth noting that the ternary classification outcome in this study (around 75% accuracy) still compares favourably to some earlier transformer baselines. For instance, as shown by Metzler et al. (2022) BERT and XLNet models classified suicide-related tweets into six risk categories and reported only about 73% accuracy (with F_1 scores 0.70–0.85 in most classes). So, although this study's generative models did not match the latest fine-tuned systems, they nevertheless outperformed or equalled certain prior studies' results. Additionally, the binary classification tasks (depression vs. control and suicidal vs. non-suicidal) yielded high accuracies (roughly 91–95%), which approach the performance of dedicated models on similar two-class problems.

In conclusion, however, this approach shows mixed outcomes for the implementation of RAG for mental health detection. Retrieval added practical clinical context for detecting suicide and marginally for the three-class task in GPT-4o-mini, but on balance; it introduced more noise - especially for depression detection and for the Mistral-7B-Instruct-v0.1 model. For future implementation, it is necessary to filter or customize retrieved snippets to ensure they genuinely help rather than hinder. Retrieval seems most useful when the task is already challenging (suicide, ternary) and the model has strong underlying clinical knowledge (GPT-4o-mini). In more manageable tasks (depression vs. normal) or for the less robust model (Mistral-7B-Instruct-v0.1), retrieved snippets can introduce irrelevant context that leads to misinterpretation.

6. CONCLUSIONS

6.1. SYNTHESIS OF THE DEVELOPED WORK

This study set out to evaluate whether RAG can effectively support the detection of psychological disorders in social media text, compared to more established prompting and fine-tuning strategies. To that end, multiple pipeline configurations were implemented using a Kaggle mental health dataset. These configurations included baseline LLMs, RAG, and a combined RAG + few-shot approach. Evaluation metrics such as macro-averaged F_1-score were used to account for class imbalance and ensure comparability across models.

The research question - whether RAG can be used effectively to classify mental health conditions - was indeed answered. While the implementation of RAG on its own yielded moderate and generally lower performance compared to prompt-only baselines, the combined RAG + few-shot setup demonstrated improved results. However, these improvements may be primarily attributable to the few-shot component rather than the retrieval mechanism itself. This finding suggests that while RAG offers a theoretically interpretable and context-rich classification framework, its added value for this specific task remains limited without further optimization of the retrieval strategy or underlying knowledge base.

Nonetheless, the overall objective of the study, exploring alternative LLM strategies for mental health text classification beyond traditional fine-tuning, was achieved. The results provide empirical insights into the strengths and limitations of RAG in a low-resource, real-world context. Limitations, such as class imbalance, retrieval noise, and the absence of supervised adaptation, are discussed in detail and serve as a basis for future improvements. As such, the study makes a valuable contribution to the emerging field of explainable and scalable LLM applications for mental health.

6.2. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This study opens several avenues for future exploration while acknowledging some important limitations.

One key challenge identified was the conceptual and linguistic overlap between depression and suicidal ideation, which likely contributed to misclassifications in the multiclass setting. These two conditions often share symptoms and language cues, making it difficult for models to distinguish between them. Future work could potentially address this issue by investigating advanced strategies, such as separate fine-tuning for each class or hierarchical classification pipelines, to improve the handling of semantic proximity.

A promising direction would be a two-step hierarchical approach. Here, in the first step, posts could be classified into normal vs at-risk, and in the second step, at-risk posts would be routed to a second specialized classifier that could be trained to distinguish between depression and suicide. This hierarchical strategy would help to improve the fine-grained distinction that this

multiclass model struggled with and reduce confusion between overlapping classes, which would consequently enhance overall classification performance.

Third, there is a notable gap between formal clinical descriptions of disorders and the informal language used on social media. The retrieval-augmented approach implemented knowledge from sources like the ICD-11, but these formal medical texts do not always align with how individuals express distress online. Social media posts often include slang, sarcasm and context-dependent meanings, which are difficult for models to interpret (Sykora et al., 2020). Colloquial expressions often deviate from formal medical terminology; for instance, "I DID IT - I ATE A FULL MEAL TODAY!" conveys excitement yet maps poorly to any ICD-11 symptom descriptor, illustrating how this vocabulary gap can hinder accurate recognition.

Furthermore, the dataset used, sourced from Kaggle, comes with limited metadata on the labelling origin. The absence of clinical annotation introduces a risk of noisy labels or weak supervision, which could influence model reliability. It was also observed in prior work that when data from multiple sources are combined, their labels while valid in their original context—may lack consistency in a unified dataset (Sedlakova et al., 2023). In this study, the issue was faced that the boundaries between what constitutes (for example) depression versus suicidal or other conditions were sometimes blurry due to varying annotation standards. Future collaborations with clinical professionals, such as psychologists or psychiatrists, to validate and annotate samples could substantially improve both the precision and real-world applicability of future models.

The dataset used in this study suffered a notable class imbalance. To preserve real-world validity, reflecting the natural distribution of mental health-related posts on online platforms, the decision was made not to apply resampling techniques such as oversampling, augmentation, or SMOTE. Instead, model evaluation relied on the macro-averaged F_1-score, a metric known to mitigate the influence of class imbalance by equally weighting each class regardless of its frequency (Hinojosa Lee et al., 2024). Nevertheless, it is important to acknowledge that severe imbalance may still affect the model's output, especially in tasks involving subtle class distinctions. In this study, however, no fine-tuning was conducted, and thus, no gradient-based learning process could be influenced by the imbalance directly. That said, the prompting-based and RAG approaches might still exhibit implicit biases toward dominant classes during inference due to the underlying distribution of examples. To further validate the robustness of this approach, future work could incorporate controlled resampling strategies and compare performance outcomes. This would help isolate the effects of class imbalance and ensure that data distribution artifacts do not inadvertently skew the findings.

And finally, another key limitation of fully automated RAG-based classifiers is their residual FNR, which in a clinical context can have extensive consequences. To address this, future implementations should incorporate a human-in-the-loop framework: licensed psychologists would review suicidal or high-risk classifications before any downstream action. In parallel, decision thresholds can be dynamically tuned - prioritizing higher recall on suicidal ideation

predictions - to further reduce the chance of missed cases. Equally important is close collaboration between researchers, practicing clinicians, and individuals with lived experience of mental illness. Together, they can assess the trade-offs between the potential benefits of social-media-based screening and its ethical, privacy, and safety risks.

6.3. ETHICAL CONSIDERATIONS

This study aligns with the ethical principles governing the use of data, model reliability, and the responsible deployment of LLMs in mental health assessment.

The dataset was accessed via the Kaggle API, and therefore, its status as a publicly available resource with documented accessibility is ensured.

Given the high stakes of detecting depression and suicidal ideation, automated systems were explicitly acknowledged to mitigate the risk of misclassification. Any prospective real-world implementation would necessitate the following:

(1) Rigorous clinical validation

(2) Continuous supervision by licensed mental health professionals

(3) Strict compliance with ethical frameworks and regulatory standards to mitigate risks of misinformation or unintended harm.

BIBLIOGRAPHICAL REFERENCES

Abdollahi, J., Davari, N., Panahi, Y., & Gardaneh, M. (2022). Detection of Metastatic Breast Cancer from Whole-Slide Pathology Images Using an Ensemble Deep-Learning Method. *Archives of Breast Cancer*, *9*(3). https://doi.org/10.32768/abc.202293364-376

Ahammad, S. H., Rajesh, V., Jafar Khan, P., Sumanth, P., Sivaram, G., Inthiyaz, S., & Saikumar, K. (2020). Chexnet reimplementation for pneumonia detection using pytorch. *International Journal of Pharmaceutical Research*, *12*(2). https://doi.org/10.31838/ijpr/2020.12.02.0023

Alghazzawi, D., Zhang, Z., & Liu, W. (2025). Explainable AI-based suicidal and non-suicidal ideation detection from social media text with enhanced ensemble technique. *Journal of Affective Disorders*, *321*, 45–57. https://doi.org/10.1038/s41598-024-84275-6

American Psychological Association. (2024, January). Trends and pathways to access mental health care. *Monitor on Psychology*. https://www.apa.org/monitor/2024/01/trends-pathways-access-mental-health-care

American Psychological Association. (2022, November). *2022 practitioner insights: Increasing demand, stress, and burnout*. https://www.apa.org/news/press/releases/stress/2022/state-of-mental-health

Anonymous. (2024, August). Fake news detection with retrieval-augmented generative artificial intelligence. In Proceedings of Workshop '4 (pp. 1–9). *Association for Computing Machinery*. https://openreview.net/pdf?id=wRCEh8mO57

Badran, N., Le, J., Le, T., & Uchiya, T. (2025). Improving stress detection with synthetic datasets: GPT-4o-Mini and transformer model evaluation. In N. T. Nguyen, H. Xiong, K.-H. Kim, & V. C. Leung (Eds.), *Intelligent information and database systems* (Lecture Notes in Computer Science, Vol. 15683, pp. 108–119). Springer. https://doi.org/10.1007/978-981-96-6008-7_9

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia*, *1*(1). https://doi.org/10.1038/npjschz.2015.30

Buddhitha, P., & Inkpen, D. (2023). Multi-task learning to detect suicide ideation and mental disorders among social media users. *Frontiers in Research Metrics and Analytics*, *8*. https://doi.org/10.3389/frma.2023.1152535

Centers for Disease Control and Prevention. (2022). *WISQARS Leading Causes of Death Reports, National and Regional, 1999–2020*. National Center for Injury Prevention and Control. https://www.cdc.gov/injury/wisqars/

Cha, J., Kim, S., & Park, E. (2022). A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. *Humanities and Social Sciences Communications*, *9*(1). https://doi.org/10.1057/s41599-022-01313-2

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

Dell'Osso, B., Glick, I. D., Baldwin, D. S., & Altamura, A. C. (2013). Can long-term outcomes be improved by shortening the duration of untreated illness in psychiatric Disorders? a conceptual framework. In *Psychopathology* (Vol. 46, Issue 1). https://doi.org/10.1159/000338608

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639). https://doi.org/10.1038/nature21056

Ghali, J. P. E., Shima, K., Moriyama, K., Mutoh, A., & Inuzuka, N. (2024). *Enhancing retrieval processes for language generation with augmented queries* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2402.16874

Gould, M., Jamieson, P., & Romer, D. (2003). Media contagion and suicide among the young. *American Behavioral Scientist*, *46*(9). https://doi.org/10.1177/0002764202250670

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. In *Current Opinion in Behavioral Sciences* (Vol. 18). https://doi.org/10.1016/j.cobeha.2017.07.005

Guo, Y., Ovadje, A., Al-Garadi, M. A., & Sarker, A. (2024). Evaluating large language models for health-related text classification tasks with public social media data. Journal of the American Medical Informatics Association, 31(10), 2181–2189. https://doi.org/10.1093/jamia/ocae210

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., Li, K., … & Sheu, Y.-H. (2024). *Large language models for mental health applications: Systematic review*. *JMIR Mental Health, 11*, Article e57400. https://doi.org/10.2196/57400

Gül, I., Lebret, R., & Aberer, K. (2024). *Stance detection on social media with fine-tuned large language models* [Preprint]. arXiv. https://arxiv.org/abs/2404.12171

Hawton, K., & James, A. (2005). Suicide and deliberate self-harm in young people. *BMJ, 330*(7496), 891–894. https://doi.org/10.1136/bmj.330.7496.891

Hawton, K., Saunders, K. E. A., & O'Connor, R. C. (2012). Self-harm and suicide in adolescents. In *The Lancet* (Vol. 379, Issue 9834). https://doi.org/10.1016/S0140-6736(12)60322-5

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted F_1-scores. *Applied Sciences, 14*(21), Article 9863. https://doi.org/10.3390/app14219863

Hossain, M. A., Islam, M. R., & Rahman, M. M. (2025). Multi-task opinion-enhanced hybrid BERT model for mental health analysis. *IEEE Transactions on Affective Computing*. https://doi.org/10.1038/s41598-025-86124-6

Hurtado, J. F. (2023). Harnessing retrieval-augmented generation (RAG) for uncovering knowledge gaps [Preprint]. *arXiv*.

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. In *Annual Review of Clinical Psychology* (Vol. 6). https://doi.org/10.1146/annurev.clinpsy.3.022806.091532

Hyman, S., Chisholm, D., Kessler, R. C., Patel, V., & Whiteford, H. (2006). Mental disorders. In *Disease control priorities related to mental, neurological, developmental and substance abuse disorders* (pp. 1–20). World Health Organization.

Instagram. (2019, February 7). Changes we're making to do more to support and protect the most vulnerable people who use Instagram. *Instagram Blog*. https://about.instagram.com/blog/announcements/supporting-and-protecting-vulnerable-people-on-instagram

Jeong, C. (2023). A study on the implementation of generative AI services using an enterprise data-based LLM application architecture. *Advances in Artificial Intelligence and Machine Learning, 3*(4), 1588–1618. https://doi.org/10.54364/aaiml.2023.1191

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active Retrieval Augmented Generation. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/2023.emnlp-main.495

Jovanovic, M., & Campbell, M. (2022). Generative Artificial Intelligence: Trends and Prospects. *Computer*, *55*(10). https://doi.org/10.1109/MC.2022.3192720

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*. https://doi.org/10.1016/j.nlp.2023.100048

Kaywan, P., Ahmed, K., Ibaida, A., Miao, Y., & Gu, B. (2023). Early detection of depression using a conversational AI bot: A non-clinical trial. *PLoS ONE*, *18*(February 2). https://doi.org/10.1371/journal.pone.0279743

Kerasiotis, M., Ilias, L., & Askounis, D. (2024). *Depression detection in social media posts using transformer-based models and auxiliary features*. *Social Network Analysis and Mining, 14*, Art. 196. https://doi.org/10.1007/s13278-024-01360-4

Kermani, A., Perez-Rosas, V., & Metsis, V. (2025). A Systematic Evaluation of LLM Strategies for Mental Health Text Analysis: Fine-tuning vs. Prompt Engineering vs. RAG. *arXiv preprint arXiv:2503.24307*.

Kumar, S., Sharma, N., Kumar, R., & Singh, S. (2024). GENERATIVE AI & LLMs.

Kuzlu, M., Xiao, Z., Sarp, S., Catak, F. O., Gurler, N., & Guler, O. (2023). The Rise of Generative Artificial Intelligence in Healthcare. *12th Mediterranean Conference on Embedded Computing, MECO 2023*. https://doi.org/10.1109/MECO58584.2023.10155107

Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. A. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta-analytic evaluation of recent meta-analyses. In *World Psychiatry* (Vol. 21, Issue 1). https://doi.org/10.1002/wps.20941

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *2020-December*.

Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020). SenseMood: Depression detection on social media. *ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval*. https://doi.org/10.1145/3372278.3391932

Mazhar, A., Hasan Shaik, Z., Srivastava, A., Ruhnke, P., Vaddavalli, L., Katragadda, S. K., Yadav, S., & Akhtar, M. S. (2025, April). *Figurative-cum-Commonsense knowledge infusion for multimodal mental health meme classification*. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*. ACM. https://doi.org/10.1145/3696410.3714778

McGrath, J. J., Al-Hamzawi, A., Alonso, J., Altwaijri, Y., Andrade, L. H., Bromet, E. J., Bruffaerts, R., de Almeida, J. M. C., Chardoul, S., Chiu, W. T., Degenhardt, L., Demler, O. v., Ferry, F., Gureje, O., Haro, J. M., Karam, E. G., Karam, G., Khaled, S. M., Kovess-Masfety, V., … Zaslavsky, A. M. (2023). Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. *The Lancet Psychiatry*, *10*(9). https://doi.org/10.1016/S2215-0366(23)00193-1

Metzler, H., Baginski, H., Niederkrotenthaler, T., & Garcia, D. (2022). Detecting Potentially Harmful and Protective Suicide-Related Content on Twitter: Machine Learning Approach. *Journal of Medical Internet Research*, *24*(8). https://doi.org/10.2196/34705

Moher, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, *62*(10), 1006–1012. https://doi.org/10.1016/j.jclinepi.2009.06.005

Molfenter, T., Heitkamp, T., Murphy, A. A., Tapscott, S., Behlman, S., & Cody, O. J. (2021). Use of Telehealth in Mental Health (MH) Services During and After COVID-19. *Community Mental Health Journal*, *57*(7). https://doi.org/10.1007/s10597-021-00861-2

Montejo-Ráez, A., Molina-González, M. D., Jiménez-Zafra, S. M., García-Cumbreras, M. Á., & García-López, L. J. (2024). A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, *53*, 100654. https://doi.org/10.1016/j.cosrev.2024.100654

Morgan, C., Webb, R. T., Carr, M. J., Kontopantelis, E., Green, J., Chew-Graham, C. A., Kapur, N., & Ashcroft, D. M. (2017). Incidence, clinical management, and mortality risk following self harm among children and adolescents: Cohort study in primary care. *BMJ (Online)*, *359*. https://doi.org/10.1136/bmj.j4351

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021, April). *Classification of mental illnesses on social media using RoBERTa*. In E. Holderness, A. J. Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, & F. Rinaldi (Eds.), *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis* (S. 59–68). Association for Computational Linguistics.

Muriello, D., Donahue, L., Ben-David, D., Ozertem, U., & Shilon, R. (2018, February 21). *Under the hood: Suicide prevention tools powered by AI.* Engineering at Meta. https://engineering.fb.com/2018/02/21/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020). Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. In *Journal of Technology in Behavioral Science* (Vol. 5, Issue 3). https://doi.org/10.1007/s41347-020-00134-x

National Institute of Mental Health. (2025, March). Suicide. https://www.nimh.nih.gov/health/statistics/suicide

Neralla, V., & de Vroe, S. B. (2024). *Evaluating Poro-34B-Chat and Mistral-7B-Instruct-v0.1: LLM system description for ELOQUENT at CLEF 2024*. In G. Faggioli, N. Ferro, P. Galuščáková, & A. G. Seco de Herrera (Eds.), *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings, Vol. 3740, pp. 708–711). CEUR-WS.org. https://ceur-ws.org/Vol-3740/paper-68.pdf

Ogunleye, B., Sharma, H., & Shobayo, O. (2024). *Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection*. *Big Data and Cognitive Computing*, **8**(9), 112. https://doi.org/10.3390/bdcc8090112

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. In *The BMJ* (Vol. 372). https://doi.org/10.1136/bmj.n71

Powell, A. C., Torous, J. B., Firth, J., & Kaufman, K. R. (2020). Generating value with mental health apps. *BJPsych Open*, *6*(2). https://doi.org/10.1192/bjo.2019.98

Qasim, I., Rizvi, A. S., & Tariq, M. (2025). Detection of depression severity in social media text using transformer-based models. *Computers in Human Behavior Reports*, *8*, 100215. https://doi.org/10.3390/info16020114

Rane, N. L., Tawde, A., Choudhary, S. P., & Rane, J. (2023). Contribution and performance of ChatGPT and other large language models (LLM) for scientific and research advancements: A double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science, 5*(10), 875–899. https://doi.org/10.56726/IRJMETS45312

Ravenda, A., Braun, D., & Ultes, S. (2025). Are LLMs effective psychological assessors? Leveraging adaptive RAG for interpretable mental health screening through psychometric practice. In *Proceedings of the 2025 Conference of the Association for Computational Linguistics (ACL)*

Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *Journal of King Saud University - Computer and Information Sciences*, *34*(10). https://doi.org/10.1016/j.jksuci.2021.11.010

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Leveraging large language models in tourism: A comparative study of the latest GPT Omni models and BERT NLP for customer review classification and sentiment analysis. *Information, 15*(12), Article 792. https://doi.org/10.3390/info15120792

Scangos, K. W., State, M. W., Miller, A. H., Baker, J. T., & Williams, L. M. (2023). New and emerging approaches to treat psychiatric disorders. In *Nature Medicine* (Vol. 29, Issue 2). https://doi.org/10.1038/s41591-022-02197-0

Scherr, S., Arendt, F., Frissen, T., & Oramas, M. J. (2020). Detecting Intentional Self-Harm on Instagram: Development, Testing, and Validation of an Automatic Image-Recognition Algorithm to Discover Cutting-Related Posts. *Social Science Computer Review*, *38*(6). https://doi.org/10.1177/0894439319836389

Sedlakova, J., Daniore, P., Wintsch, A. H., Wolf, M., Stanikic, M., Haag, C., Sieber, C., Schneider, G., Staub, K., Ettlin, D. A., Grübner, O., Rinaldi, F., & von Wyl, V. (2023). Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digital Health*, *2*(10). https://doi.org/10.1371/journal.pdig.0000347

Stein, D. J., Shoptaw, S. J., Vigo, D. v., Lund, C., Cuijpers, P., Bantjes, J., Sartorius, N., & Maj, M. (2022). Psychiatric diagnosis and treatment in the 21st century: paradigm shifts versus incremental integration. *World Psychiatry*, *21*(3). https://doi.org/10.1002/wps.20998

Stene-Larsen, K., & Reneflot, A. (2019). Contact with primary and mental health care prior to suicide: A systematic review of the literature from 2000 to 2017. In *Scandinavian Journal of Public Health* (Vol. 47, Issue 1). https://doi.org/10.1177/1403494817746274

Sutar, R., Kumar, A., & Yadav, V. (2023). Suicide and prevalence of mental disorders: A systematic review and meta-analysis of world data on case-control psychological autopsy studies. In *Psychiatry Research* (Vol. 329). https://doi.org/10.1016/j.psychres.2023.115492

Sykora, M., Elayan, S., & Jackson, T. W. (2020). A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. *Big Data and Society*, *7*(2). https://doi.org/10.1177/2053951720972735

Torous, J., Nicholas, J., Larsen, M. E., Firth, J., & Christensen, H. (2018). Clinical review of user engagement with mental health smartphone apps: Evidence, theory and improvements. In *Evidence-Based Mental Health* (Vol. 21, Issue 3). https://doi.org/10.1136/eb-2018-102891

Turecki, G., Brent, D. A., Gunnell, D., O'Connor, R. C., Oquendo, M. A., Pirkis, J., & Stanley, B. H. (2019). Suicide and suicide risk. In *Nature Reviews Disease Primers* (Vol. 5, Issue 1). https://doi.org/10.1038/s41572-019-0121-0

Twenge, J. M., & Campbell, W. K. (2018). Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-

based study. *Preventive Medicine Reports*, *12*. https://doi.org/10.1016/j.pmedr.2018.10.003

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2018). Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time. *Clinical Psychological Science*, *6*(1). https://doi.org/10.1177/2167702617723376

UNICEF. (2021). *Policy note: Age restrictions on social media* [PDF]. UNICEF. https://www.unicef.org/media/170666/file/Policy%20note_age%20restrictions%20social%20media-new.pdf.pdf

Vajre, P., Wang, S., & Liu, Y. (2021). PsychBERT: A mental health language model for social media mental health behavioral analysis. *IEEE Journal of Biomedical and Health Informatics*, *25*(10), 3873–3882. https://doi.org/10.1109/BIBM52615.2021.9669469

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., le Scao, T., Gugger, S., … Rush, A. M. (2020). *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

World Health Organization. (2022, March 2). COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

World Health Organization. (2022). *World mental health report: Transforming mental health for all*. https://www.who.int/publications/i/item/9789240049338

X. (n.d.). What to do about self-harm and suicide concerns on X. X Help Center. Retrieved June 3, 2025, from https://help.x.com/en/safety-and-security/self-harm-and-suicide

Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2402.13178

Yang, K., Zhang, T., & Ananiadou, S. (2022). A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media. *Information Processing and Management*, *59*(4). https://doi.org/10.1016/j.ipm.2022.102961

Yu, H., & McGuinness, S. (2024). An experimental study of integrating fine-tuned LLMs and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence, 7*, 1–16. https://doi.org/10.21037/jmai-23-1

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2024). Retrieval-augmented generation for AI-generated content: A survey [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2402.19473

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa